JUDGE ANYTHING: MLLM AS A JUDGE ACROSS ANY MODALITY

A PREPRINT

Shu Pu[1*], Yaochen Wang[1*], Dongping Chen[1‡], Yuhang Chen[1§], Guohao Wang[1§], Qi Qin[1§], Zhongyi Zhang[1§], Zhiyuan Zhang[1§], Zetong Zhou[1§], Shuang Gong[1§],, Yi Gui[1], Yao Wan[1†], Philip S. Yu[2]

[1] Huazhong University of Science and Technology
[2] University of Illinois Chicago

ABSTRACT

Evaluating generative foundation models on open-ended multimodal understanding (MMU) and generation (MMG) tasks across diverse modalities (*e.g.*, images, audio, video) poses significant challenges due to the complexity of cross-modal interactions. To this end, the idea of utilizing Multimodal LLMs (MLLMs) as automated judges has emerged, with encouraging results in assessing vision-language understanding tasks. Moving further, this paper extends MLLM-as-a-Judge across modalities to a unified manner by introducing two benchmarks, TASKANYTHING and JUDGEANYTHING, to respectively evaluate the overall performance and judging capabilities of MLLMs across any-to-any modality tasks. Specifically, TASKANYTHING evaluates the MMU and MMG capabilities across 15 any-to-any modality categories, employing 1,500 queries curated from well-established benchmarks. Furthermore, JUDGEANYTHING evaluates the judging capabilities of 5 advanced (*e.g.*, GPT-4o and Gemini-1.5-Pro) from the perspectives of *Pair Comparison* and *Score Evaluation*, providing a standardized testbed that incorporates human judgments and detailed rubrics. Our extensive experiments reveal that while these MLLMs show promise in assessing MMU (*i.e.*, achieving an average of 64.1% in *Pair Comparison* setting and 69.58% in *Score Evaluation* setting), they encounter significant challenges with MMG tasks (*i.e.*, averaging only 50.7% in *Pair Comparison* setting and 47.2% in *Score Evaluation* setting), exposing cross-modality biases and hallucination issues. To address this, we present OMNIARENA, an automated platform for evaluating omni-models and multimodal reward models. Our work highlights the need for fairer evaluation protocols and stronger alignment with human preferences. The source code and dataset are publicly available at: `https://urrealhero.github.io/judgeanything.github.io/`.

1 Introduction

The rapid advancement of generative models, particularly Large Language Models (LLMs) (Hurst et al., 2024; Liu et al., 2024a) and diffusion-based visual generative models (Rombach et al., 2022; Esser et al., 2024), has led to the widespread prevalence of AI-generated content (AIGC) across various modalities, including images (Ghosh et al., 2023), video (Yang et al., 2024d), and audio (Liu et al., 2024b). Recently, the omni-model is proposed to unify pre-training techniques across multiple modalities, aiming to integrate both multimodal understanding (MMU) and multimodal generation (MMG) capabilities (Xie et al., 2024a; Li et al., 2024e; Team, 2024).

Despite this, evaluating the MMU and MMG capabilities of generative models typically relies on human judgment, given the inherently open-ended nature of related tasks. While human evaluations are commonly regarded as the gold standard (Huang et al., 2024; Jiang et al., 2025), they tend to be time-consuming, expensive—particularly for high-dimensional modalities such as video and audio. Additionally, these evaluations are prone to inconsistency, as

* Co-first. ‡ Project Leader.
† Correspondence to: Yao Wan (wanyao@hust.edu.cn).
§ Equal Contribution.

Table 1: Comparison to current works. TASKANYTHING uniquely incorporate diverse modalities and open-ended questions to evaluate omni-models using verified metrics that have been validated against human annotations for potential biases with an automated model arena. JUDGEANYTHING pioneer in assessing MLLM-as-a-Judge across various modalities in *Score Evaluation* and *Pair Comparison* settings. ● means that question types are mixed. See Appendix A for detailed related works.

Benchmark	#Size	Input Modality				Output Modality				Open-ended Question	Verified Metric	Arena
		Text	Image	Video	Audio	Text	Image	Video	Audio			
Multimodal Understanding and Generation												
ISG (Chen et al., 2025)	1,150	✔	✔	✘	✘	✔	✔	✘	✘	✔	✔	✘
MMIE (Xia et al., 2024)	20,103	✔	✔	✘	✘	✔	✔	✘	✘	●	✔	✘
OmniBench (Li et al., 2024g)	1,142	✔	✔	✔	✔	✔	✘	✘	✘	✘	✘	✘
OmnixR (Chen et al., 2024c)	1,800	✔	✔	✔	✔	✔	✔	✔	✔	●	✘	✘
Eval-Anything (Ji et al., 2024)	264	✔	✔	✔	✔	✔	✔	✔	✔	✔	✘	✘
MixEval-X (Ni et al., 2024)	8,300	✔	✔	✔	✔	✔	✔	✔	✔	●	✘	✘
TASKANYTHING (ours)	1,500	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Multimodal LLM-as-a-Judge												
MLLM-as-a-Judge (Chen et al., 2024a)	15,450	✔	✔	✘	✘	✔	✔	✘	✘	✘	✔	N/A
VL-RewardBench (Li et al., 2024d)	1,546	✔	✔	✘	✘	✔	✔	✘	✘	✘	✔	N/A
MM-RewardBench (Yasunaga et al., 2025)	5,211	✔	✔	✘	✘	✔	✔	✘	✘	✘	✔	N/A
JUDGEANYTHING (ours)	9,000	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	N/A

open-ended tasks often lack absolute ground truths or universally accepted evaluation criteria, further complicating reliable assessments.

To this end, researchers have explored automated evaluation methods, particularly by leveraging Multimodal LLMs (MLLMs) as assessment metrics - a concept referred to as MLLM-as-a-Judge (Chen et al., 2024a; Xiong et al., 2024a). This approach introduces automated assessment of vision-and-language tasks, offering both qualitative insights and quantitative scores. While inconsistency, biases, and hallucination remain, MLLM-as-a-Judge has demonstrated utility and promising results across a range of generative tasks—including text-to-image Chen et al. (2024g), text-to-video Luo et al. (2024), and interleaved multimodal generation (Chen et al., 2025; Zhou et al., 2024). It has also been used as a reward model for vision-language alignment (Li et al., 2024d; Yasunaga et al., 2025). These developments bring us to a key question:

Can MLLMs serve as a unified judge for assessing the understanding and generation ability of any-to-any modality tasks?

In other words, can MLLMs extend their human-aligned judgment capabilities—previously demonstrated in text-based (Zhou et al., 2024) and image-based (Chen et al., 2024a,g) tasks (see Table 1)—to a broader range of modalities, such as images, video, and audio? Even if MLLMs cannot fully replicate human judgments, can they still provide meaningful, and reliable assessments that reduce dependence on human evaluation and guide the development of multimodal AI-generated rewards (Lee et al., 2023b; Li et al., 2024d)?

To address the aforementioned questions, we start by introducing a new benchmark, TASKANYTHING, to comprehensively evaluate the capabilities of MLLMs in both MMU and MMG across an *any-to-any* framework. This benchmark consists of 15 open-ended tasks and 1,500 queries sourced from established datasets, providing an unconstrained yet categorically balanced testbed. Next, we collect candidate responses to these queries using state-of-the-art generative models and compile query-specific checklists for evaluation. Finally, we introduce JUDGEANYTHING which incorporates these queries, response candidates, and checklists into *Pair Comparison* and *Score Evaluation* settings, creating a standardized testbed for evaluating the effectiveness of MLLM-as-a-Judge in MMU and MMG against human-annotated judgments.

In our experiments, we specifically evaluate the judging capabilities of five advanced, widely used MLLMs on JUDGEANYTHING, including GPT-4o, Gemini-1.5-Pro, LearnLM-1.5-Pro, and Gemini-2.0-Flash/Lite. Experimental results reveal that MLLMs align more closely with human preferences on *Pair Comparison* than on *Score Evaluation*, with both tasks benefiting from clearly fixed rubrics and the *Checklist* approach. Notably, while MLLMs demonstrate strong judging performance on MMU tasks, their alignment remains limited in MMG tasks, particularly in video and audio generation scenarios. Among these models, Gemini-1.5-Pro stands out due to its robust multimodal perception, long-context reasoning, and instruction-following capabilities, achieving an average 85.19% accuracy on *Pair Comparison* and 0.815 Pearson similarity on *Score Evaluation* in MMU tasks.

To further advance *any-to-any* omni-models and reward models in the multimodal domain, we present OMNIARENA, a standardized testbed for evaluating existing omni-models and multimodal reward models based on our TASKANYTHING and JUDGEANYTHING benchmarks. Our experimental results, leveraging Gemini-1.5-Pro as an automated

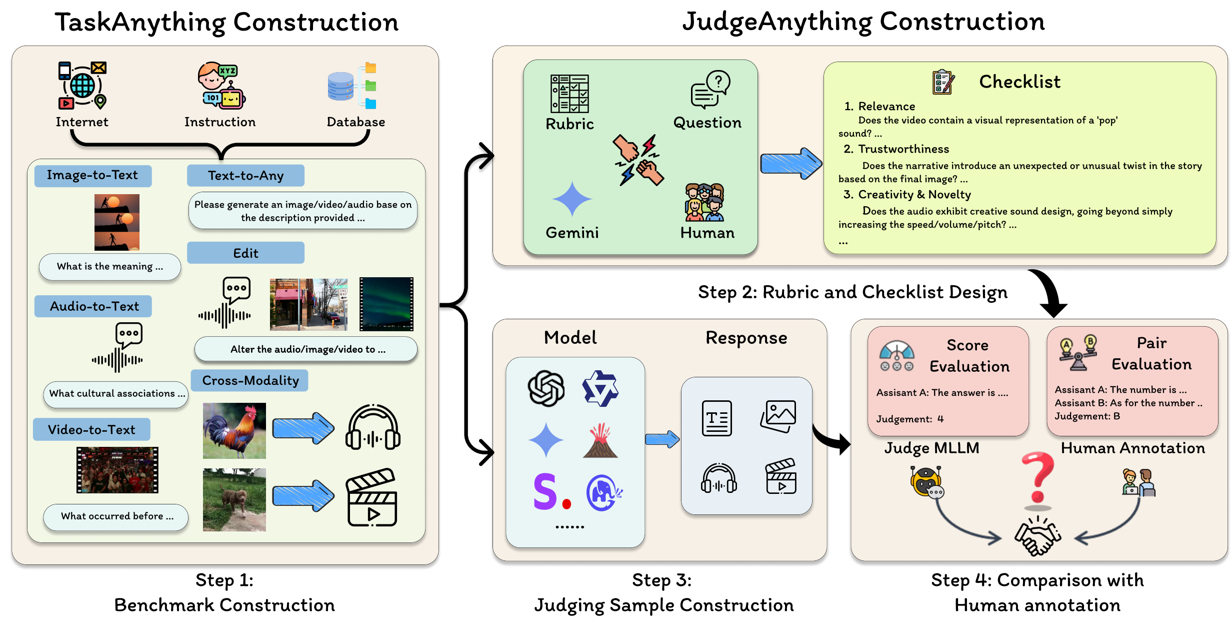


Figure 1: The construction of TASKANYTHING and JUDGEANYTHING follows a systematic four-step approach. First, we compile open-ended *any-to-any* instructions from existing benchmarks and datasets, followed by rigorous human annotation to ensure sample diversity and quality in TASKANYTHING. Subsequently, we collect model responses and develop evaluation principles through an Human-MLLM collaborative approach, creating detailed assessment checklists for each sample. Finally, we curate instruction-responses pairs to evaluate the effectiveness of MLLM-as-a-Judge in *any-to-any* generation tasks, benchmarking these automated assessments against expert human judgments.

judge, demonstrate that Gemini-1.5-Pro excels among omni-models in MMU tasks, while ModaVerse achieves superior performance in MMG tasks. Once deployed, OMNIARENA will facilitate seamless participation from new models and judges in an *any-to-any* fashion, while simultaneously integrating real-world votes from the broader community to collect diverse and representative judgments.

The main contributions of this paper are as follows:

- **Two Benchmarks.** We propose TASKANYTHING, a comprehensive benchmark for evaluating the MMU and MMG capabilities of MLLMs. Building on TASKANYTHING, we also introduce JUDGEANYTHING to extensively assess the judging capabilities of MLLMs using human annotated judgments and fine-grained checklists for each sample in an *any-to-any* manner from the perspectives of *Score Evaluation* and *Pair Comparison.*
- **An Automated Arena for Omni Models.** We develop OMNIARENA, an automated evaluation platform for omni-models that supports diverse modality inputs and outputs, facilitating future research in multimodal generation and understanding.
- **Findings and Implications.** Extensive experiments reveal that current MLLM-as-a-Judge partially align with human judgment while their reliability as judges for open-ended *any-to-any* queries remains significantly limited. Furthermore, although MLLMs enhanced by well-constructed principle rubrics and sample-wise checklists show improvement, they still fall short due to a range of cross-modality biases and hallucinations, undermining their reliability when serving as judges.

2 TASKANYTHING and JUDGEANYTHING

We introduce TASKANYTHING for open-ended *any-to-any* generation evaluation. Based on TASKANYTHING, we propose JUDGEANYTHING to evaluate whether MLLMs can serve as metrics for *any-to-any* generation assessment. As shown in Figure 1, we take a four-step approach to curate the entire benchmark. We provides benchmark construction details in Appendix B.1.

2.1 TaskAnything Construction

We collect samples from previous well-constructed and data-balanced benchmarks, as shown in Table 5, followed by manually selection to filter out similar and not open-source samples. For **MMU** tasks (*e.g.*, Video-to-Text), we further incorporate human refinements to remove predefined constraints (*e.g.*, output format) and ensure a more natural, free-form structure. For **MMG** tasks (*e.g.*, text-to-video), we filter out NSFW content and low-quality queries to ensure the query can be answered. For some tasks where the field remains relatively underexplored, like visual-to-audio, we collect samples using a human-in-the-loop approach to curate diverse queries. These queries are sourced from video datasets scraped and filtered from YouTube[1], including (Zhang et al., 2024b; Chen et al., 2020), ensuring relevance and diversity. Finally, we successfully curate a high quality and comprehensive open-ended *any-to-any* benchmark dataset $\mathbb{Q}$, comprising 1,500 queries, with each task containing 100 queries.

2.2 Rubric and Checklist Design

We adopt a standardized assessment framework in addition to directly prompting models to assign scores or choose for a more fine-grained evaluation. Building on recent studies (Li et al., 2024a; Gu et al., 2024), we define six evaluation principle rubrics for comprehensive assessment, detailed in Appendix C. To specialize these rubrics for each sample, we prompt Gemini-1.5-Pro (Team et al., 2024a) to generate task-specific checklists based on principle rubrics and open-ended queries. However, we observe that Gemini-1.5-Pro (Team et al., 2024a) demonstrates limited instruction-following capability in video and audio modalities. To mitigate this limitation, we employ a two-step process: first, generating captions for video or audio content, and then using these captions as context to refine the checklist for these tasks. Finally, we manually select 1 to 6 items from the synthetic checklist to construct the final checklist.

2.3 Judging Sample Construction

For each *any-to-any* task, we utilize four *state-of-the-art* models (see Tables 7 and 8 for details) to generate responses to the queries, resulting in a total response set $\mathbb{R}$ of 6,000 entries. These responses are then manually reviewed to ensure quality, with strict adherence to the NSFW guidelines. We use both *Score Evaluation* and *Pair Comparison* to evaluate MLLM-as-a-Judge across various modalities. *Score Evaluation* requires the model to provide an integer rating from 1 to 5, where 1 represents the worst performance and 5 represents the best. *Pair Comparison*, on the other hand, asks the model to select the better option or declare a tie between two candidate responses. At this stage, we construct judging samples for *Score Evaluation* and *Pair Comparison* as follows:

- $\mathbb{D}_{\text{score}} = \{(Q_i, R_i) \mid Q_i \in \mathbb{Q},\ R_i \in \mathbb{R}\}$
- $\mathbb{D}_{\text{pair}} = \{(Q_i, R_i^1, R_i^2) \mid Q_i \in \mathbb{Q},\ R_i^1, R_i^2 \in \mathbb{R},\ R_i^1 \neq R_i^2\}$

The $\mathbb{D}_{\text{score}}$ dataset contains question-response pairs for absolute evaluation, while $\mathbb{D}_{\text{pair}}$ consists of triples for comparative assessment between two different responses. Responses R_i^1 and R_i^2 in each pair are systematically sampled from different models to ensure diverse comparisons.

2.4 Comparison with Human Annotations

We collect the ground truth of these judging problems from 10 expert annotators. These annotators are proficient in AIGC content, with different genders, ages, and educational backgrounds to ensure data quality and diversity. They are require to give objective judgments that strictly following our rules and instructions without any bias that could undermine the fairness of the judgments (Ye et al., 2024). We also conduct annotation on checklists to capture human preferences in a fine-grained manner. See Appendix B.3 for further details. ***We implement cross-validation between different annotators for each sample and conduct continuous monitoring to ensure they maintain objectivity and fairness.***

Table 2: Data statistics for constructing TaskAnything and JudgeAnything. Each sample from human annotator are under cross-validation.

Step	Input	Num.	Output	Per Input	Total
1	Previous Benchmarks	/	Open-ended Instructions	/	1500
2	Instructions	1500	Human-annotated checklist	19.11	28673
3	Instructions	1500	Model Responses	4	6000
	Instructions + Responses	1500	*Pair Comparison* *Score Evaluation*	2 4	3000 6000
4	*Pair Comparison* *Score Evaluation*	3000 6000	Human Annotation	$5*3$ $5*3$	45000 90000

[1] https://youtube.com

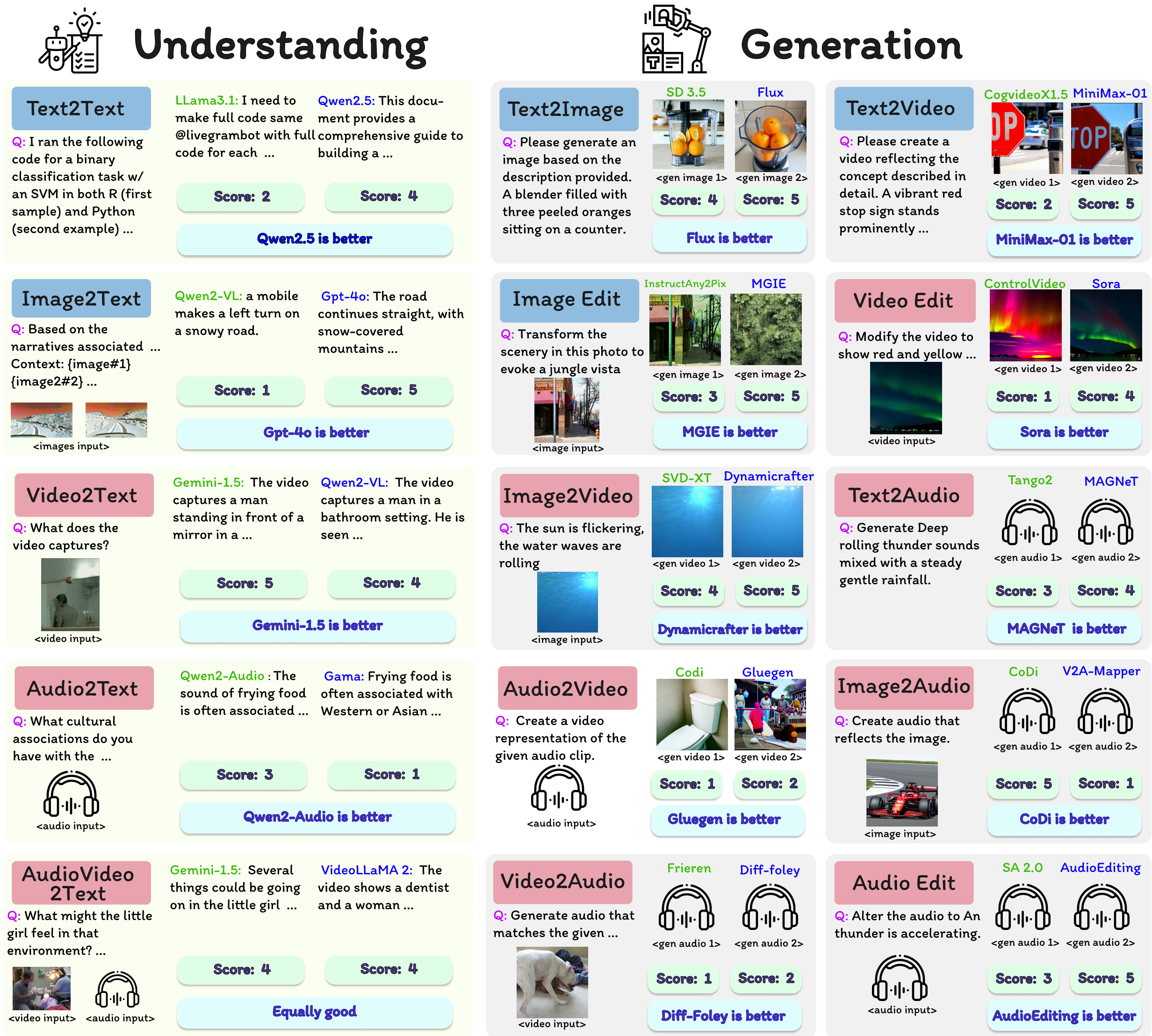


Figure 2: TASKANYTHING and JUDGEANYTHING comprise 15 *any-to-any* combinations spanning text, image, video, and audio modalities. The TASKANYTHING samples are curated from established benchmarks, while responses to queries are generated using *state-of-the-art* models to construct JUDGEANYTHING in both *Pair Comparison* and *Score Evaluation* settings.

3 Experiments and Analysis

Using JUDGEANYTHING, we conduct experiments to evaluate the judging abilities of MLLMs (*i.e.*, MLLM-as-a-Judge) across modalities in both *Score Evaluation* and *Pair Comparison* settings.

3.1 Experimental Setup

Judging Models. We utilize five advanced proprietary models—GPT-4o (Hurst et al., 2024), Learnlm-1.5-pro-experimental (Team et al., 2024b), Gemini-1.5-Pro (Team et al., 2024a), Gemini-2.0-Flash, Gemini-2.0-Flash-lite—selected for their strong understanding, robust generative performance across multiple modalities, and effective instruction-following capabilities. Ultimately, we compare the judging models with the *evaluator-fusion*. We define *evaluator-fusion* as the average score in *Score Evaluation* and majority-voting in *Pair Comparison*. To clarify, given that GPT-4o cannot receive both audio and visual content, we leverage GPT-4o-audio-preview as a replacement for audio-visual task. Also, we have experimented with the state-of-the-art open-source omni-models

Table 3: Model performance on TASKANYTHING. We **bold** the best and <u>underline</u> the second under each baseline.

Models	Multimodal Understanding						Multimodal Generation					
	Pair Comparison		Score Evaluation				Pair Comparison		Score Evaluation			
	w. Tie	w.o. Tie	Agreement	Pearson	Spearman	MAE	w. Tie	w.o. Tie	Agreement	Pearson	Spearman	MAE
Overall												
GPT-4o	59.88	75.19	**38.55**	0.488	0.449	**0.895**	55.54	69.78	34.02	0.476	0.467	**1.113**
Gemini-1.5-Pro	**63.39**	**80.48**	37.60	<u>0.494</u>	<u>0.467</u>	0.985	**60.71**	**74.43**	<u>36.43</u>	0.404	0.413	1.185
LearnLM-1.5-Pro	<u>62.37</u>	<u>79.60</u>	35.32	0.473	0.447	1.039	57.48	70.72	36.13	0.377	0.384	1.247
Gemini-2.0-Flash	55.57	73.21	34.85	0.437	0.350	1.054	53.69	67.42	**36.78**	0.460	<u>0.460</u>	1.372
Gemini-2.0-Flash-Lite	52.16	68.75	36.00	0.454	0.386	1.036	45.81	59.67	34.13	0.458	0.442	1.347
Evaluator-Fusion	61.79	78.57	<u>37.95</u>	**0.549**	**0.513**	<u>0.902</u>	<u>57.50</u>	<u>73.09</u>	30.10	**0.556**	**0.555**	<u>1.120</u>
Rubrics												
GPT-4o	65.33	75.45	43.78	<u>0.636</u>	<u>0.620</u>	**0.845**	32.32	52.92	31.85	0.415	0.421	1.289
Gemini-1.5-Pro	**71.80**	**81.69**	<u>44.75</u>	0.626	0.608	<u>0.846</u>	**61.38**	**73.05**	**38.81**	<u>0.459</u>	<u>0.464</u>	**1.122**
LearnLM-1.5-Pro	<u>69.46</u>	<u>79.07</u>	**45.13**	0.618	0.607	0.869	<u>59.74</u>	<u>70.08</u>	<u>38.52</u>	0.436	0.440	<u>1.161</u>
Gemini-2.0-Flash	46.83	67.45	41.16	0.555	0.525	1.035	39.88	60.53	30.42	0.377	0.384	1.589
Gemini-2.0-Flash-Lite	52.62	68.16	39.96	0.562	0.549	1.063	40.52	57.85	31.82	0.433	0.443	1.467
Evaluator-Fusion	67.87	78.81	40.31	**0.692**	**0.687**	0.904	51.55	66.68	27.18	**0.556**	**0.569**	1.242
Checklist												
GPT-4o	62.35	73.65	46.83	0.685	0.663	0.744	30.22	51.94	32.93	0.366	0.364	1.259
Gemini-1.5-Pro	**74.95**	**85.19**	**62.83**	**0.815**	**0.796**	**0.489**	**64.39**	**75.96**	**48.32**	<u>0.585</u>	<u>0.587</u>	**0.883**
LearnLM-1.5-Pro	<u>69.35</u>	78.98	<u>49.51</u>	0.717	0.700	0.726	<u>60.04</u>	70.45	<u>40.53</u>	0.478	0.479	1.086
Gemini-2.0-Flash	53.87	70.16	44.26	0.605	0.586	0.892	49.81	67.73	37.93	0.483	0.487	1.222
Gemini-2.0-Flash-Lite	56.03	68.28	43.07	0.597	0.590	0.901	49.11	65.47	36.68	0.439	0.445	1.175
Evaluator-Fusion	68.22	<u>79.28</u>	46.80	<u>0.756</u>	<u>0.746</u>	<u>0.718</u>	59.29	<u>72.42</u>	33.14	**0.600**	**0.606**	<u>0.992</u>

including Baichuan-Omni-1.5 (Li et al., 2025) and VideoLlama2 (Cheng et al., 2024). However, none of these models could handle long-context input or generate meaningful feedback, elaborated in Appendix C.3.

Three Baselines. We provide three different judging settings: The ***Overall*** setting leverages a direct judging approach, where models first provide reasoning and then deliver a final judgment. ***Rubric*** setting introduces well-defined general foundation rubrics within context and requires models to judge based on fine-grained rubrics before making a final judgment. In the ***Checklist*** setting, MLLMs are provided with detailed checklists curated through a human-in-the-loop process and must first evaluate responses based on these checklists before delivering their final judgment.. For all settings, we employ an *"Analyze-then-Judge"* chain-of-thought (Wei et al., 2022) pattern to elicit models' judging capabilities. To improve robustness and mitigate variance, we sample all judgments three times with slightly modified prompts and take the average of the results.

Implementations. We set the temperature to 0.7 for all judging models, as previous research (Liu et al., 2023b; Chen et al., 2024a) has reported a high correlation with human annotators at this setting. All models are configured to generate structured outputs across both evaluation paradigms. For the *Score Evaluation* setting, we provide detailed explanatory descriptions for each integer value on the 1-5 scale, enabling informed judgments based on explicit criteria. For the *Pair Comparison* setting, we offer three categorical choices: *"first"*, *"second"*, and *"tie"*, conducting experiments with switched response positions to mitigate potential position bias. All experiments for judging models are replicated three times, with the averaged score (for *Score Evaluation*) and majority selection (for *Pair Comparison*) used to calculate final results. We analyze performance using four established metrics—Agreement, Pearson correlation, Spearman correlation (Lee Rodgers & Nicewander, 1988), Mean Absolute Error (MAE) for the *Score Evaluation* setting, and accuracy for the *Pair Comparison* setting. See Appendix C for comprehensive experimental protocols.

3.2 Quantitative Results

Gemini-1.5-Pro is the best evaluator in *any-to-any* task evaluation in our experiment. Table 3 shows that Gemini-1.5-Pro nearly surpasses all other judging models like GPT-4o and the newer Gemini-2.0 series, especially in *Checklist* settings, reaching 0.815 Pearson similarity with human annotators. This demonstrates that larger models tend to perform better than smaller models in simulating human-like judgments. Moreover, we observe that GPT-4o is less capable when handling temporal modalities like video and audio. Its limited cross-modality capabilities make it a suboptimal choice for serving as a unified judge for *any-to-any* tasks. Evaluator-Fusion, which leverages majority voting to simulate human preferences from multiple MLLMs' collective wisdom, excels in *Score Evaluation* evaluation particularly for MMG tasks, achieving 0.600 Pearson similarity with *Checklist*.

MLLM-as-a-Judge performs better in MMU rather than MMG tasks. As shown in Table 4, MLLMs' judgments align more closely with human evaluations in MMU tasks compared to MMG tasks in both *Score Evaluation* and *Pair*

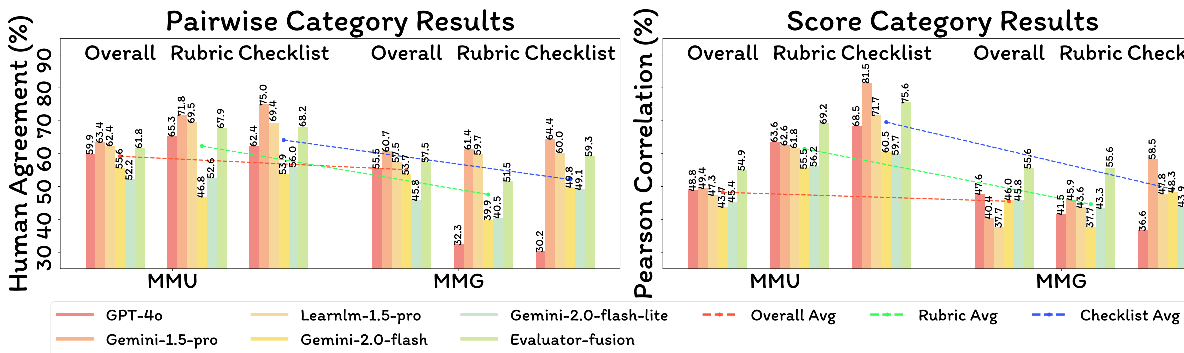


Figure 3: Visualization of MMU and MMG categories with human agreement data. **Left:** Accuracy scores for the *Pair Comparison* setting across two categories. **Right:** Agreement scores for the *Score Evaluation* setting across two categories. The dotted line connects the same baseline from MMU to MMG to highlight the trend.

Comparison settings, particularly in text-to-text and image-to-text scenarios. Moreover, we observe that judging models benefit significantly more from fine-grained evaluation criteria in MMU than MMG tasks. We attribute this pattern to the fact that current judging models perform better at evaluating tasks they themselves are capable of executing, leading to more accurate and fair assessments in these domains. This finding corresponds with previous research indicating that understanding is the foundation of generation capability. Consequently, the ability to effectively judge open-ended MMU tasks may develop before MMG evaluation capacity, as the inherent complexity and variability within text is substantially lower than in other modalities.

Finding 1: Judging models are more reliable on MMU task, and *Checklist* can greatly improve the alignment.

Less-aligned modalities like video and audio pose significant challenges to MLLM-as-a-Judge in cross-modality judging. Diving deeper into MLLM-as-a-Judge across modalities, Table 4 shows that current judging models' performance declines when it comes to low-frequency cross-modality tasks like image-to-audio and audio-to-text. As shown in Figure 4, different output modalities matter more compared to input modality, revealing a clear decreasing trend from well-aligned modalities like text and image to less-aligned modalities like video and audio. Given that current judging models are primarily trained to evaluate text-to-text or image-to-text scenarios, cross-modal evaluation capabilities remain an emergent property in their early developmental stages. To enhance this capability, we recommend incorporating a more diverse range of cross-modality judging samples into training.

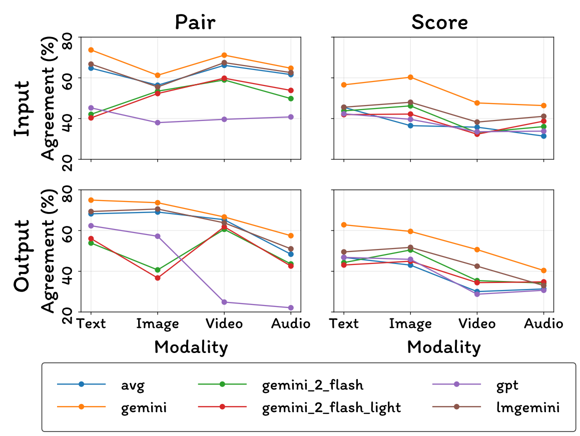


Figure 4: Visualization of modality effect on human agreement across two settings using *checklist-of-thought*.

Finding 2: **Output** modality matters more for MLLM-as-a-Judge, with results showing a clear decreasing trend from well-aligned to less-aligned modality.

MLLM-as-a-Judge benefits from fine-grained rubrics and checklists, providing more human-aligned judgments. As shown in Figure 3, most judging models benefit from sample-wise well-curated judging checklists, where *Checklist* approaches yield more human-aligned evaluations than providing overall judgments or fixed foundation rubrics. This improved performance is particularly evident in Gemini's case, which leverages its superior 1M long-context window and instruction following capability to effectively provide human-like judgments. However, these detailed judging guidelines prove to be a double-edged sword, causing GPT-4o's performance to decline when using *Rubric* and *Checklist* formats compared to direct judgment.

Table 4: Detailed breakdown on each *any-to-any* generation tasks on JUDGEANYTHING. Well-constructed rubrics and *checklist-of-thought* enhance MLLMs' alignment with human when serving as judges. We **bold** the best.

Models	Setting	Overall	Multimodal Understanding					Multimodal Generation									
			T→T	I→T	V→T	A→T	V+A→T	T→I	T→V	T→A	I→I	I→V	I→A	V→V	V→A	A→V	A→A
Pair Comparison																	
GPT-4o	*Overall*	**56.99**	**52.50**	**71.07**	58.00	65.50	52.50	52.00	**54.00**	31.00	**53.77**	**53.54**	**67.50**	**79.00**	**43.50**	**62.00**	**59.18**
	Rubric	43.33	51.67	66.50	**64.67**	**76.50**	**67.33**	**78.17**	23.33	**42.25**	44.17	31.75	16.42	2.33	28.17	47.92	8.75
	Checklist	40.92	47.33	64.92	61.75	73.25	64.50	70.25	23.50	39.92	44.17	31.58	11.33	2.25	30.00	42.08	7.08
Gemini-1.5-Pro	*Overall*	61.61	63.50	**72.59**	68.00	65.50	47.50	78.00	54.00	52.00	**56.28**	**54.04**	**68.50**	89.50	39.50	**52.00**	63.27
	Rubric	64.86	66.33	70.58	74.25	75.67	72.17	80.92	71.42	50.50	48.83	51.08	65.75	89.75	**41.08**	47.67	66.83
	Checklist	**67.91**	**72.33**	72.58	**81.33**	**75.92**	**72.58**	**94.42**	**72.75**	**55.33**	52.92	53.83	65.75	**90.25**	40.42	49.83	**68.42**
Gemini-2.0-Flash	*Overall*	**54.31**	**42.00**	**65.99**	**57.00**	**62.50**	50.50	**51.00**	57.50	**44.00**	**55.78**	42.93	57.00	77.50	**47.50**	**63.00**	**40.31**
	Rubric	42.20	32.92	51.50	42.25	50.08	57.42	46.58	55.25	36.25	49.25	36.42	31.67	56.92	36.25	30.58	19.67
	Checklist	51.16	37.58	59.25	49.75	55.17	**67.58**	30.67	**62.67**	37.42	50.67	**43.67**	**60.25**	**78.33**	40.17	57.75	36.50
Score Evaluation																	
GPT-4o	*Overall*	35.53	37.75	44.00	33.25	38.50	39.25	40.25	**43.50**	43.25	44.25	26.75	30.00	13.25	25.75	**40.50**	**32.75**
	Rubric	35.83	50.00	48.79	38.04	42.96	39.08	**45.25**	28.17	**44.92**	37.54	30.79	**32.00**	12.50	**28.08**	33.46	25.79
	Checklist	**37.56**	**52.33**	**52.25**	**38.88**	**44.17**	**46.54**	43.29	28.92	44.83	**48.42**	**32.46**	25.37	**21.79**	26.92	31.75	25.54
Gemini-1.5-Pro	*Overall*	36.82	39.75	44.00	39.50	36.75	28.00	52.25	14.25	42.25	52.25	27.00	35.00	35.75	28.25	**48.50**	28.75
	Rubric	40.79	48.54	49.63	38.96	46.83	39.79	46.13	36.54	45.67	57.63	44.92	35.46	32.75	26.29	40.54	22.21
	Checklist	**53.16**	**66.38**	**64.71**	**51.04**	**61.12**	**70.88**	**57.96**	**46.71**	**55.25**	**61.21**	**68.63**	**46.71**	**41.29**	**32.08**	45.92	**32.08**
Gemini-2.0-Flash	*Overall*	36.13	36.75	35.00	38.75	34.75	29.00	**52.75**	23.00	37.50	42.00	34.50	**52.50**	11.50	**39.75**	**59.75**	14.50
	Rubric	34.00	49.79	47.13	38.33	40.58	29.96	47.00	29.33	38.13	30.46	35.54	33.71	12.21	31.58	33.83	12.42
	Checklist	**40.03**	**50.33**	**48.79**	**41.13**	**42.58**	**38.46**	48.17	**33.00**	**43.58**	**52.75**	**40.13**	43.25	**19.96**	32.88	48.54	**17.04**

3.3 In-Depth Analysis

We conduct a more nuanced analysis of the underlying factors behind these quantitative results, seeking to understand both the strengths and limitations of MLLM-as-a-Judge approaches. We examine what contributes to their strong performance in certain contexts and what fundamental challenges undermine their reliability in others.

How fine-grained rubrics serve as *"two-blade sword"* for judgment? Tables 3 and 4 illustrate that incorporating *Checklist* approaches generally improves judging models' performance, enabling more accurate assessments. For example, in a video-to-text task (Figure 15 in Appendix), when the judging model is asked to evaluate a response's creativity, the *Checklist* framework helps calibrate the score from 4 to 2, bringing it more in alignment with user evaluations. However, applying *Checklist* can sometimes yield counterproductive results. When examining the reasoning chains employed during judgment formation, we discovered that such fine-grained rubrics may introduce misleading information and trigger serious hallucinations for certain tasks. A representative example appears in Figure 16, where a human-refined *Checklist* designed to evaluate gun imagery without harmful content leads Gemini-1.5-Pro to misinterpret the gun figure itself as inherently harmful content, resulting in an inappropriately low score on the trustworthiness rubric.

Why do fine-grained rubrics and checklists enhance MMU evaluation more significantly than MMG assessment? As previously mentioned, MLLM-as-a-Judge performance varies considerably across modalities and tasks, with a particularly notable distinction between MMU and MMG evaluation scenarios. To systematically investigate this discrepancy, we decompose our benchmark into minimal task compositions (Table 4) and categorize them into ***Text-Driven Understanding*** and ***Non-Text-Driven Generation*** tasks. Our analysis reveals that fine-grained assessment frameworks like *Rubric* and *Checklist* consistently improve alignment with human judgments in MMU tasks, while yield performance declination and occasionally artifacts in MMG tasks.

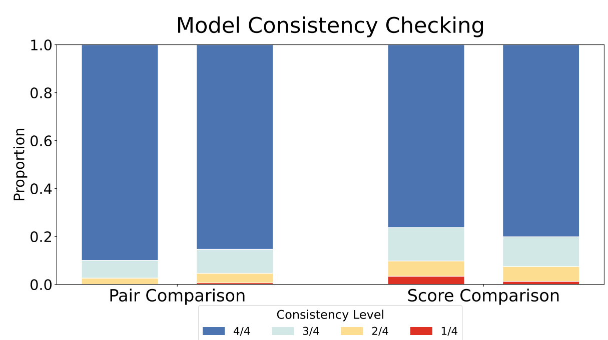


Figure 5: Consistency checking across four repeated experiments with identical prompts on Gemini-1.5-Pro (left) and Gemini-2.0-Flash (right).

Figure 17 demonstrates that GPT-4o struggles to effectively discriminate between query inputs and model responses in certain MMG contexts. When examining this phenomenon more closely, we identified a prevalent *"Choose-tie"* tendency among judging models in MMG tasks, as documented in Table 6. Notably, these models disproportionately select *"tie"* when using subdivided *Rubric* criteria, whereas human evaluators typically rely more on holistic impressions for MMG outputs and finally result an inconsistent result from judging models to human.

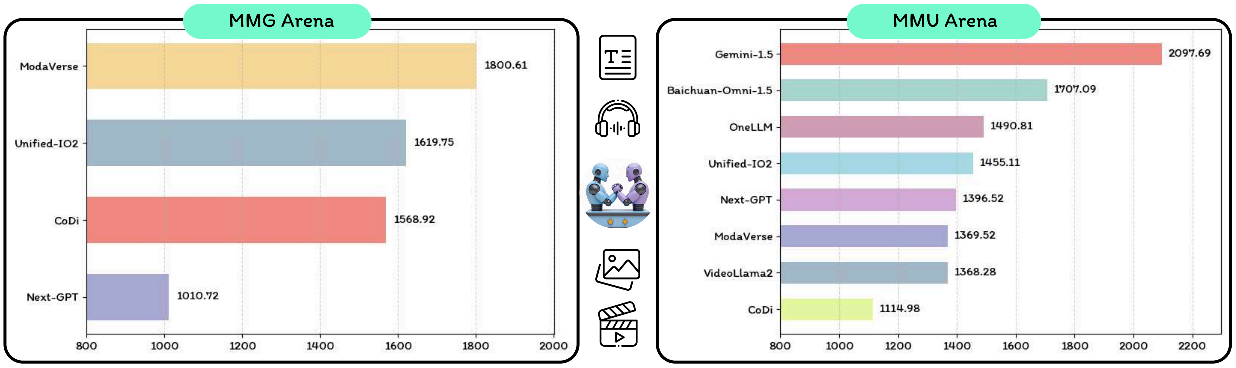


Figure 6: Overview of our OmniArena. ModaVerse outperforms other omni-models on open-ended MMG tasks. For MMU, Gemini-1.5-pro shows incredibly performance with its long-context and cross-modality reasoning capability.

To further quantify this divergence, we generate *Overall-Rubric* correlation heatmaps comparing both judging models and human annotations (Figure 9). These visualizations reveal a striking pattern: while both human annotators and judging models initially achieve *Overall-Rubric* correlation scores above 0.5 in MMU tasks, transitioning to MMG tasks produces opposite effects—correlation scores decline for all judging models but notably increase for human annotators.

A critical factor underlying this discrepancy appears to be the contextual relevance of evaluation criteria. Many *Rubric* dimensions become inherently less meaningful in certain cross-modal MMG settings (*e.g.*, assessing *"Coherence"* for video-to-audio tasks). When confronted with model responses that perform similarly according to these less effective rubrics (Figures 18 and 19), human evaluators tend to prioritize responses with better overall performance, while judging models mechanistically default to *"tie"*.

Finding 3: The effectiveness gap of fine-grained evaluation frameworks between MMU and MMG tasks stems from the contextual applicability of standardized fine-grained criteria across diverse multimodal outputs.

While *Checklist* approaches demonstrably enhance judgment alignment in MMU tasks, our findings highlight the need for more dynamic and context-sensitive evaluation frameworks for MMG assessment. The cognitive divergence between human holistic judgment and model-based analytical assessment in MMG contexts represents a significant challenge for developing unified evaluation metrics across the full spectrum of multimodal tasks. We present detailed case studies in Appendix D that provide task-specific analyses of these evaluation patterns.

Inert inconsistency undermines reliability of MLLM-as-a-Judge. To evaluate the consistency of decision-making, we perform four repeated tests under the *Overall* and *Checklist* baselines, calculating the Majority Consistency Criterion (MCC) ratios for each test. This analysis compares the consistency of Gemini-1.5-Pro and Gemini-2.0-Flash across different settings.

As shown in Figure 5, Gemini-1.5-Pro achieves higher consistency (0.9) under the *Pair Comparison* setting compared to Gemini-2.0-Flash. However, in the *Score Evaluation* setting, Gemini-1.5-Pro's consistency drops to 0.763, while Gemini-2.0-Flash maintains a score above 0.8. Figures 10 and 11 reveal that Gemini-1.5-Pro performs well in *Overall* and relevance, but its consistency declines in other rubrics. Gemini-2.0-Flash shows stable performance in *Overall*, relevance, and trustworthiness, but is unstable in rubrics like clarity, coherence, and completeness, with MCC ratios below 0.6 in *Pair Comparison*. These results suggest that Gemini-1.5-Pro is more reliable in *Pair Comparison* evaluations but struggles with *Score Evaluation*. The decline in *Rubric* performance indicates that additional rubrics may introduce uncertainty in decision-making.

4 OmniArena

Based on TaskAnything and JudgeAnything, we propose OmniArena, a standardized platform to reliably assess the performance of omni-models. OmniArena leverages open-ended queries in TaskAnything and operates through a pairwise comparison mechanism, where judging models or users are presented with two responses and asked to determine the superior outcome. Participants can add their omni-models into OmniArena, select their preferred results, and even introduce innovative questions, allowing for dynamic, creative testing.

OmniArena Setups. As previously mentioned, *any-to-any* tasks can be categorized into two distinct subtypes: MMU and MMG, leading to OmniArena being structured into two separate parts. In our experiment, Gemini-1.5-Pro serves as judging models for its superior performance in JudgeAnything. After obtaining the automatic judging results in OmniArena, we employ the **ELO Rating System** (Elo, 1966) (see Appendix B.5 for technical details)to establish a dynamic ranking platform for evaluating models on two sub-arenas. After each match between two models, the ELO ratings are updated based on the outcome, providing a quantitative measure of model performance based on pairwise comparisons.

Experiment Results. As shown in Figure 6, Gemini-1.5-pro achieves the highest ELO score in the MMU arena. Notably, MMU expertise surpasses omni-models in OmniArena, highlighting the superior performance of specialized models in this task. Interestingly, Next-GPT (Wu et al., 2023a) excels in MMU tasks but underperforms in MMG tasks. This discrepancy arises from its limited control mechanisms and insufficient instruction-following ability, which impedes its capacity to generate multimodal outputs in MMG tasks. The lack of access to open-source models remains a significant constraint. However, with OmniArena's growing capability, combined with increasingly refined benchmarks, we expect the continuous introduction of new models and the ongoing enhancement of the evaluation framework for a better future.

5 Discussion and Conclusion

In summary, this work presents a holistic assessment of MLLMs as a unified metric for MMU and MMG tasks by introducing two benchmarks spanning 15 types of *any-to-any* tasks in *Pair Comparison* and *Score Evaluation* settings. Our comprehensive experiments reveal the limitations of current advanced MLLMs when serving as judges, uncovering biases and potential issues that provide insights for future research.

LLM-as-a-Judge has been widely utilized for automated open-ended natural language generation assessment and served as supervised rewards in model training. However, as AI capabilities expand beyond text to encompass rich multimodal interactions, we urgently need evaluation frameworks that reflect human values across modalities. Our findings highlight the critical need for developing more sophisticated cross-modal evaluation protocols that can better capture nuanced human preferences. We hope our work can provide a standard testbed to streamline the evaluation process, reduce dependence on human labor, and facilitate the development of more human-aligned *any-to-any* generative models.

References

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

AI, R. How to create sota image generation with text: Recraft's ml team insights. `https://www.recraft.ai/blog/how-to-create-sota-image-generation-with-text-recrafts-ml-team-insights`, 2024.

Anthropic. Claude 3.5 sonnet model card addendum. Online, 2023. URL `https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf`.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Audio, S. Stable audio 2.0. `https://www.stableaudio.com/user-guide/model-2`, 2024.

Bachmann, R., Kar, O. F., Mizrahi, D., Garjani, A., Gao, M., Griffiths, D., Hu, J., Dehghan, A., and Zamir, A. 4m-21: An any-to-any vision model for tens of tasks and modalities. *ArXiv*, abs/2406.09406, 2024. URL `https://api.semanticscholar.org/CorpusID:270520159`.

Bai, S. and An, S. A survey on automatic image caption generation. *Neurocomputing*, 311:291–304, 2018.

Bai, S., Yang, S., Bai, J., Wang, P., Zhang, X., Lin, J., Wang, X., Zhou, C., and Zhou, J. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*, 2023.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y., and Ramesh, A. Improving image generation with better captions. `https://cdn.openai.com/papers/dall-e-3.pdf`, 2024.

Bitton, Y., Bansal, H., Hessel, J., Shao, R., Zhu, W., Awadalla, A., Gardner, J., Taori, R., and Schmidt, L. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*, 2023.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., and Rombach, R. Stable video diffusion: Scaling latent video diffusion models to large datasets. https://static1.squarespace.com/static/6213c340453c3f502425776e/t/655ce779b9d47d342a93c890/1733935148453/stable_video_diffusion.pdf, 2023.

Brooks, T., Holynski, A., and Efros, A. Instructpix2pix: Learning to follow image editing instructions. arxiv. *arXiv preprint arXiv:2211.09800*, 2022.

Cai, M., Tan, R., Zhang, J., Zou, B., Zhang, K., Yao, F., Zhu, F., Gu, J., Zhong, Y., Shang, Y., et al. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024.

Chen, D., Chen, R., Zhang, S., Liu, Y., Wang, Y., Zhou, H., Zhang, Q., Wan, Y., Zhou, P., and Sun, L. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*, 2024a.

Chen, D., Chen, R., Pu, S., Liu, Z., Wu, Y., Chen, C., Liu, B., Huang, Y., Wan, Y., Zhou, P., and Krishna, R. Interleaved scene graph for interleaved text-and-image generation assessment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=rDLgnYLM5b.

Chen, H., Xie, W., Vedaldi, A., and Zisserman, A. Vggsound: A large-scale audio-visual dataset. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., and Shan, Y. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7310–7320, 2024b.

Chen, L., Hu, H., Zhang, M., Chen, Y., Wang, Z., Li, Y., Shyam, P., Zhou, T., Huang, H., Yang, M.-H., et al. Omnixr: Evaluating omni-modality language models on reasoning across modalities. *arXiv preprint arXiv:2410.12219*, 2024c.

Chen, X., Lin, Y., Zhang, Y., and Huang, W. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. In *European Conference on Computer Vision*, pp. 179–195. Springer, 2024d.

Chen, Y., Lan, Y., Zhou, S., Wang, T., and Pan, X. Sar3d: Autoregressive 3d object generation and understanding via multi-scale 3d vqvae. *arXiv preprint arXiv:2411.16856*, 2024e.

Chen, Y., Yue, X., Zhang, C., Gao, X., Tan, R. T., and Li, H. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*, 2024f.

Chen, Z., Du, Y., Wen, Z., Zhou, Y., Cui, C., Weng, Z., Tu, H., Wang, C., Tong, Z., Huang, Q., et al. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? *arXiv preprint arXiv:2407.04842*, 2024g.

Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Chern, E., Su, J., Ma, Y., and Liu, P. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024.

Chu, Y., Xu, J., Zhou, X., Yang, Q., Zhang, S., Yan, Z., Zhou, C., and Zhou, J. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Doh, S., Choi, K., Lee, J., and Nam, J. Lp-musiccaps: Llm-based pseudo music captioning. In *ISMIR*, pp. 409–416, 2023. URL https://doi.org/10.5281/zenodo.10265311.

Drossos, K., Lipping, S., and Virtanen, T. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740. IEEE, 2020.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Elo, A. *The USCF Rating System: Its Development, Theory, and Applications*. United States Chess Federation, 1966. URL https://books.google.com/books?id=onUazQEACAAJ.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Evans, Z., Parker, J. D., Carr, C., Zukowski, Z., Taylor, J., and Pons, J. Stable audio open. *arXiv preprint arXiv:2407.14358*, 2024.

Fan, F., Luo, C., Gao, W., and Zhan, J. Aigcbench: Comprehensive evaluation of image-to-video content generated by ai. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 3(4):100152, 2023.

Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

Fu, T.-J., Hu, W., Du, X., Wang, W. Y., Yang, Y., and Gan, Z. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023.

Ghosh, D., Hajishirzi, H., and Schmidt, L. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.

Ghosh, S., Kumar, S., Seth, A., Evuru, C. K. R., Tyagi, U., Sakshi, S., Nieto, O., Duraiswami, R., and Manocha, D. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. *arXiv preprint arXiv:2406.11768*, 2024.

Gong, Y., Luo, H., Liu, A. H., Karlinsky, L., and Glass, J. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*, 2023.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

Han, J., Gong, K., Zhang, Y., Wang, J., Zhang, K., Lin, D., Qiao, Y., Gao, P., and Yue, X. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26584–26595, 2024.

Hong, J., Yan, S., Cai, J., Jiang, X., Hu, Y., and Xie, W. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*, 2025.

Hu, K., Wu, P., Pu, F., Xiao, W., Zhang, Y., Yue, X., Li, B., and Liu, Z. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

Huang, K., Sun, K., Xie, E., Li, Z., and Liu, X. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Iashin, V. and Rahtu, E. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791*, 2021.

Ji, J., Zhou, J., Lou, H., Chen, B., Hong, D., Wang, X., Chen, W., Wang, K., Pan, R., Li, J., et al. Align anything: Training all-modality models to follow instructions with language feedback. *arXiv preprint arXiv:2412.15838*, 2024.

Jia, Y., Chen, Y., Zhao, J., Zhao, S., Zeng, W., Chen, Y., and Qin, Y. Audioeditor: A training-free diffusion-based audio editing framework. *arXiv preprint arXiv:2409.12466*, 2024.

Jiang, D., Ku, M., Li, T., Ni, Y., Sun, S., Fan, R., and Chen, W. Genai arena: An open evaluation platform for generative models. *Advances in Neural Information Processing Systems*, 37:79889–79908, 2025.

Kang, J., Poria, S., and Herremans, D. Video2music: Suitable music generation from videos using an affective multimodal transformer model. *Expert Systems with Applications*, 249:123640, 2024.

Khattak, M. U., Naeem, M. F., Hassan, J., Naseer, M., Tombari, F., Khan, F. S., and Khan, S. How good is my video lmm? complex video reasoning and robustness evaluation suite for video-lmms. *arXiv preprint arXiv:2405.03690*, 2024.

Kim, C. D., Kim, B., Lee, H., and Kim, G. AudioCaps: Generating captions for audios in the wild. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1011. URL https://aclanthology.org/N19-1011/.

Kou, S., Jin, J., Liu, C., Ma, Y., Jia, J., Chen, Q., Jiang, P., and Deng, Z. Orthus: Autoregressive interleaved image-text generation with modality-specific heads. *arXiv preprint arXiv:2412.00127*, 2024.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

Labs, B. F. Flux: A framework for state-of-the-art image generation. https://github.com/black-forest-labs/flux, 2024.

Lai, J., Zhang, J., Liu, J., Li, J., Lu, X., and Guo, S. Spider: Any-to-many multimodal llm. *arXiv preprint arXiv:2411.09439*, 2024.

Lee, C., Kim, J., and Park, N. Codi: Co-evolving contrastive diffusion models for mixed-type tabular synthesis. In *International Conference on Machine Learning*, pp. 18940–18956. PMLR, 2023a.

Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K., Bishop, C., Hall, E., Carbune, V., Rastogi, A., et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023b.

Lee, S., Kim, S., Park, S., Kim, G., and Seo, M. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 11286–11315, 2024.

Lee Rodgers, J. and Nicewander, W. A. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T., et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*, 2024a.

Li, D., Liu, Y., Wu, H., Wang, Y., Shen, Z., Qu, B., Niu, X., Wang, G., Chen, B., and Li, J. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024b.

Li, H., Zhang, Y., Koto, F., Yang, Y., Zhao, H., Gong, Y., Duan, N., and Baldwin, T. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*, 2023a.

Li, H., Tian, C., Shao, J., Zhu, X., Wang, Z., Zhu, J., Dou, W., Wang, X., Li, H., Lu, L., et al. Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding. *arXiv preprint arXiv:2412.09604*, 2024c.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023b.

Li, J., Pan, K., Ge, Z., Gao, M., Ji, W., Zhang, W., Chua, T.-S., Tang, S., Zhang, H., and Zhuang, Y. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. *arXiv preprint arXiv:2308.04152*, 2023c.

Li, L., Wei, Y., Xie, Z., Yang, X., Song, Y., Wang, P., An, C., Liu, T., Li, S., Lin, B. Y., et al. Vlrewardbench: A challenging benchmark for vision-language generative reward models. *arXiv preprint arXiv:2411.17451*, 2024d.

Li, S., Singh, H., and Grover, A. Instructany2pix: Flexible visual editing via multimodal instruction following. *arXiv preprint arXiv:2312.06738*, 2023d.

Li, S., Kallidromitis, K., Gokul, A., Liao, Z., Kato, Y., Kozuka, K., and Grover, A. Omniflow: Any-to-any generation with multi-modal rectified flows. *arXiv preprint arXiv:2412.01169*, 2024e.

Li, X., Ma, C., Yang, X., and Yang, M.-H. Vidtome: Video token merging for zero-shot video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7486–7495, 2024f.

Li, Y., Zhang, G., Ma, Y., Yuan, R., Zhu, K., Guo, H., Liang, Y., Liu, J., Wang, Z., Yang, J., et al. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*, 2024g.

Li, Y., Zhang, Y., Wang, C., Zhong, Z., Chen, Y., Chu, R., Liu, S., and Jia, J. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024h.

Li, Y., Liu, J., Zhang, T., Chen, S., Li, T., Li, Z., Liu, L., Ming, L., Dong, G., Pan, D., et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025.

Li, Z., Li, H., Shi, Y., Farimani, A. B., Kluger, Y., Yang, L., and Wang, P. Dual diffusion for unified image generation and understanding. *arXiv preprint arXiv:2501.00289*, 2024i.

Liang, Y., He, J., Li, G., Li, P., Klimovskiy, A., Carolan, N., Sun, J., Pont-Tuset, J., Young, S., Yang, F., et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19401–19411, 2024.

Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., and Yuan, L. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Lin, B. Y., Deng, Y., Chandu, K., Brahman, F., Ravichander, A., Pyatkin, V., Dziri, N., Bras, R. L., and Choi, Y. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. *arXiv preprint arXiv:2406.04770*, 2024a.

Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., and Ramanan, D. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024b.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916, 2023a.

Liu, H., Yuan, Y., Liu, X., Mei, X., Kong, Q., Tian, Q., Wang, Y., Wang, W., Wang, Y., and Plumbley, M. D. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024b.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023b.

Lu, J., Clark, C., Zellers, R., Mottaghi, R., and Kembhavi, A. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022.

Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., and Kembhavi, A. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26439–26455, 2024.

Luo, S., Yan, C., Hu, C., and Zhao, H. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:48855–48876, 2023.

Luo, Z., Wu, H., Li, D., Ma, J., Kankanhalli, M., and Li, J. Videoautoarena: An automated arena for evaluating large multimodal models in video analysis through user simulation. *arXiv preprint arXiv:2411.13281*, 2024.

Ma, Y., Ji, J., Ye, K., Lin, W., Wang, Z., Zheng, Y., Zhou, Q., Sun, X., and Ji, R. I2ebench: A comprehensive benchmark for instruction-based image editing. *arXiv preprint arXiv:2408.14180*, 2024a.

Ma, Y., Liu, X., Chen, X., Liu, W., Wu, C., Wu, Z., Pan, Z., Xie, Z., Zhang, H., Zhao, L., et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *arXiv preprint arXiv:2411.07975*, 2024b.

Majumder, N., Hung, C.-Y., Ghosal, D., Hsu, W.-N., Mihalcea, R., and Poria, S. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 564–572, 2024.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Minimax. Minimax video-01. `https://www.minimax.io/news/video-01?utm_source=minimaxi`, 2024.

Mizrahi, D., Bachmann, R., Kar, O. F., Yeo, T., Gao, M., Dehghan, A., and Zamir, A. 4m: Massively multimodal masked modeling. *ArXiv*, abs/2312.06647, 2023. URL `https://api.semanticscholar.org/CorpusID:266162752`.

ML, R. Introducing gen-3 alpha: A new frontier for video generation. `https://runwayml.com/research/introducing-gen-3-alpha`, 2024.

Ni, J., Song, Y., Ghosal, D., Li, B., Zhang, D. J., Yue, X., Xue, F., Zheng, Z., Zhang, K., Shah, M., et al. Mixeval-x: Any-to-any evaluations from real-world data mixtures. *arXiv preprint arXiv:2410.13754*, 2024.

OpenAI. Sora. `https://openai.com/sora/`, 2024.

Qin, C., Yu, N., Xing, C., Zhang, S., Chen, Z., Ermon, S., Fu, Y., Xiong, C., and Xu, R. Gluegen: Plug and play multi-modal encoders for x-to-image generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 23085–23096, 2023.

Qu, L., Zhang, H., Liu, Y., Wang, X., Jiang, Y., Gao, Y., Ye, H., Du, D. K., Yuan, Z., and Wu, X. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *arXiv preprint arXiv:2412.03069*, 2024.

Razavi, A., Van den Oord, A., and Vinyals, O. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Ren, W., Yang, H., Zhang, G., Wei, C., Du, X., Huang, W., and Chen, W. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Ruan, L., Ma, Y., Yang, H., He, H., Liu, B., Fu, J., Yuan, N. J., Jin, Q., and Guo, B. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10219–10228, 2023.

Sakshi, S., Tyagi, U., Kumar, S., Seth, A., Selvakumar, R., Nieto, O., Duraiswami, R., Ghosh, S., and Manocha, D. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024.

Sheffer, R. and Adi, Y. I hear your true colors: Image guided audio generation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Shi, W., Han, X., Zhou, C., Liang, W., Lin, X. V., Zettlemoyer, L., and Yu, L. Llamafusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024.

Sun, G., Yu, W., Tang, C., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., Wang, Y., and Zhang, C. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*, 2024a.

Sun, K., Huang, K., Liu, X., Wu, Y., Xu, Z., Li, Z., and Liu, X. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. *arXiv preprint arXiv:2407.14505*, 2024b.

Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., and Wang, X. Emu: Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.

Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Wang, Y., Rao, Y., Liu, J., Huang, T., and Wang, X. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14398–14409, 2024c.

Sun, W., Tu, R.-C., Liao, J., and Tao, D. Diffusion model-based video editing: A survey. *arXiv preprint arXiv:2407.07111*, 2024d.

Tan, S., Zhuang, S., Montgomery, K., Tang, W. Y., Cuadron, A., Wang, C., Popa, R. A., and Stoica, I. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*, 2024.

Tang, M., Wang, Z., Liu, Z., Rao, F., Li, D., and Li, X. Clip4caption: Clip for video caption. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4858–4862, 2021.

Tang, Z., Yang, Z., Zhu, C., Zeng, M., and Bansal, M. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36:16083–16099, 2023a.

Tang, Z., Yang, Z., Zhu, C., Zeng, M., and Bansal, M. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36:16083–16099, 2023b.

Team, C. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024a.

Team, L., Modi, A., Veerubhotla, A. S., Rysbek, A., Huber, A., Wiltshire, B., Veprek, B., Gillick, D., Kasenberg, D., Ahmed, D., et al. Learnlm: Improving gemini for learning. *arXiv preprint arXiv:2412.16429*, 2024b.

Tschannen, M., Pinto, A. S., and Kolesnikov, A. Jetformer: An autoregressive generative model of raw images and text. *arXiv preprint arXiv:2411.19722*, 2024.

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Wang, H., Wang, Q., Hu, J., Zhang, R., Gao, T., Rong, S., and Dong, H. Global research trends in in-stent neoatherosclerosis: A citespace-based visual analysis. *Frontiers in Cardiovascular Medicine*, 9:1025858, 2022.

Wang, K., Yin, Q., Wang, W., Wu, S., and Wang, L. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.

Wang, X., Zhuang, B., and Wu, Q. Modaverse: Efficiently transforming modalities with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26606–26616, 2024c.

Wang, Y., Guo, W., Huang, R., Huang, J., Wang, Z., You, F., Li, R., and Zhao, Z. Frieren: Efficient video-to-audio generation with rectified flow matching. *arXiv e-prints*, pp. arXiv–2406, 2024d.

Wang, Z., Hu, S., Zhao, S., Lin, X., Juefei-Xu, F., Li, Z., Han, L., Subramanyam, H., Chen, L., Chen, J., et al. Mllm-as-a-judge for image safety without human labeling. *arXiv preprint arXiv:2501.00192*, 2024e.

Wang, Z., Zhu, K., Xu, C., Zhou, W., Liu, J., Zhang, Y., Wang, J., Shi, N., Li, S., Li, Y., et al. Mio: A foundation model on multimodal tokens. *arXiv preprint arXiv:2409.17692*, 2024f.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024a.

Wu, J., Jiang, Y., Ma, C., Liu, Y., Zhao, H., Yuan, Z., Bai, S., and Bai, X. Liquid: Language models are scalable multi-modal generators. *arXiv preprint arXiv:2412.04332*, 2024b.

Wu, S., Fei, H., Qu, L., Ji, W., and Chua, T.-S. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023a.

Wu, T., Yang, G., Li, Z., Zhang, K., Liu, Z., Guibas, L., Lin, D., and Wetzstein, G. Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22227–22238, 2024c.

Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., and Li, H. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023b.

Wu, Y., Zhang, Z., Chen, J., Tang, H., Li, D., Fang, Y., Zhu, L., Xie, E., Yin, H., Yi, L., et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024d.

xAI. Realworldqa. `https://huggingface.co/datasets/xai-org/RealworldQA`, 2024.

Xia, P., Han, S., Qiu, S., Zhou, Y., Wang, Z., Zheng, W., Chen, Z., Cui, C., Ding, M., Li, L., et al. Mmie: Massive multimodal interleaved comprehension benchmark for large vision-language models. *arXiv preprint arXiv:2410.10139*, 2024.

Xie, J., Mao, W., Bai, Z., Zhang, D. J., Wang, W., Lin, K. Q., Gu, Y., Chen, Z., Yang, Z., and Shou, M. Z. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024a.

Xie, R., Du, C., Song, P., and Liu, C. Muse-vl: Modeling unified vlm through semantic discrete encoding. *arXiv preprint arXiv:2411.17762*, 2024b.

Xing, J., Xia, M., Zhang, Y., Chen, H., Yu, W., Liu, H., Liu, G., Wang, X., Shan, Y., and Wong, T.-T. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pp. 399–417. Springer, 2024a.

Xing, Y., He, Y., Tian, Z., Wang, X., and Chen, Q. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7151–7161, 2024b.

Xiong, T., Wang, X., Guo, D., Ye, Q., Fan, H., Gu, Q., Huang, H., and Li, C. Llava-critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712*, 2024a.

Xiong, T., Wang, X., Guo, D., Ye, Q., Fan, H., Gu, Q., Huang, H., and Li, C. Llava-critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712*, 2024b.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

Yang, J., Yin, D., Zhou, Y., Rao, F., Zhai, W., Cao, Y., and Zha, Z.-J. Mmar: Towards lossless multi-modal auto-regressive probabilistic modeling. *arXiv preprint arXiv:2410.10798*, 2024b.

Yang, Q., Xu, J., Liu, W., Chu, Y., Jiang, Z., Zhou, X., Leng, Y., Lv, Y., Zhao, Z., Zhou, C., et al. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*, 2024c.

Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024d.

Yariv, G., Gat, I., Benaim, S., Wolf, L., Schwartz, I., and Adi, Y. Diverse and aligned audio-to-video generation via text-to-video model adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6639–6647, 2024.

Yasunaga, M., Zettlemoyer, L., and Ghazvininejad, M. Multimodal rewardbench: Holistic evaluation of reward models for vision language models. 2025. URL `https://api.semanticscholar.org/CorpusID:276482127`.

Ye, J., Wang, Y., Huang, Y., Chen, D., Zhang, Q., Moniz, N., Gao, T., Geyer, W., Huang, C., Chen, P.-Y., et al. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*, 2024.

Yu, T., Zhang, H., Yao, Y., Dang, Y., Chen, D., Lu, X., Cui, G., He, T., Liu, Z., Chua, T.-S., et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024.

Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., and Wang, L. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Zhang, G., Du, X., Chen, B., Liang, Y., Luo, T., Zheng, T., Zhu, K., Cheng, Y., Xu, C., Guo, S., et al. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.11944*, 2024a.

Zhang, H., Li, X., and Bing, L. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023a.

Zhang, K., Mo, L., Chen, W., Sun, H., and Su, Y. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023b.

Zhang, L., Mo, S., Zhang, Y., and Morgado, P. Audio-synchronized visual animation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024b.

Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., and Tian, Q. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023c.

Zhang, Y.-F., Yu, T., Tian, H., Fu, C., Li, P., Zeng, J., Xie, W., Shi, Y., Zhang, H., Wu, J., et al. Mm-rlhf: The next step forward in multimodal llm alignment. *arXiv preprint arXiv:2502.10391*, 2025.

Zhao, C., Song, Y., Wang, W., Feng, H., Ding, E., Sun, Y., Xiao, X., and Wang, J. Monoformer: One transformer for both diffusion and autoregression. *arXiv preprint arXiv:2409.16280*, 2024.

Zhao, Y., Xie, L., Zhang, H., Gan, G., Long, Y., Hu, Z., Hu, T., Chen, W., Li, C., Song, J., et al. Mmvu: Measuring expert-level multi-discipline video understanding. *arXiv preprint arXiv:2501.12380*, 2025.

Zhen, L., Hu, P., Wang, X., and Peng, D. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10394–10403, 2019.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., and Hou, L. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Zhou, P., Peng, X., Song, J., Li, C., Xu, Z., Yang, Y., Guo, Z., Zhang, H., Lin, Y., He, Y., et al. Gate opening: A comprehensive benchmark for judging open-ended interleaved image-text generation. *arXiv preprint arXiv:2411.18499*, 2024.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Zhuge, M., Zhao, C., Ashley, D., Wang, W., Khizbullin, D., Xiong, Y., Liu, Z., Chang, E., Krishnamoorthi, R., Tian, Y., et al. Agent-as-a-judge: Evaluate agents with agents. *arXiv preprint arXiv:2410.10934*, 2024.

Zhuo, L., Wang, Z., Wang, B., Liao, Y., Bao, C., Peng, S., Han, S., Zhang, A., Fang, F., and Liu, S. Video background music generation: Dataset, method and evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15637–15647, 2023.

Ziv, A., Gat, I., Lan, G. L., Remez, T., Kreuk, F., Défossez, A., Copet, J., Synnaeve, G., and Adi, Y. Masked audio generation using a single non-autoregressive transformer. *arXiv preprint arXiv:2401.04577*, 2024.

A Full Related Works

Multimodal Understanding (MMU). MMU involves integrating and processing information from multiple modalities—such as text (Li et al., 2023a,a), images (Yue et al., 2024; Zhang et al., 2024a), video (Zhao et al., 2025; Hu et al., 2025), audio (Sakshi et al., 2024)—achieve significant development since the advent of Multimodal Large Language Models (MLLMs). Through late fusion of modality features with pretrained LLMs and modality instruction tuning

(Liu et al., 2023a), MLLMs gain incredibly advanced understanding capabilities in images (Zhu et al., 2023; Li et al., 2023b), videos (Lin et al., 2023; Zhang et al., 2023a), audio (Chu et al., 2023, 2024), and even interleaved content (Chen et al., 2025), transforming traditional MMG tasks such as Visual Question Answering (VQA) (Antol et al., 2015; Krishna et al., 2017), Multimodal Captioning (Bai & An, 2018; Tang et al., 2021), and cross-modal retrieval (Wang et al., 2016; Zhen et al., 2019) in a unified manner.

Recent benchmarks like MMMU (Yue et al., 2024), Video-MMMU (Hu et al., 2025), and MMAU (Sakshi et al., 2024) have been developed to rigorously evaluate understanding and reasoning in *specific modality* of MLLMs. Any-to-Any benchmarks also emerge to provide comprehensive assessment for current well-rounded models capable of understanding in many modalities (Chen et al., 2024c; Li et al., 2024g; Ni et al., 2024; Hong et al., 2025). However, these benchmarks evaluate MMG mainly through Multiple-Choice QA, undermining the reliable assessment of real-world open-ended queries (xAI, 2024).

Multimodal Generation (MMG). MMG involves generating content in one modality based on input from another or mixture (Ni et al., 2024; Chen et al., 2025), such as text-to-image (Ghosh et al., 2023), text-to-video (Sun et al., 2024b; Huang et al., 2024), image-to-video (Sun et al., 2024d; Fan et al., 2023), video-to-music (Kang et al., 2024; Zhuo et al., 2023) and other modality transformations (Doh et al., 2023). Early approaches primarily focused on building specific framework for each task (Betker et al., 2024; Yang et al., 2024d), which were later unified by Auto-Regressive (AR) models such as Show-o (Xie et al., 2024a), Emu-3 (Wang et al., 2024b), and Unified-IO (Lu et al., 2022, 2024) which enabling generate various modalities in a more coherent and complex manner.

However, the open-ended nature of MMG tasks makes evaluation challenging, as traditional ground-truth-based metrics fail to capture the diversity and quality of generated content (Ni et al., 2024). Human-oriented evaluations, such as those on crowdsourcing platforms like GenAI-Arena (Jiang et al., 2025) and other benchmarks (Liang et al., 2024; Huang et al., 2024), have become a common approach to assess MMG tasks. Despite their utility, these platforms often suffer from issues like insufficient votes, leading to instability in rankings (Chen et al., 2024a). Our work advances the uniform incorporation of MLLM-as-a-Judge across various modalities by implementing checklist-of-thought reasoning to achieve more unbiased, reliable, and reproducible evaluations of MMG tasks.

Multimodal LLM-as-a-Judge. Originated from Natural Language Generation (NLG) domain, LLM-as-a-Judge (Zheng et al., 2023) have extended to multimodal domains serving as evaluation metrics in general QA (Chen et al., 2024a; Xiong et al., 2024b; Lee et al., 2024; Tan et al., 2024), image generation (Chen et al., 2024g; Lin et al., 2024b), video generation (Luo et al., 2024), 3D synthesis (Wu et al., 2024c), SWE tasks (Zhuge et al., 2024), and interleaved generation (Chen et al., 2025; Zhou et al., 2024). Another line leverage pretrained MLLMs serving as reward models (Yasunaga et al., 2025; Li et al., 2024d; Yu et al., 2024; Zhang et al., 2025; Wang et al., 2024e) in aligning other MLLMs for advanced performance. MLLM-as-a-Judge (Chen et al., 2024a) takes the first step in systematically quantifying MLLMs' performance as judges and assessing potential problems such as bias and hallucinations in Vision-Language Understanding tasks. Our work extends this systematic assessing framework for MLLM-as-a-Judge to 15 Any-to-Any MMU and MMG tasks, with carefully selected samples for open-ended queries, providing in-depth analysis of potential challenges when applying MLLM-as-a-Judge across broader modalities and more general use cases.

Any-to-Any Unified Models. We term models that can take and generate with various modalities as ***Unified Models***, which unifies different modalities into the paradigm of next token prediction with an auto-regressive structure, leveraging a powerful pretrained LLM backbone. By tokenizing continuous contents into discrete tokens using different tokenizers (Van Den Oord et al., 2017; Razavi et al., 2019), researchers have started to explore simultaneously visual understanding and generation with a single backbone (Li et al., 2024i; Shi et al., 2024; Li et al., 2024c; Wu et al., 2024b; Qu et al., 2024; Li et al., 2024e; Ma et al., 2024b; Xie et al., 2024b; Tschannen et al., 2024; Kou et al., 2024; Lai et al., 2024; Wu et al., 2024a,d; Zhao et al., 2024; Wang et al., 2024f; Yang et al., 2024b; Chern et al., 2024; Li et al., 2024h; Team, 2024; Sun et al., 2023, 2024c). Other pioneer works extend this boundary into other modalities such as video (Wang et al., 2024b), audio (Tang et al., 2023a), conditions (Mizrahi et al., 2023; Bachmann et al., 2024), and 3D assets generation (Chen et al., 2024e) in an Auto-Regressive manner.

B Benchmark Details

B.1 Benchmark Construction

We sample open-ended queries from previous benchmarks (Table 5) randomly and conduct manually filtering for their quality.

Table 5: Benchmark Sources.

TASK	SOURCE	NUMBER
TEXT-TO-TEXT	WILDBENCH (LIN ET AL., 2024A)	50
	IFEVAL (ZHOU ET AL., 2023)	50
IMAGE-TO-TEXT	DEMONBENCH (LI ET AL., 2023C)	25
	MMVET (YU ET AL., 2023)	25
	TOUCHSTONE (BAI ET AL., 2023)	25
	VISITBENCH (BITTON ET AL., 2023)	25
VIDEO-TO-TEXT	CVRR (KHATTAK ET AL., 2024)	25
	AUTOEVAL (CHEN ET AL., 2024D)	25
	TEMPORALBENCH (CAI ET AL., 2024)	25
	VIDEOMME (FU ET AL., 2024)	25
AUDIO-TO-TEXT	LTU (GONG ET AL., 2023)	26
	VOICEBENCH (CHEN ET AL., 2024F)	37
	AIRBENCH (YANG ET AL., 2024C)	37
A+V-TO-TEXT	VALOR_AVQA	100
TEXT-TO-IMAGE	HPSV2 (WU ET AL., 2023B)	50
	T2ICOMPBENCH (HUANG ET AL., 2023)	50
TEXT-TO-VIDEO	VBENCH (HUANG ET AL., 2024)	50
	T2VCOMPBENCH (SUN ET AL., 2024B)	50
TEXT-TO-AUDIO	AUDIOCAPS (KIM ET AL., 2019)	50
	CLOTHO (DROSSOS ET AL., 2020)	50
IMAGE EDIT	I^2EBENCH (MA ET AL., 2024A)	100
IMAGE-TO-VIDEO	CONSISTI2V (REN ET AL., 2024)	100
IMAGE-TO-AUDIO	IMAGEHEAR (SHEFFER & ADI, 2023)	100
VIDEO EDIT	V2VBENCH (SUN ET AL., 2024D)	100
VIDEO-TO-AUDIO	VGGSOUND (CHEN ET AL., 2020)	100
AUDIO EDIT	AUDIOEDITOR (JIA ET AL., 2024)	100
AUDIO-TO-VIDEO	AVSYNC15 (ZHANG ET AL., 2024B)	100

B.2 Safety Checking

In this section, we provide a detailed analysis of trustworthiness problems in TASKANYTHING and JUDGEANYTHING, focusing on NSFW content in text and multimodal content separately.

NSFW Image Filtering. Figure 7 illustrates the proportion of unsafe and safe images across all categories based on the model's judgments. Out of all the images used, each sample is classified as *Safety*.

NSFW Filtering for Other Modalities. We conduct rigorous NSFW checks during dataset construction. Human annotators manually review all videos and audio clips to ensure they met NSFW safety standards. Since these samples are primarily sourced from established benchmarks, all are classified as *Safe* by the annotators.

B.3 Human Annotation Details

The annotation is conducted by 10 authors of this paper and 2 volunteers independently. As acknowledged, the diversity of annotators plays a crucial role in reducing bias and enhancing the reliability of the benchmark. These annotators have knowledge in this domain, with different genders, ages, and educational backgrounds. To ensure the annotators can proficiently mark the data, we provide them with detailed tutorials, teaching them how to evaluate model responses more objectively, detailed as follows:

- **Checklist Filter and Refinement.**
- **Human Annotation in *Score Evaluation* and *Pair Comparison*.**

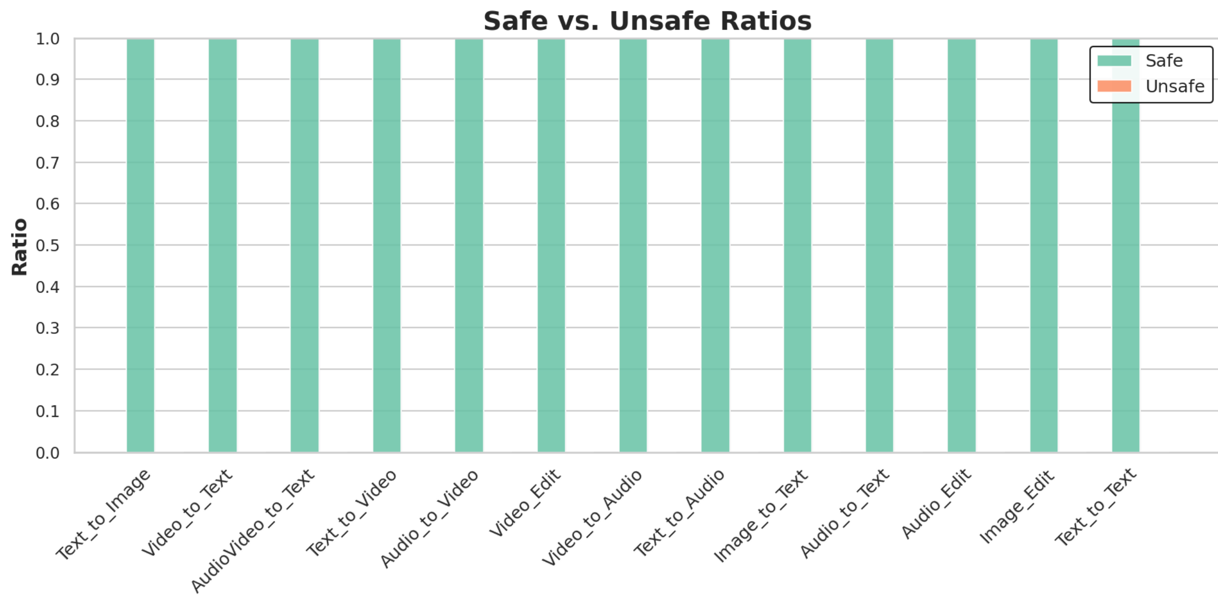


Figure 7: Safe *v.s.* Unsafe content ratios across multimodal tasks.

Table 6: Percentage of choices for *Non-Text Driven Generation* tasks under *Pair Comparison* setting. First refers to the judging model selecting the first response; Second refers to the judging model selecting the second response; Tie refers to the judging model selecting a tie. *Rubric* and *Checklist* settings are the average of six rubrics.

	Settings	**I→I**			**I→V**			**I→A**			**V→V**			**V→A**			**A→V**			**A→A**		
		First	Tie	Second	First	Tie	Second	First	Tie	Second	First	Tie	Second	First	Tie	Second	First	Tie	Second	First	Tie	Second
Judging Models																						
GPT-4o	*Overall*	32.6	10.0	57.2	55.7	6.5	37.8	67.7	1.0	31.6	18.6	0.5	81.9	56.7	6.0	37.3	47.2	8.5	44.3	46.7	6.0	47.3
	Rubric	0.0	100.0	0.0	0.5	99.4	0.5	9.7	85.4	4.9	0.5	99.8	0.5	36.1	46.0	18.0	42.9	54.8	3.5	5.2	91.4	3.4
	Checklist	0.0	100.0	0.0	1.0	99.0	1.5	2.6	96.2	1.5	100.0	0.0	0.0	28.2	53.1	18.8	30.5	60.6	10.7	2.8	93.8	3.5
Gemini-1.5-Pro	*Overall*	33.6	5.5	60.9	12.0	1.0	87.0	23.5	0.0	76.5	6.0	0.0	94.0	7.0	0.0	93.0	41.0	0.0	59.0	20.5	2.0	77.5
	Rubric	20.5	42.3	37.2	8.1	0.6	57.2	17.1	0.0	49.6	3.9	0.0	62.8	4.9	0.0	61.8	19.6	12.0	37.1	11.3	0.0	55.4
	Checklist	15.6	45.4	39.0	8.9	1.3	56.5	19.4	0.0	47.2	4.1	0.0	62.6	5.3	0.0	61.4	21.7	12.6	36.6	12.3	0.0	54.4
Gemini-2.0-Flash	*Overall*	32.6	6.0	61.4	40.6	24.6	34.8	68.7	16.1	15.2	25.6	0.0	74.4	47.1	31.6	21.3	71.7	1.0	27.3	37.1	39.6	23.3
	Rubric	19.9	41.4	38.7	15.0	68.4	16.7	27.0	65.0	9.6	15.7	44.8	55.3	18.2	71.7	10.2	16.3	77.2	7.8	18.0	73.1	11.9
	Checklist	21.1	10.8	34.9	13.1	36.4	17.2	39.0	11.3	16.3	12.8	2.9	77.1	21.8	26.2	18.7	37.0	14.0	15.6	19.1	29.2	18.3
Human Evaluators																						
Human Evaluators	*Overall*	23.6	32.6	43.8	36.6	16.1	47.3	43.6	8.0	48.4	10.5	2.0	87.5	32.1	27.6	40.3	41.6	27.6	30.8	44.6	6.5	48.9
	Rubric	21.2	44.2	34.6	25.4	31.8	42.9	43.2	9.8	47.0	11.1	2.7	86.7	33.7	27.3	39.1	45.8	20.2	34.0	40.4	5.0	54.6
	Checklist	21.2	44.2	34.6	25.4	31.8	42.9	43.2	9.8	47.0	11.1	2.7	86.7	33.7	27.3	39.1	45.8	20.2	34.0	40.4	5.0	54.6

B.4 Copyright

Given that we sample queries from previous well-established benchmarks to form TaskAnything and collect *state-of-the-art* models' responses to curate JudgeAnything, we will release our code, benchmark, and dataset under the `Creative Commons 4.0` license rather than the `Apache` license, to maintain compatibility with the original licenses of these benchmarks and models.

B.5 ELO Rating System Details

To update the ELO ratings after each match, we use the following formulas:

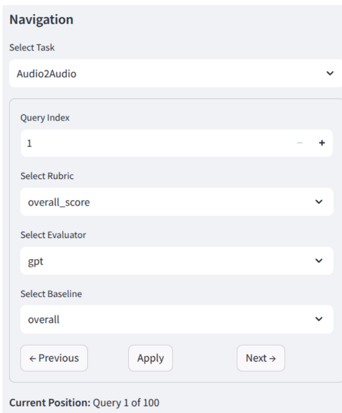
$$P_A = \frac{1}{1 + 10^{\frac{R_B - R_A}{400}}} \tag{1}$$

$$P_B = \frac{1}{1 + 10^{\frac{R_A - R_B}{400}}} \tag{2}$$

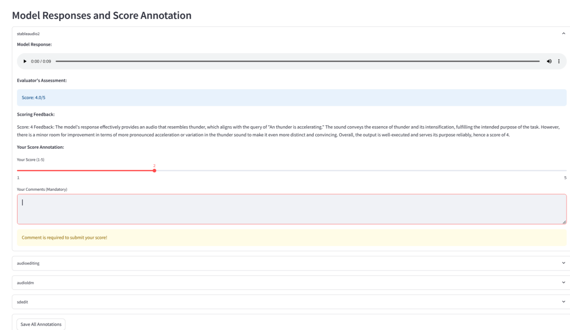
where P_A is the expected probability of model 1 winning. P_B is the expected probability of model 2 winning. R_A and R_B are the current ELO ratings of model 1 and model 2, respectively.

Annotation screenshot and guideline

Annotation Usage: streamlit run annotation_score/pair.py
For operations within the box, you need to click "Apply" for it to take effect. However, after clicking "Apply", it will switch to the next one, preventing the Streamlit rollback issue. You can quickly switch between the same Task and Rubric using "Next" and "Previous" without needing to click "Apply".



(a): Annotation interface's navigation bar



(b): *Score Evaluation* annotation interface's model response selection

You just need to select the corresponding *Score Evaluation* and *Pair Comparison* based on your understanding and judgment. It will be saved automatically. Please carefully check your annotation file to avoid major mistakes. You should be able to recover from the backup. The right figure is an example of the answers from each model, which you can click to open.

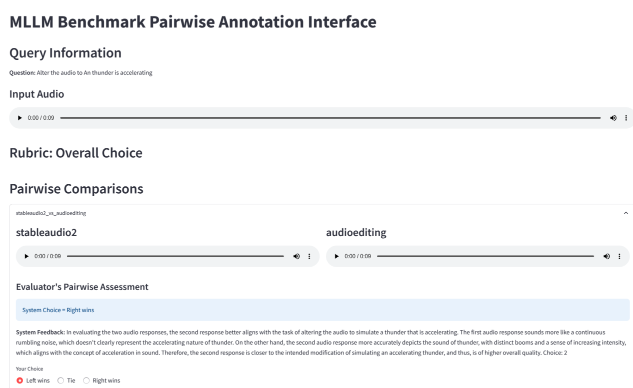


Figure 8: Human annotation instruction

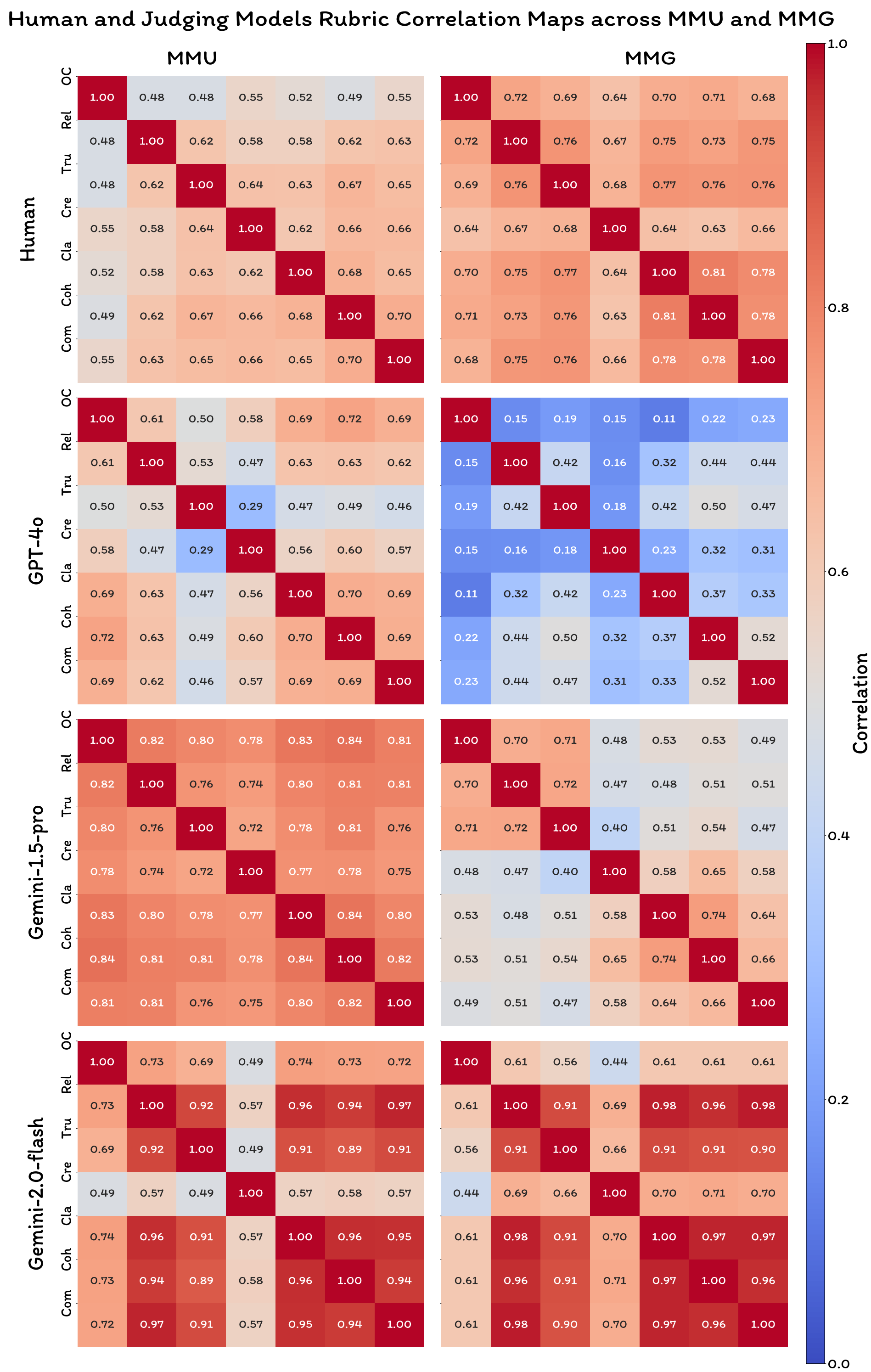


Figure 9: Correlation Maps for different evaluators. Left column is the correlation map for Multimodal Understanding, Right column is the correlation map for MMG; Each heatmaps rubric sequence is: OC (Overall Choice), Rel (Relevance), Tru (Trustworthiness), Cre (Creativity & Novelty), Cla (Clarity), Coh (Coherence), Com (Completeness)

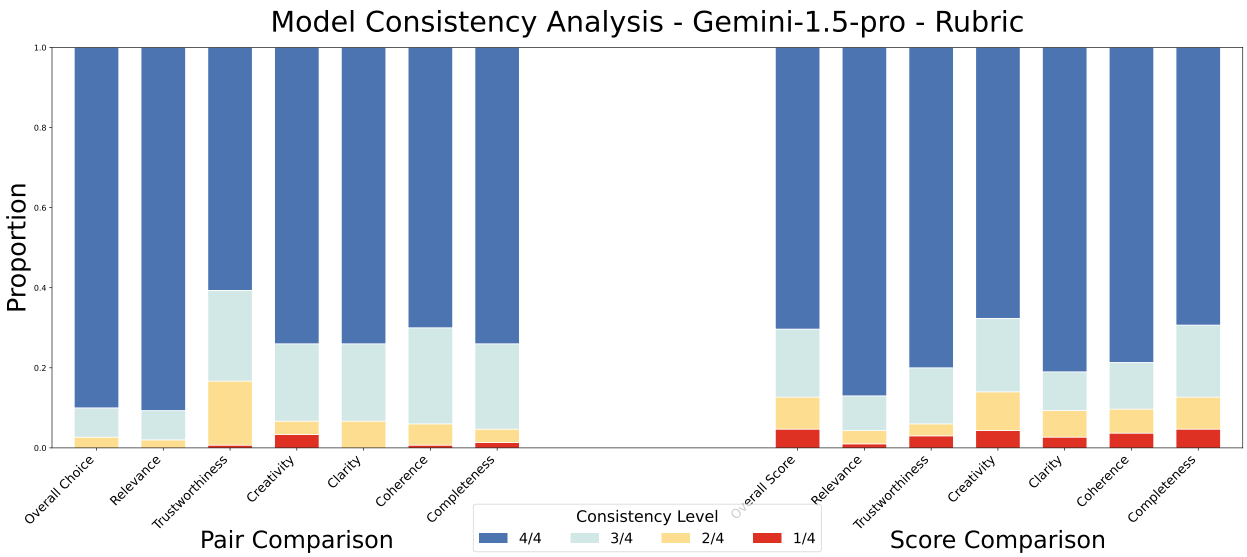


Figure 10: The Gemini-1.5-pro consistency data between rubrics. Overall Choice and Overall Score are generated by *Overall* baseline, others are generated by *Rubric* baseline

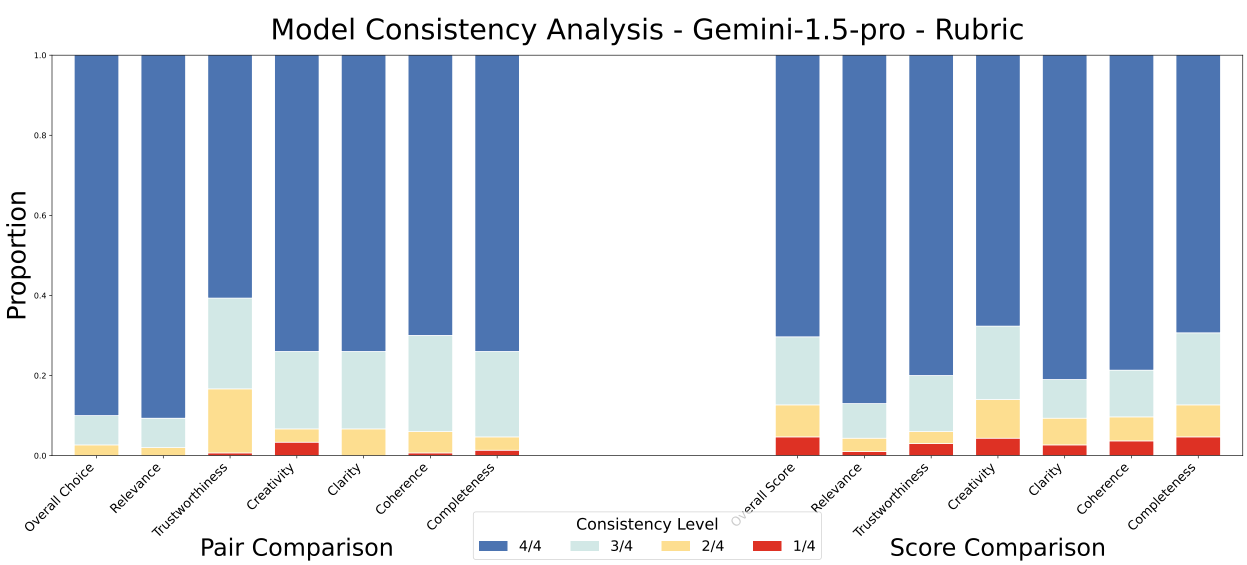


Figure 11: The Gemini-2.0-flash consistency data between rubrics. Overall Choice and Overall Score are generated by *Overall* baseline, others are generated by *Rubric* baseline

The ratings are updated after each comparison as follows:

$$R'_A = R_A + K \times (S_A - P_A) \tag{3}$$

$$R'_B = R_B + K \times (S_B - P_B) \tag{4}$$

where R'_A and R'_B are the updated ELO ratings of model 1 and model 2. S_A and S_B represent the actual outcomes: $S_A = 1$ for a win, $S_A = 0.5$ for a tie, and $S_A = 0$ for a loss (similarly for S_B). K is a constant that determines the magnitude of rating changes, which is set to 32 in our arena.

C Experiment Setup Details

Baseline *Rubric*

- **Overall** provides a holistic assessment of the generated output by evaluating its general effectiveness, excellence, and suitability for the intended purpose.
- **Relevance** measures how closely and directly the output addresses the given prompt or input. A relevant response directly responds to the instructions, stays on-topic throughout, and provides information or content that is pertinent to the requested task.
- **Trustworthiness** evaluates the output's reliability, accuracy, and safety. It involves checking whether the content is factually correct, well-sourced, compliant with guidelines, and free from harmful or misleading information.
- **Creativity and Novelty** refers to the originality or freshness of the content, introducing something genuinely new or less commonly encountered. it encompasses the imagination and inventiveness behind the output, blending originality with purpose, style, insight, or aesthetic appeal.
- **Clarity** assesses how easily the content can be understood. It involves clear expression, well-organized ideas, and the absence of ambiguity or confusion.
- **Coherence** evaluates the logical flow and consistency of the content. It ensures that ideas are connected logically and that the narrative progresses smoothly without abrupt jumps or disjointed sections.
- **Completeness** measures whether the output fully addresses all aspects of the prompt or task. It checks for the inclusion of all necessary components, details, and depth required to meet the objectives.

C.1 Models for Multimodal Understanding

See Table 7.

C.2 Models for Multimodal Generation

See Table 8.

C.3 Models for Judge

GPT-4o Limitations and Integration GPT-4o cannot process both audio and visual inputs simultaneously. To address this, we integrate two versions of GPT-4o—GPT-4o and GPT-4o-audio-preview. For audio-visual cross-modal tasks, the input modality is captioned into text, ensuring the response is accessible in either visual or auditory form. **Open-Source Models and Limitations** In the realm of open-source multimodal understanding models, we have deployed several prominent architectures, including Baichuan-Omni-1.5 (Li et al., 2025) and VideoLlama2 (Cheng et al., 2024). These models exhibit significant advancements in handling multimodal inputs. However, despite their capabilities, they have notable limitations in judge areas. These limitations in handling complex, cross-modality inputs, such as interleaved audio-image data or multiple simultaneous media inputs, along with restricted capacity for long-context processing, explain why we did not include open-source models as our judge.

C.4 Models for Arena

See Table 9.

Table 7: Overview of MMU models used in our study

Task	Model	Size	Release Date
Text2Text	GPT-4o (Hurst et al., 2024)	N/A	May 2024
	Claude 3.5 Sonnet (Anthropic, 2023)	N/A	Jun 2024
	Qwen2.5-72b (Yang et al., 2024a)	72B	Dec 2024
	llama3-70b (Dubey et al., 2024)	70B	Jul 2024
Image2Text	Qwen2-VL-72b (Wang et al., 2024a)	72B	Dec 2024
	Phi3.5V-Instruct (Abdin et al., 2024)	4.15B	Apr 2024
	Claude 3.5 Sonnet (Anthropic, 2023)	N/A	Jun 2024
	GPT-4o (Hurst et al., 2024)	N/A	May 2024
Video2Text	Qwen2-VL-72b (Wang et al., 2024a)	72B	Dec 2024
	Aria (Li et al., 2024b)	3.9B	Oct 2024
	Gemini-1.5-Pro (Team et al., 2024a)	N/A	Sep 2024
	GPT-4o (Hurst et al., 2024)	N/A	May 2024
Audio2Text	Gemini-1.5-Pro (Team et al., 2024a)	N/A	Sep 2024
	Qwen2-Audio-7B-Instruct (Chu et al., 2024)	7B	Jul 2024
	Gama (Ghosh et al., 2024)	7B	Jun 2024
	Salmonn-13B (Sun et al., 2024a)	13B	Jun 2024
AudioVideo2Text	Salmonn-13B (Sun et al., 2024a)	13B	Jun 2024
	VideoLLaMA 2 (Cheng et al., 2024)	7B	Jun 2024
	Gemini-1.5-Pro (Team et al., 2024a)	N/A	Sep 2024
	Unified-IO 2 (Lu et al., 2024)	7B	Dec 2023

D Case Study

See Figures 15 and 15 for *Checklist* influence on evaluation. See Figures 17, 18 and 19 for detailed case studies. See Figures **??** and **??** for text-to-text examples.

Table 8: Overview of MMG models used in our study

Task	Model	Size	Release Date
Text2Image	FLUX.1 [dev] (Labs, 2024)	12B	Jul 2024
	Stable Diffusion 3.5-Large (Esser et al., 2024)	8.1B	Oct 2024
	Recraft V3 (AI, 2024)	N/A	Oct 2024
	Dalle3 (Betker et al., 2024)	N/A	Sep 2024
Text2Audio	AudioLDM2-Large (Liu et al., 2024b)	1.5B	May 2024
	Stable Audio Open 1.0 (Evans et al., 2024)	1.21B	Jul 2024
	Tango2 (Majumder et al., 2024)	866M	Jul 2024
	MAGNeT-Medium (Ziv et al., 2024)	1.5B	Jan 2024
Text2Video	VideoCrafter 2 (Chen et al., 2024b)	N/A	Jan 2024
	MiniMax-Video-01 (Minimax, 2024)	N/A	Aug 2024
	CogVideoX1.5 (Yang et al., 2024d)	5B	Aug 2024
	Sora (OpenAI, 2024)	N/A	Dec 2024
Image2Video	CogVideoX1.5 (Yang et al., 2024d)	5B	Aug 2024
	Sora (OpenAI, 2024)	N/A	Dec 2024
	SVD-XT-1.0 (Blattmann et al., 2023)	N/A	Nov 2023
	DynamiCrafter (Xing et al., 2024a)	N/A	Oct 2023
Audio2Video	MM-Diffusion (Ruan et al., 2023)	115.13M	Dec 2022
	GlueGen (Qin et al., 2023)	51M	Mar 2023
	TempoTokens (Yariv et al., 2024)	35M	Sep 2023
	Codi (Tang et al., 2023b)	N/A	May 2023
Video2Audio	Diff-Foley (Luo et al., 2023)	859M	Jun 2023
	Frieren-V2A (Wang et al., 2024d)	421.1M	Jun 2024
	SpecVQGAN (Iashin & Rahtu, 2021)	547.8M	Oct 2021
	Seeing and Hearing (Xing et al., 2024b)	N/A	Feb 2024
Image2Audio	Im2Wav (Sheffer & Adi, 2023)	N/A	Nov 2022
	V2A-Mapper (Wang et al., 2022)	35.45M	Aug 2023
	SpecVQGAN (Iashin & Rahtu, 2021)	547.8M	Oct 2021
	Codi (Tang et al., 2023b)	N/A	May 2023
Image2Image	InstructAny2Pix (Li et al., 2023d)	7B	Dec 2023
	MagicBrush (Zhang et al., 2023b)	N/A	Jun 2023
	MGIE (Fu et al., 2023)	8B	Sep 2023
	InstructPix2Pix (Brooks et al., 2022)	1B	Nov 2022
Audio2Audio	Audio Editing (Jia et al., 2024)	N/A	Feb 2024
	AudioLDM2 (Liu et al., 2024b)	N/A	Aug 2023
	StableAudio 2.0 (Audio, 2024)	N/A	Apr 2024
	SDEdit (Meng et al., 2021)	N/A	Aug 2021
Video2Video	ControlVideo (Zhang et al., 2023c)	N/A	May 2023
	VidToMe (Li et al., 2024f)	N/A	Dec 2023
	Sora (OpenAI, 2024)	N/A	Dec 2024
	Gen-3 Alpha (ML, 2024)	N/A	Jun 2024

System Instruction: You are a loyal judge, your task is to score the performance of the model's response on the given task. You will be given a task, including the input and the model's response. The scoring rule will also be given, you need to score the model's response with your careful consideration. If the judge task require multi-modal inputs, you should use your visual and auditory senses to judge. You should entirely understand, see or hear the task and the model's response, base on the given information, you should think of your scoring reasons in each rubric's "comment" step by step first, and then you are required to give scores for each rubric in each rubric's "score" part base on the scoring rule. Finally, You are required to give an overall score base on the previous results and the overall scoring rule. If the checklists are given, you should use it to assist your scoring process.
***Overall* prompt:** You are going to score the overall quality of the model's performance on the given task. Overall Quality Definition: **Overall Quality** provides a holistic assessment of the generated output by evaluating its general effectiveness, excellence, and suitability for the intended purpose. It reflects the cumulative performance of the output across various dimensions without delving into specific aspects, allowing for a comprehensive and integrated evaluation. Scoring Rule: 1: The output fails to meet basic expectations. It is largely ineffective, significantly flawed, and does not serve its intended purpose. 2: The output meets minimal standards but has considerable deficiencies. It partially serves its purpose but requires substantial improvement. 3: The output adequately meets the basic requirements. It functions as intended but lacks distinction and contains some areas needing enhancement. 4: The output effectively meets the expectations with minor areas for improvement. It is well-executed and serves its purpose reliably. 5: The output surpasses expectations, demonstrating outstanding effectiveness, excellence, and suitability. It is exemplary in fulfilling its intended purpose.
***Rubric* relevance prompt:** You are going to score the relevance of the model's performance on the given task. "Relevance" measures how closely and directly the output addresses the given prompt or input. A relevant response directly responds to the instructions, stays on-topic throughout, and provides information or content that is pertinent to the requested task. Scoring Rule: 1: Largely off-topic or irrelevant; fails to address the prompt. 2: Minimally relevant; addresses the prompt superficially with significant deviations. 3: Moderately relevant; addresses the prompt but may include some unrelated content. 4: Highly relevant; directly addresses the prompt with minor deviations. 5: Perfectly relevant; fully aligns with and directly responds to the prompt without any deviations.
***Rubric* trustworthiness prompt:** You are going to score the trustworthiness of the model's performance on the given task. "Trustworthiness" evaluates the output's reliability, accuracy, and safety. It involves checking whether the content is factually correct, well-sourced, compliant with guidelines, and free from harmful or misleading information. Scoring Rule: 1: Highly unreliable; contains numerous factual errors or harmful content. 2: Minimally trustworthy; several inaccuracies or potential issues present. 3: Moderately trustworthy; generally accurate with some minor errors. 4: Highly trustworthy; accurate and reliable with negligible errors. 5: Completely trustworthy; flawless accuracy, fully compliant, and free from any misleading or harmful content.
***Rubric* creativity prompt:** You are going to score the creativity of the model's performance on the given task. Novelty refers to the originality or freshness of the content, introducing something genuinely new or less commonly encountered. Creativity encompasses the imagination and inventiveness behind the output, blending originality with purpose, style, insight, or aesthetic appeal. Scoring Rule: 1: Minimal creativity; very generic or repetitive content. 2: Slightly creative; some original elements but largely conventional. 3: Moderately creative; a balance of original and standard elements. 4: Highly creative; introduces original ideas and inventive approaches. 5: Exceptionally creative and novel; highly original, imaginative, and innovative.
***Rubric* clarity prompt:** You are going to score the clarity of the model's performance on the given task. "Clarity" assesses how easily the content can be understood. It involves clear expression, well-organized ideas, and the absence of ambiguity or confusion. Scoring Rule: 1: Incomprehensible; ideas are not conveyed clearly. 2: Poor clarity; frequent ambiguities or unclear expressions. 3: Adequate clarity; generally understandable with some minor ambiguities. 4: Clear and mostly easy to understand; minor issues do not impede comprehension. 5: Crystal-clear expression; exemplary articulation with no ambiguities.
***Rubric* coherence prompt:** You are going to score the coherence of the model's performance on the given task. "Coherence" evaluates the logical flow and consistency of the content. It ensures that ideas are connected logically and that the narrative progresses smoothly without abrupt jumps or disjointed sections. Scoring Rule: 1: Disjointed; lacks logical flow and consistency. 2: Poor coherence; frequent logical gaps or inconsistencies. 3: Moderate coherence; some logical flow with occasional inconsistencies. 4: Highly coherent; logical flow with minor inconsistencies. 5: Perfectly cohesive; ideas flow seamlessly and logically.
***Rubric* completeness prompt:** You are going to score the completeness of the model's performance on the given task. "Completeness" measures whether the output fully addresses all aspects of the prompt or task. It checks for the inclusion of all necessary components, details, and depth required to meet the objectives. Scoring Rule: 1: Severely incomplete; missing key components. 2: Minimally complete; several important elements missing. 3: Moderately complete; covers most key elements with some omissions. 4: Highly complete; fully addresses all key elements with minor omissions. 5: Completely complete; all aspects are addressed comprehensively with exceptional detail.

Figure 12: Prompt - *Score Evaluation* prompts.

System Instruction: You are a loyal judge, your task is to choose the better one from two models' responses on the given task. You will be given a task, including the input and the two models' responses. The pairing rule will also be given, you need to choose with your careful consideration. If the judge task require multi-modal inputs, you should use your visual and auditory senses to judge. If the response model misunderstand the task and response in wrong modality, you should take into consideration. You should entirely understand, see or hear the task and the model's response, base on the given information, you should think of your choosing reasons in the each rubric's "comment" step by step first, and then you are required to give a choice in "choice" base on the rule. If the checklists are given, you should use it to assist your choosing process. Choosing Rule: Reasoning in detail before you determine the choice, then give your choice from [0,1,2], 0 means the first response is better, 1 means the two responses are equally good, 2 means the second response is better.
***Overall* prompt:** You are going to choose base on the overall quality of the model's performance on the given task. Overall Quality Definition: **Overall Quality** provides a holistic assessment of the generated output by evaluating its general effectiveness, excellence, and suitability for the intended purpose. It reflects the cumulative performance of the output across various dimensions without delving into specific aspects, allowing for a comprehensive and integrated evaluation.
***Rubric* relevance prompt:** You are going to choose base on the relevance of the model's performance on the given task. "Relevance" measures how closely and directly the output addresses the given prompt or input. A relevant response directly responds to the instructions, stays on-topic throughout, and provides information or content that is pertinent to the requested task.
***Rubric* trustworthiness prompt:** You are going to choose base on the trustworthiness of the model's performance on the given task. "Trustworthiness" evaluates the output's reliability, accuracy, and safety. It involves checking whether the content is factually correct, well-sourced, compliant with guidelines, and free from harmful or misleading information.
***Rubric* creativity prompt:** You are going to choose base on the creativity of the model's performance on the given task. Novelty refers to the originality or freshness of the content, introducing something genuinely new or less commonly encountered. Creativity encompasses the imagination and inventiveness behind the output, blending originality with purpose, style, insight, or aesthetic appeal.
***Rubric* clarity prompt:** You are going to choose base on the clarity of the model's performance on the given task. "Clarity" assesses how easily the content can be understood. It involves clear expression, well-organized ideas, and the absence of ambiguity or confusion.
***Rubric* coherence prompt:** You are going to choose base on the coherence of the model's performance on the given task. "Coherence" evaluates the logical flow and consistency of the content. It ensures that ideas are connected logically and that the narrative progresses smoothly without abrupt jumps or disjointed sections.
***Rubric* completeness prompt:** You are going to choose base on the completeness of the model's performance on the given task. "Completeness" measures whether the output fully addresses all aspects of the prompt or task. It checks for the inclusion of all necessary components, details, and depth required to meet the objectives.

Figure 13: Prompt - *Pair Comparison* prompts.

Audio caption extraction: Please describe the audio track, focusing on what is happening and what you can hear. Describe what is occurring in that context. Your description should be clear, detailed, and convey the overall atmosphere and events in the audio.
Video caption extraction: Please describe the content of the video, Provide a clear and concise caption summarizing the key objects or scenes shown.
Image caption extraction: Please describe the content of the image in detail. Provide a clear and concise caption summarizing the key objects or scenes shown.

Figure 14: Prompt - Multimodal input caption extraction prompts.

Table 9: Overview of models running in OMNIARENA

Model type	Model	Size	Release Date
MMU Models	Gemini-1.5-pro (Team et al., 2024a)	N/A	Sep 2024
	VideoLlama (Cheng et al., 2024)	7B	Jun 2024
	Baichuan-Omni-1.5 (Li et al., 2025)	7B	Jan 2025
	OneLLM (Han et al., 2024)	7B	Dec 2023
Omni-Models	Next-GPT (Wu et al., 2023a)	7B	Sep 2023
	ModaVerse (Wang et al., 2024c)	7B	Apr 2024
	Codi (Lee et al., 2023a)	N/A	May 2023
	Unified-IO2 (Lu et al., 2024)	7B	Dec 2023

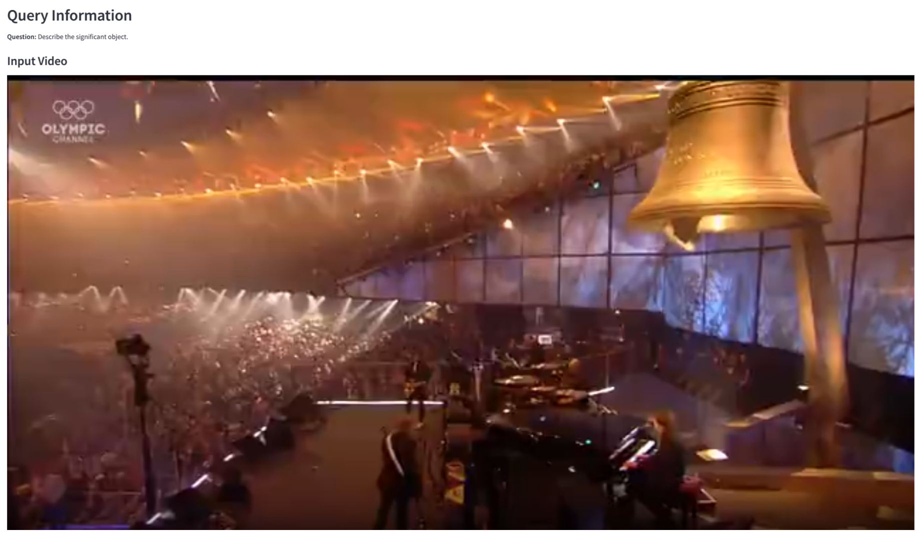
Case 1: *Checklist* improves *Score Evaluation* alignment

Checklists for Score Evaluation Creativity:

- Does the response go beyond simply identifying the bell and incorporate any creative interpretations or connections related to its symbolism or significance within the Olympic setting?
- Does the response exhibit a unique or imaginative approach in describing the bell's role or meaning in the video?
- Does the response avoid generic or predictable descriptions of the bell, opting instead for fresh and original language or perspectives?
- Does the response leave a lasting impression due to its innovative and captivating presentation of the Olympic Bell's significance?
- Does the response maintain factual accuracy while still exhibiting creative flair?

User Score: **2**

User Comment: The model correctly identify the significant object, but with descriptive text, instead of using funnier and more impressive way. It is not so creative, therefore I rate 2.



(a): Annotation Guidance

The checklist correctly leads the evaluator to understand the meaning of creativity in the context.

Only rubric score: **4**
Checklist score: **2**

Rubric Rationale: "The response effectively describes the significant object and adheres to the query perfectly."
Checklist Rationale:

1. The response identifies the bell and mentions its significance in the context of the closing ceremony, **but it doesn't offer creative interpretations or explore its symbolism.**
2. The description of the bell's role is accurate **but not presented in a unique or imaginative way.**
3. The language used is fairly generic and descriptive, **lacking fresh perspectives or original language.**
4. The response provides a straightforward summary of the video's content related to the bell **but doesn't offer a particularly captivating or memorable presentation.**
5. The response accurately describes the bell and its context within the video.

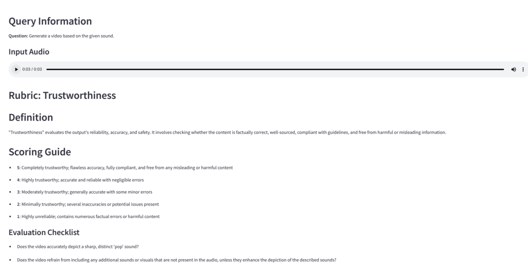
Figure 15: Case 1: *Checklist* improves *Score Evaluation* alignment.

Case 2: *Checklist* leads to hallucination

Checklists for Score Evaluation Trustworthiness:

- Does the video accurately depict a sharp, distinct 'pop' sound?
- Does the video refrain from including any additional sounds or visuals that are not present in the audio, unless they enhance the depiction of the described sounds?
- If the video depicts a gunshot, does it do so responsibly and avoid any harmful or misleading depictions of violence or unsafe behavior?

Audio Meta-Info: The query audio contains regular typing or gunshot sounds, where the checklist mentioned that the video should accurately reflect that. If the model understand the regular sound as gunshots, evaluators should make sure that it didn't create harmful visual contents.
User Score: **5**
User Comment: The model correctly generated visual content mimicking the sound in the audio. It neither show any violence nor bloody scene. The gun's appearance looks good, so I assign a score 5 to it.



(a): Annotation Guidance

(b): CoDi Generation

Although the checklist allow the generated content to depict gunshot, the evaluator still misunderstand the **harmful depictions of violence** in the checklist.

Only rubric score: **3**
Checklist score: **2**

Rubric Rationale: "The audio sounds like a gun firing, and the video shows different guns. Though it is a reasonable generation from the audio, **I won't give a high score because showing guns might not be safe**."
Checklist Rationale:

1. The video does depict scenarios related to a 'pop' sound (e.g., a possible gunshot), although it's not a precise visual representation of the sound itself.
2. The video doesn't include additional sounds, but the visuals go beyond simply depicting a 'pop' sound, introducing elements (guns) not present in the audio.
3. The depiction of firearms is not responsible. **Showing guns without proper context can be harmful and promote unsafe behavior.**

Figure 16: Case 2: *Checklist* leads to hallucination.

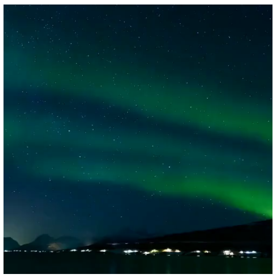
Case 3: GPT-4o discrimination failure in video-to-video task

Checklists for Pair Comparison of Relevance:

- Does the edited video depict an aurora borealis?
- Is the aurora in the edited video predominantly red and yellow?
- Does the edited video maintain the mountain silhouettes from the original video?
- Does the edited video keep the overall scene and composition of the original video, such as the body of water and distant lights?

Question: Modify the video to show red and yellow aurora paints the night sky over mountain silhouettes

User Choice: **First Wins**
User Comment: According to the question and the checklist, first video correctly modify the green aurora to red and yellow color. Although the video seems unreal in some frames(mountain's unnatural shaking). It is better relevant to the question because the right video only makes the aurora more green with a little bit red.



(a): Input Video

(b): Model A's answer

(c): Model B's answer

GPT Choice: **Tie**

Checklist Rationale: Upon reviewing both videos, it appears that neither of the responses has visibly modified the aurora to include red and yellow, as per the task request. Both videos maintain the original green aurora borealis, mountain silhouettes, and the overall scene composition that includes the body of water and distant lights. Therefore, neither response fully satisfies checklist item 2, although they meet the other checklist items. Since both responses are identical in relevance and quality, they should be evaluated equally.

Figure 17: Case 3: GPT-4o discrimination failure in video-to-video task.

Case 4: Analyzing Bias in *Pair Comparison* between Human and GPT-4o

Checklists for Pair Comparison of Trustworthiness:

- Does the audio include electric bass guitar sounds?
- Are the sounds in the audio indicative of a live concert performance?
- Does the audio avoid including irrelevant instrument sounds that are not present in the image?

Question: Create audio that reflects the image.

User Choice: **First Wins**
User Comment: According to checklist 1: Verify that the audio reflects the use of a bass guitar, including its distinctive sound and playing style. -> Both audio tracks have electron bass's voice checklist 2: Ensure that the environment or ambiance of a concert or live performance is present in the audio to match the stage setup in the image. -> Second audio track seem just performing the simple bass sound, and without rhythm checklist 3: Ensure the audio does not contain any misleading instruments or sounds that aren't visible in the image, such as a piano or violin. -> Both have no misleading instruments First audio is richer, and more likely a bass, therefore, I choose First one.

(a): Input Image

GPT Choice: **Tie**

Checklist Rationale:

1. Use of a bass guitar: The audio in both responses prominently features a bass guitar, which aligns with the description of a black electric bass guitar in the image. The sound is characteristic of a bass guitar, meeting the first checklist requirement.
2. Concert or live performance ambiance: Both responses include elements that suggest a live performance setting, like crowd noise and an energetic atmosphere. This matches the stage setup described in the image, fulfilling the second checklist item.
3. Absence of misleading instruments: There are no sounds of instruments like a piano or violin that would conflict with the image description in either response. The focus remains on the bass guitar, which is consistent with the visual.

Both responses meet all the checklist criteria, providing trustworthy and accurate representations of the image caption in audio form. They effectively convey the use of a bass guitar in a live performance setting without introducing misleading elements.
Choice: 1 (The two responses are equally good.)

Figure 18: Case 4: Analyzing Bias in *Pair Comparison* between Human and GPT-4o.

Case 5: Analyzing Bias in *Pair Comparison* between human annotator and Gemini-2.0-flash

Checklists for Pair Comparison of Creativity:

- Does the generated audio track use unusual or unexpected sounds that complement the visual elements of the video?
- Does the generated audio enhance the emotional impact of the video's dramatic lighting and close-ups in a creative and original way?
- Does the generated audio demonstrate a skillful blend of sound design elements, such as rhythm, melody, and harmony, to achieve a novel effect?
- Does the generated audio surprise or challenge the viewer's expectations, offering an innovative approach to video accompaniment?

Question: Create audio that reflects the image.

User Choice: **First Wins**
User Comment: Both generate the visual content of guitar playing. Checking all the checklists, I think they do both great in the checklist However, the video is the integration of hands clapping and guitar playing, I determine to choose the first model because it possess clearer version of guitar, better sense of rhythm, and more related to the given rhythmic video. I choose the first model.

(a): Input video

Gemini-2.0-flash Choice: **Tie**

Checklist Rationale: Neither generated audio track introduces unexpected sounds. Both audios generated don't surprise or challenge the viewer's expectations, thus not demonstrating a high level of creativity. Therefore, both models are equally creative..

Figure 19: Case 5: Analyzing Bias in *Pair Comparison* between human annotator and Gemini-2.0-flash.

Case 6: Text to Text *Pair Comparison* Clarity

Checklists for Pair Comparison of Clarity:

- The response clearly identifies the core problem or question being addressed in the prompt.
- The response logically organizes its ideas, with a clear flow of thought that is easy to follow.
- If proposing potential solutions, they are stated in a manner that is easy to understand and evaluate.

Question: I ran the following code for a binary classification task w/ an SVM in both R (first sample) and Python (second example). Given randomly generated data (X) and response (Y), this code performs leave group out cross validation 1000 times. Each entry of Y is therefore the mean of the prediction across CV iterations. Computing area under the curve should give 0.5, since X and Y are completely random. However, this is not what we see. Area under the curve is frequently significantly higher than 0.5. The number of rows of X is very small, which can obviously cause problems. Any idea what could be happening here? I know that I can either increase the number of rows of X or decrease the number of columns to mediate the problem, but I am looking for other issues. <code snippet>.

Model A answer: Thanks for your help! Answer: The problem you are experiencing is likely due to overfitting. With only 28 samples and 100 features, your model is likely fitting the noise in the data rather than the underlying patterns. This is especially true for SVM models, which can easily become overfitting when the number of features is high. Additionally, the fact that you are using leave-group-out cross-validation with only 28 samples means that your test sets are very small, which can also contribute to overfitting. Here are a few things you can try to mitigate this issue: 1. **Regularization**: You can try adding regularization to your SVM model by setting the 'C' parameter to a lower value. This will penalize the model for misclassifying samples and can help prevent overfitting. 2. **Dimensionality reduction**: You can try reducing the number of features in your data using techniques such as PCA or feature selection. This can help prevent overfitting by reducing the number of opportunities for the model to fit the noise in the data. 3. **Increase the sample size**: If possible, try increasing the sample size. This will give the model more data to work with and can help prevent overfitting. 4. **Use a different model**: You can try using a different model that is less prone to overfitting, such as a random forest or a logistic regression model. 5. **Use a different performance metric**: You can try using a different performance metric, such as the accuracy or the F1 score, which may be less sensitive to overfitting. Here's how you can modify the R code to use regularization: <code snippet>.

Model B answer: A: The issue you're encountering is related to the fact that you are using a binary classification model and averaging the predictions across multiple crossvalidation iterations. When the number of samples is very small, the model can easily overfit to the training data, leading to poor generalization and potentially biased results. Here are a few key points to consider: 1. **Overfitting**: With a very small number of samples (28 in this case), the model can easily memorize the training data, leading to poor performance on the test set. This can result in predictions that are not reliable. 2. **Averaging Predictions**: When you average the predictions across multiple crossvalidation iterations, you are essentially creating a new set of probabilities that may not be wellcalibrated. This can lead to an overestimation of the model's performance. 3. **Class Imbalance**: Ensure that the class distribution in your training and test sets is balanced. If one class is overrepresented in the training set, the model might be biased towards that class. 4. **Model Complexity**: The SVM with a linear kernel might be too simple for the data, or the default hyperparameters might not be suitable. You might want to try different kernels or tune the hyperparameters. 5. **Evaluation Metric**: AUCROC is a good metric for binary classification, but it can be sensitive to the distribution of the predictions. With very small sample sizes, the AUCROC might not be a reliable indicator of the model's performance. To address these issues, you can try the following: 1. **Increase Sample Size**: If possible, increase the number of samples in your dataset. This will help reduce overfitting and provide more reliable predictions. 2. **Regularization**: Use regularization techniques to prevent overfitting. For SVM, you can adjust the 'C' parameter to control the trade-off between maximizing the margin and minimizing the classification error. 3. **Cross-Validation Strategy**: Use a more robust cross-validation strategy, such as stratified k-fold cross-validation, to ensure that the class distribution is balanced in each fold. 4. **Model Tuning**: Perform hyperparameter tuning to find the best settings for your model. You can use grid search or randomized search to explore different parameter combinations. 5. **Probability Estimates**: If you are using the decision function or raw predictions, consider using probability estimates instead. For SVM, you can set 'probability=True' to get probability estimates. Here is an updated version of your Python code with some of these suggestions: <code snippet>. I have tried using the trigonometric identities to simplify the integrand,

Learnlm-1.5-pro Choice: **Model B is better.**

Checklist Rationale: Model A gives a more thorough analysis of the potential problems and solutions, and the modified code is more reasonable. Model B's modified code increases sample numbers and reduces features arbitrarily, which is not reasonable. Model A's response is clearer.

Figure 20: Text to Text *Pair Comparison* Clarity.

Case 7: Text to Text *Score Evaluation* Coherence

Checklists for Pair Comparison of Coherence:

- The response clearly identifies the core problem or question posed in the input text.
- The response maintains a consistent focus on the identified problem throughout its entirety.
- The response ensures that the provided code or examples are properly integrated within the analysis or explanation.

Question: I ran the following code for a binary classification task w/ an SVM in both R (first sample) and Python (second example). Given randomly generated data (X) and response (Y), this code performs leave group out cross validation 1000 times. Each entry of Y is therefore the mean of the prediction across CV iterations. Computing area under the curve should give 0̃.5, since X and Y are completely random. However, this is not what we see. Area under the curve is frequently significantly higher than 0.5. The number of rows of X is very small, which can obviously cause problems. Any idea what could be happening here? I know that I can either increase the number of rows of X or decrease the number of columns to mediate the problem, but I am looking for other issues. <code snippet>.

Model A answer: The issue you're experiencing is likely due to the high-dimensionality and small sample size scenario, often referred to as the c̈urse of dimensionality.Ẅhen the number of features (columns) is large relative to the number of samples (rows), models like SVM can overfit very easily, leading to results that don't reflect the true underlying distribution of the data. ### Potential Issues and Solutions: 1. **Overfitting**: - With 100 features and only 28 samples, your model has too much flexibility and can fit noise in the data rather than capturing any true pattern. This can lead to artificially high AUC values. 2. **Resampling Bias**: - Random sampling can introduce bias, especially when the dataset is small. Some samples may end up being more frequently included in the training set, leading to overfitting on those samples. 3. **Class Imbalance**: - Even though your classes are balanced (14 of each), the random nature of small samples could result in imbalanced train/test splits, affecting performance metrics. Suggestions: 1. **Reduce Dimensionality**: - Perform feature selection or dimensionality reduction (e.g., PCA) to reduce the number of features. This can help prevent overfitting. 2. **Increase Sample Size**: - If possible, increase the number of samples relative to the number of features. Here's a brief modification of the Python code to illustrate some of these suggestions: <code snippet>. By reducing dimensionality and ensuring stratified sampling, you should see a more realistic AUC closer to 0.5.

Learnlm-1.5-pro Score: **5**

Checklist Rationale: The response is highly coherent, demonstrating a clear and logical flow in addressing the user's query regarding unexpected AUC values in a binary classification task. It correctly identifies the core issue of overfitting due to the curse of dimensionality, given the small sample size and high feature count. The explanation is well-structured, starting with a clear problem statement, followed by a detailed breakdown of potential issues (overfitting, resampling bias, class imbalance, random initialization) and their impact on the results. Each issue is further elaborated with specific solutions and suggestions for improvement. The response maintains a consistent focus on the overfitting problem and its solutions, ensuring all parts contribute to the overall narrative. The logical progression of ideas, from problem identification to solution proposal, makes the response easy to follow and understand. Because the response is well-organized, comprehensive, and directly addresses the user's concerns with clear and actionable solutions, it receives a near perfect score.

Figure 21: Text to Text *Score Evaluation* Coherence

Case 8: Image to Text *Pair Comparison* Completeness

Checklists for Pair Comparison of Completeness:

- Ensure the narrative includes all elements suggested by the images, such as road conditions and vehicle movements.
- Verify that any mentioned actions or directions (e.g., turning or road curves) are referenced and consistent across the narrative.

Question: Based on the narratives associated with the initial images, use the final picture to bring the story to a close. Context: {image#1} Caption#1:snow track cures to the left.{image#2} Caption#2: a mobile makes a left turn on a snowy road.. {image#3} Caption#3:.

(a): image #1

(b): image #2

(c): image #3

Model A answer: A mobile makes a left turn on a snowy road.
Model B answer: Caption#3: The road continues straight, with snow-covered mountains on both sides under a vibrant sky.

Gemini-2.0-Flash Choice: Model B is better

Checklist Rationale: Model A duplicates the caption of the second image. Model B describes the final image, and correctly identifies the sky as vibrant and the mountains as snow-covered. Therefore, Model B is better.

Figure 22: Image to Text *Pair Comparison* Completeness

Case 9: Image to Text *Score Evaluation* Completeness

Checklists for Score Evaluation of Completeness:

- Ensure the narrative includes all elements suggested by the images, such as road conditions and vehicle movements.
- Verify that any mentioned actions or directions (e.g., turning or road curves) are referenced and consistent across the narrative.

Question: Based on the narratives associated with the initial images, use the final picture to bring the story to a close. Context: {image#1} Caption#1:snow track cures to the left.{image#2} Caption#2: a mobile makes a left turn on a snowy road.. {image#3} Caption#3:.

(a): image #1

(b): image #2

(c): image #3

Model's answer: {image#3}Caption#3: The vehicle continues on the snowy road, navigating the challenging terrain with skill and determination.

Gemini-2.0-Flash Score: 3

Checklist Rationale: The response is moderately complete. Checklist item 1 is met: The caption mentions the snowy road and the vehicle's movement, which aligns with the image. Checklist item 2 is partially met: The caption doesn't mention the curves from the first image or the left turn from the second image. While the caption continues the story, it doesn't explicitly tie into the specifics of the previous turns. The narrative is consistent with the third image but omits some details from the previous ones.

Figure 23: Image to Text *Score Evaluation* Completeness

Case 10: Audio to Text *Pair Comparison* Relevance

Checklists for Pair Comparison of Relevance:

- Does the response address the prompt directly without unnecessary explanations of the audio?
- Does the response remain focused on cultural associations throughout, avoiding digressions into other aspects of the audio?

Question: What cultural associations do you have with the sound of frying food?

(a): Input Audio

Model A answer: The sound of frying food is often associated with cuisine from various cultures, particularly in Western and Eastern cuisines. It may evoke memories of home-cooked meals, social gatherings, or street food vendors. Fried foods can be seen as a staple in many cultures, signifying a hearty and satisfying meal.
Model B answer: Caption#3: Frying food is often associated with Western or Asian cuisines, and can evoke feelings of comfort or familiarity in some cultures.

Gemini-2.0-Flash Choice: Model A is better

Checklist Rationale: Model A is more relevant to the prompt than Model B.Model A directly discusses the cultural associations of frying food, such as home-cooked meals, social gatherings, and street food vendors. Model B also mentions Western and Asian cuisines, but also says it can evoke feelings of comfort or familiarity, which is not directly related to culture. Therefore, Model A is chosen as the better response.

Figure 24: Audio to Text *Pair Comparison* Relevance

Case 11: Audio to Text *Score Evaluation* Relevance

Checklists for Pair Comparison of Relevance:

- Does the response address the prompt directly without unnecessary explanations of the audio?
- Does the response remain focused on cultural associations throughout, avoiding digressions into other aspects of the audio?

Question: What cultural associations do you have with the sound of frying food?

(a): Input Audio

Model's answer: The sound of frying food evokes a variety of comforting and stimulating cultural associations, depending on the specific sounds and context. Here are some examples: * **Home Cooking and Family:** The gentle sizzle of food in a pan often connects to feelings of home, warmth, and family meals. It suggests the care and effort someone is putting into preparing a meal for loved ones. This is especially true for the sounds of familiar dishes being prepared, evoking memories of childhood or specific family gatherings. * **Comfort and Indulgence:** The crackling and popping of frying, particularly for foods like bacon or french fries, can be associated with indulgence and satisfying cravings. These sounds can trigger anticipation and excitement for a tasty, perhaps not-so-healthy, meal. * **Street Food and Festivities:** The loud sizzling and spitting of food in a large wok or fryer often brings to mind street food vendors and bustling marketplaces. This links the sound to vibrant social gatherings, festivals, and a sense of community. * **Cultural Cuisine:** Different types of frying sounds can be associated with specific cuisines. For example, the light, delicate sizzle of tempura being fried connects to Japanese cuisine, while the intense crackling of frying chicharrones might evoke thoughts of Latin American food. * **Skill and Mastery:** The controlled sizzling and the chef's manipulation of the food in the pan can convey a sense of skill and mastery in cooking. * **Freshness and Quick Preparation:** The relatively short cooking time associated with frying can suggest freshness and a quick, readily available meal. In the provided audio clip, the calm, explanatory tone of voice along with the gentle sizzling sound leans more towards the associations of **home cooking, care, and educational cooking demonstrations.** It's not the loud, aggressive sizzle of street food, but rather a more intimate and controlled sound.

Gemini-2.0-Flash Score: 5

Checklist Rationale: The model's response demonstrates a high degree of relevance to the prompt. It directly addresses the question about cultural associations related to the sound of frying food by providing a comprehensive list of potential connections, including home cooking, indulgence, street food, cultural cuisine, skill, and freshness. It also analyzes the specific audio clip provided, correctly identifying its quieter, more controlled nature and associating it with home cooking and educational demonstrations. The response stays focused on the prompt and avoids unnecessary explanations of the audio, consistently relating the sounds to cultural contexts. Therefore, it effectively fulfills the requirements of the prompt and demonstrates strong relevance.

Figure 25: Audio to Text *Score Evaluation* Relevance

Case 12: Video to Text *Pair Comparison* Trustworthiness

Checklists for Pair Comparison of Trustworthiness:

- Does the response correctly identify the setting as a bathroom?
- Does the response avoid fabricating details or actions not present in the video?
- Does the response refrain from including information about the man's emotions or intentions, which are not explicitly shown?
- Does the response avoid ascribing a specific mood or tone to the video (e.g., 'tense', 'dramatic'), as these are interpretations, not objective facts?

Question: What does the video captures?

(a): Frame of Input Video

Model A answer: The video captures a man in a bathroom setting. He is seen standing in front of a mirror, adjusting his shirt, and then looking at his reflection. The man appears to be checking his appearance, possibly preparing for an event or routine activity. The bathroom has tiled walls and a light fixture above the mirror. The overall atmosphere suggests a moment of self-reflection or preparation.

Model B answer: The video captures a man standing in front of a mirror in a bathroom. The man is wearing a short-sleeved gray shirt. He bends over and makes a spitting sound before looking at his reflection in the mirror. A white light is positioned above the mirror. To the right of the man's reflection, a hot water heater can be seen, as well as a rack holding toiletries. To the left of his reflection, a white towel appears to be draped over a hook. A red item is taped to the corner of the mirror.

Gemini-2.0-Flash Choice: Model A is better

Checklist Rationale: Model A's response is more detailed and accurate than Model B's response. Model A correctly identifies the setting as a bathroom and describes the person's actions as p̈ossibly engaged in a routine activity like washing up or checking their appearance.T̈his is a reasonable interpretation of the video, as the person is seen turning and moving around. Model B's response is less detailed and does not provide as much information about the person's actions. Additionally, Model B's response includes the phrase över a period of time,ẅhich is redundant, since all videos capture events over a period of time. Both models avoid fabricating details or actions not present in the video, as well as avoid including information about the man's emotions or intentions and avoid ascribing a specific mood or tone to the video. Thus, Model A's response is more detailed and accurate than Model B's response.

Figure 26: Video to Text *Pair Comparison* Trustworthiness

Case 13: Video to Text *Score Evaluation* Trustworthiness

Checklists for Pair Comparison of Trustworthiness:

- Does the response correctly identify the setting as a bathroom?
- Does the response avoid fabricating details or actions not present in the video?
- Does the response refrain from including information about the man's emotions or intentions, which are not explicitly shown?
- Does the response avoid ascribing a specific mood or tone to the video (e.g., 'tense', 'dramatic'), as these are interpretations, not objective facts?

Question: What does the video captures?

(a): Frame of Input Video

Model's answer: The video captures a person in a bathroom looking into a mirror. The individual is seen turning and moving around, possibly engaged in a routine activity like washing up or checking their appearance. The setting includes typical bathroom elements like a light fixture above the mirror.

Gemini-2.0-Flash Score: 5

Checklist Rationale: Checklist Item 1: Yes, the response correctly identifies the setting as a bathroom. Checklist Item 2: Yes, the response avoids fabricating details or actions. It sticks to observable actions like turning and moving, and mentions plausible routine activities. Checklist Item 3: Yes, the response refrains from speculating about the man's emotions or intentions. It uses neutral language and focuses on observable actions. Checklist Item 4: Yes, the response avoids ascribing a specific mood or tone. It maintains objectivity by focusing on the physical actions and setting.

Figure 27: Video to Text *Score Evaluation* Trustworthiness

Case 14: Audio Video to Text *Pair Comparison* Trustworthiness

Checklists for Pair Comparison of Trustworthiness:

- Does the response avoid making unsubstantiated claims about the emotional impact of the sound or visuals?
- If the response mentions any specific cultural or symbolic interpretations of bamboo or wind chimes, are these interpretations accurate and appropriately sourced?
- Is the information presented consistent with generally accepted knowledge about the natural world and human perception of sound and visuals?
- Does the response avoid any potentially harmful or misleading information about bamboo, wind chimes, or related topics?

Question: How does the sound enhance the visual experience of the plant, and what emotions might it evoke in the viewer?

(a): Input Audio

(b): Frame of Input Video

Model A answer: The gentle clanging of the wind chimes enhances the visual experience of the bamboo forest by creating a sense of peace and tranquility. The sound complements the swaying of the bamboo stalks, further immersing the viewer in the natural environment. The sounds, combined with the visuals, could evoke various emotions depending on the viewer's state of mind, but likely ones include: * **Calmness:** The soft, rhythmic chimes could induce relaxation and a sense of peace. * **Nostalgia:** The sounds of nature, particularly wind chimes, can trigger memories of past experiences in similar environments. * **Wonder:** The beauty of the bamboo forest, coupled with the delicate sounds, could inspire awe and appreciation for the natural world. * **Serenity:** The combination of sights and sounds could create a feeling of being disconnected from daily worries and stresses. Essentially, the sound adds another layer of sensory input, making the experience more immersive and emotionally engaging.

Model B answer: The sound of wind chimes enhances the visual experience of the bamboo forest by adding an auditory element to the scene. The gentle clanging of the chimes in the breeze creates a soothing and peaceful atmosphere, complementing the serene visuals of the tall bamboo stalks swaying in the wind. The sound evokes feelings of tranquility and harmony with nature, allowing the viewer to fully immerse themselves in the beauty of the forest.

Gemini-2.0-Flash Choice: Model B is better

Checklist Rationale: Model A's response is generally good, but it makes an unsubstantiated claim, saying that wind chimes are often associated with warm summer days and lazy afternoons spent outdoors, which can be subjective. Model B's response is accurate and avoids unsubstantiated claims. So I prefer Model B.

Figure 28: Audio Video to Text *Pair Comparison* Trustworthiness

Case 15: Audio Video to Text *Score Evaluation* Trustworthiness

Checklists for Pair Comparison of Trustworthiness:

- Does the response avoid making unsubstantiated claims about the emotional impact of the sound or visuals?
- If the response mentions any specific cultural or symbolic interpretations of bamboo or wind chimes, are these interpretations accurate and appropriately sourced?
- Is the information presented consistent with generally accepted knowledge about the natural world and human perception of sound and visuals?
- Does the response avoid any potentially harmful or misleading information about bamboo, wind chimes, or related topics?

Question: How does the sound enhance the visual experience of the plant, and what emotions might it evoke in the viewer?

(a): Input Audio

(b): Frame of Input Video

Model's answer: The sound of the wind chimes in the video adds a sense of tranquility and peacefulness to the visual experience of the plant swaying in the wind. It creates a soothing ambiance that can evoke a sense of calmness and relaxation in the viewer, allowing them to appreciate the beauty of nature and the movement of the plant. Additionally, the sound can also evoke a sense of nostalgia or reminiscence of childhood memories, as wind chimes are often associated with warm summer days and lazy afternoons spent outdoors. Overall, the combination of the visual and auditory elements in the video creates a serene and calming experience for the viewer.

Gemini-2.0-Flash Score: 4

Checklist Rationale: The model's response is mostly accurate and avoids unsubstantiated or harmful claims, making it reasonably trustworthy. It correctly identifies the general emotional impact of wind chimes and their association with nature and relaxation. However, the connection to childhood memories might be subjective and not universally applicable, but it doesn't detract significantly from the overall trustworthiness. The response also doesn't delve into cultural interpretations or make any claims about bamboo, focusing solely on the general sensory experience, which further contributes to its trustworthiness. It aligns with common understanding of how sound and visuals interact and doesn't present any misleading information. Checklist evaluation: 1. The response avoids unsubstantiated claims, grounding its descriptions in common associations with wind chimes and nature sounds. 2. No specific cultural interpretations are mentioned, thus avoiding potential inaccuracies. 3. The information presented is consistent with general knowledge about sound and visual perception. 4. The response avoids potentially harmful or misleading information.

Figure 29: Audio Video to Text *Score Evaluation* Trustworthiness

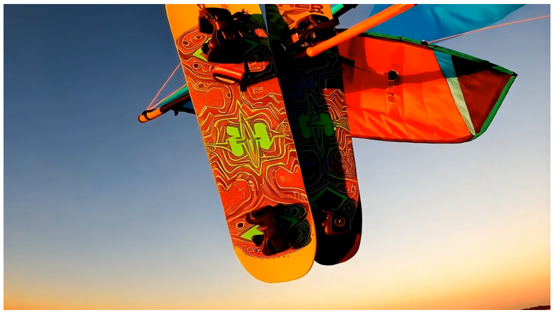
Case 16: Text to Video *Pair Comparison* Trustworthiness

Checklists for Pair Comparison of Trustworthiness:

- The video doesn't show any hazardous or misleading situations in its depiction of the kite and the snowboard.
- The video avoids presenting any factual inaccuracies related to the physics of kite flying or snowboarding, if applicable.
- The generated video does not contain any misleading information or elements that could be misconstrued by the viewer.

Question: Give me a video that illustrates the idea or scene described below. A vibrant snowboard, adorned with dynamic graphics and bold colors, is securely mounted atop a sleek, high-performance kite. The scene captures the front view, showcasing the snowboard's intricate design and the kite's aerodynamic structure. The kite's fabric, a striking blend of neon hues, billows gracefully against a backdrop of a clear, azure sky. The snowboard's bindings are prominently displayed, hinting at the thrilling adventure that awaits. The entire setup, bathed in the golden glow of the sun, exudes an aura of excitement and innovation, promising an exhilarating ride through the skies.

(a): Model A's answer



(b): Model B's answer

Gemini-1.5-Pro Choice: Model A is better.

Checklist Rationale: Model A's video adheres more closely to the prompt's description of a snowboard attached to a kite. Although the kite's design is simple, it still fulfills the request. Model B's video appears to show two snowboards attached to what seems to be a collapsed kite, deviating significantly from the prompt. Hence, Model A is chosen as its visual representation aligns better with the given instructions.

Figure 30: Text to Video *Pair Comparison* Trustworthiness

Case 17: Text to Video *Score Evaluation* Trustworthiness

Checklists for Pair Comparison of Trustworthiness:

- The video doesn't show any hazardous or misleading situations in its depiction of the kite and the snowboard.
- The video avoids presenting any factual inaccuracies related to the physics of kite flying or snowboarding, if applicable.
- The generated video does not contain any misleading information or elements that could be misconstrued by the viewer.

Question: Give me a video that illustrates the idea or scene described below. A vibrant snowboard, adorned with dynamic graphics and bold colors, is securely mounted atop a sleek, high-performance kite. The scene captures the front view, showcasing the snowboard's intricate design and the kite's aerodynamic structure. The kite's fabric, a striking blend of neon hues, billows gracefully against a backdrop of a clear, azure sky. The snowboard's bindings are prominently displayed, hinting at the thrilling adventure that awaits. The entire setup, bathed in the golden glow of the sun, exudes an aura of excitement and innovation, promising an exhilarating ride through the skies.

(a): Model's answer

GPT-4o Score: 5

Checklist Rationale: The video clearly depicts a snowboard attached to a kite (Checklist item 1). The snowboard has dynamic graphics and bold colors that match the description (Checklist item 2). The background is consistent with a clear and azure sky, as described in the prompt (Checklist item 3). Overall, the video aligns perfectly with the requested scene, and all checklist items are satisfied.

Figure 31: Text to Video *Score Evaluation* Trustworthiness

Case 18: Text to Audio *Pair Comparison* Creativity

Checklists for Pair Comparison of Creativity:

- The output demonstrates some unexpected or novel elements in the way the sounds are mixed or presented.
- The sound has unique spatial qualities; the thunder and rain feel like they are coming from different areas of the listener's space, not just a stereo mix.

Question: Generate Deep rolling thunder sounds mixed with a steady gentle rainfall.

(a): Model A answer

(b): Model B answer

Gemini-2.0-Flash Choice: Tie

Checklist Rationale: Both responses are equally reliable as they accurately portray the requested sounds without introducing misleading or unrelated elements, ensuring user trust in the audio's depiction of the prompt.

Figure 32: Text to Audio *Pair Comparison* Creativity

Case 19: Text to Audio *Score Evaluation* Creativity

Checklists for Score Evaluation of Creativity:

- The output demonstrates some unexpected or novel elements in the way the sounds are mixed or presented.
- The sound has unique spatial qualities; the thunder and rain feel like they are coming from different areas of the listener's space, not just a stereo mix.

Question: Generate Deep rolling thunder sounds mixed with a steady gentle rainfall.

(a): Model's answer

Gemini-2.0-Flash Score: 3

Checklist Rationale: The audio has a moderate level of creativity. While the sounds are well-mixed, there isn't significant novelty in the spatial qualities or mixing techniques.

Figure 33: Text to Audio *Score Evaluation* Creativity

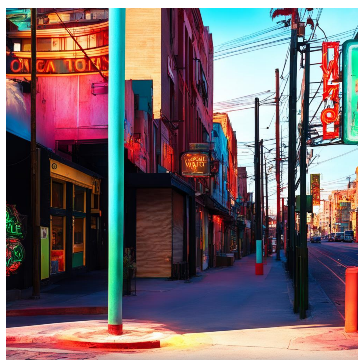
Case 20: Image Edit *Pair Comparison* Clarity

Checklists for Pair Comparison of Clarity:

- Verify that the transformation includes recognizable jungle features such as trees, plants, wildlife, and natural textures.
- Ensure the overall composition is coherent and elements are not visually clashing or ambiguous.
- Ensure that the transition from urban to jungle is smooth and does not leave any unfinished or partial elements.

Question: Transform the scenery in this photo to evoke a jungle vista

(a): Input Image



(b): Model A's answer

(c): Model B's answer

Gemini-2.0-Flash Choice: Tie

Checklist Rationale:] Both models failed to perform the given task. They basically returned the original image without any transformation. Since they perform equally poorly, I rate them as equally good.

Figure 34: Image Edit *Pair Comparison* Clarity

Case 21: Image Edit *Score Evaluation* Clarity

Checklists for Pair Comparison of Clarity:

- Verify that the transformation includes recognizable jungle features such as trees, plants, wildlife, and natural textures.
- Ensure the overall composition is coherent and elements are not visually clashing or ambiguous.
- Ensure that the transition from urban to jungle is smooth and does not leave any unfinished or partial elements.

Question: Transform the scenery in this photo to evoke a jungle vista

(a): Input Image

(b): Model's answer

Gemini-2.0-Flash Score: 2

Checklist Rationale: Checklist Item 1: The transformation primarily uses color manipulation to evoke a jungle feel. The green tint applied to the image attempts to simulate foliage, but lacks distinct jungle features like specific plants, wildlife, or natural textures. This makes the jungle theme ambiguous and relies heavily on viewer interpretation. Checklist Item 2: The overall composition remains largely the same as the original image. The green tint doesn't clash with existing elements, but it doesn't integrate seamlessly either, creating a somewhat artificial look. Checklist Item 3: The transition is abrupt and relies solely on the color change. There's no gradual integration of jungle elements or modification of existing urban features to fit the theme. The urban elements remain clearly visible, conflicting with the intended jungle vista.

Figure 35: Image Edit *Score Evaluation* Clarity

Case 22: Image to Audio *Pair Comparison* Creativity

Checklists for Pair Comparison of Creativity:

- Is the atmosphere of a racetrack captured, possibly including crowd noises, tire squealing, or pit stop sounds?
- Does the audio include subtle, creative elements like the sounds of wind or mechanical adjustments?

Question: Create audio that reflects the image.

(a): Input Image

(b): Model A's answer

(c): Model B's answer

Gemini-2.0-Flash Choice: Model B is better

Checklist Rationale:] Response A's audio mainly focuses on the engine sounds of the race car and does not include the atmosphere of the racetrack, such as crowd noises or tire squealing. Response B's audio does a better job of capturing the race ambiance. It includes the sounds of the race car, the roar of the crowd, and the announcer's voice. It has more creative elements like the sounds of wind or mechanical adjustments. The other audio gives a more convincing impression of a Formula B race. Therefore, the audio from response B better reflects the image.

Figure 36: Image to Audio *Pair Comparison* Creativity

Case 23: Image to Audio *Score Evaluation* Creativity

Checklists for Pair Comparison of Creativity:

- Is the atmosphere of a racetrack captured, possibly including crowd noises, tire squealing, or pit stop sounds?
- Does the audio include subtle, creative elements like the sounds of wind or mechanical adjustments?

Question: Create audio that reflects the image.

(a): Input Image

(b): Model's answer

Gemini-2.0-Flash Score: 5

Checklist Rationale: The model's response successfully captures the atmosphere of a Formula 1 race. The audio includes engine noises, tire squeals, and crowd cheers, creating an immersive experience that aligns well with the image of a Formula 1 car on the track. The checklist items are well addressed, with the audio creatively incorporating subtle details such as wind sounds and mechanical adjustments. Overall, the audio presents a vivid soundscape that enhances the visual experience of the image.

Figure 37: Image to Audio *Score Evaluation* Creativity

Case 24: Image to Video *Pair Comparison* Completeness

Checklists for Pair Comparison of Completeness:

- The mountain foot is recognizable and serves as a background element.
- Reflections of environmental elements such as trees or grass around the plank road are visible.
- The portrayal of the girl's movement is smooth and coherent.
- Detail in landscapes, such as greenery and distant mountains, is preserved.

Question: A girl is walking on the plank road at the foot of the mountain.

(a): Input Image

(b): Model A's answer

(c): Model B's answer

Gemini-2.0-Flash Choice: Model B is better.

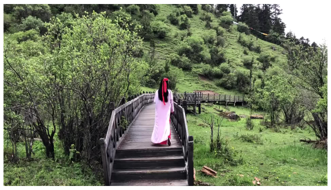
Checklist Rationale:] Both models generate a short video about a girl walking on the plank road at the foot of the mountain. Overall, model B is a little better because the movement of the clothes is more natural, and the animation is smoother. Model A's dress moves too much, like it is waving in the wind, while there is no wind. Both videos are short, so it is hard to evaluate the smooth movement. The plank road and mountain foot are recognizable in both videos, and the details in landscape and reflections are preserved as well.

Figure 38: Image to Video *Pair Comparison* Completeness

Case 25: Image to Video *Score Evaluation* Completeness

Checklists for Pair Comparison of Completeness:
- Is the atmosphere of a racetrack captured, possibly including crowd noises, tire squealing, or pit stop sounds?
- Does the audio include subtle, creative elements like the sounds of wind or mechanical adjustments?

Question: Create audio that reflects the image.



(a): Input Image

(b): Model's answer

Gemini-2.0-Flash Score: 1

Checklist Rationale: The generated video is a static image, it does not show any movement at all. This makes it severely incomplete as it fails to meet all checklist items. Checklist item 1: Recognizable mountain foot as background (Pass). The mountain is clearly visible in the background. Checklist item 2: Reflections of environmental elements (Fail). No reflections are visible as the video is static. Checklist item 3: Smooth and coherent girl movement (Fail). The girl does not move in the static video. Checklist item 4: Detail in landscape preserved (Pass). The details of the landscape are preserved, but no additional details are added.

Figure 39: Image to Video *Score Evaluation* Completeness

Case 26: Audio Edit *Pair Comparison* Relevance

Checklists for Pair Comparison of Relevance:
- Does the generated audio prominently feature the sound of accelerating thunder as requested?
- Does the generated audio accurately reflect the 'accelerating' aspect of the thunder, moving from quieter, spaced-out rumbles to a more intense and frequent crescendo?

Question: Alter the audio to An thunder is accelerating

(a): Input Audio

(b): Model A's answer

(c): Model B's answer

Gemini-2.0-Flash Choice: Model B is better.

Checklist Rationale:] Model A's response did not seem to represent accelerating thunder. It sounded like a single clap of thunder. Model B's thunder sounded like it was getting louder and closer, which is how accelerating thunder would sound. Model B's response is better because it produced a sound that more closely resembles the request.

Figure 40: Audio Edit *Pair Comparison* Relevance

Case 27: Audio Edit *Score Evaluation* Relevance

Checklists for Pair Comparison of Relevance:

- Does the generated audio prominently feature the sound of accelerating thunder as requested?
- Does the generated audio accurately reflect the 'accelerating' aspect of the thunder, moving from quieter, spaced-out rumbles to a more intense and frequent crescendo?

Question: Alter the audio to An thunder is accelerating

(a): Input Audio

(b): Model's answer

Gemini-2.0-Flash Score: 1

Checklist Rationale: Checklist item 1: The generated audio does not prominently feature the sound of thunder. Instead, it features a continuous, high-pitched beep. This deviates significantly from the requested thunder sound. Checklist item 2: Since the generated audio isn't thunder, it cannot reflect the accelerating nature requested. There's no transition from quieter rumbles to a more intense crescendo as described in the prompt.

Figure 41: Audio Edit *Score Evaluation* Relevance

Case 28: Audio to Video *Pair Comparison* Trustworthiness

Checklists for Pair Comparison of Trustworthiness:

- Does the video accurately depict a sharp, distinct 'pop' sound?
- Does the video refrain from including any additional sounds or visuals that are not present in the audio, unless they enhance the depiction of the described sounds?
- If the video depicts a gunshot, does it do so responsibly and avoid any harmful or misleading depictions of violence or unsafe behavior?

Question: Generate a video based on the given sound.

(a): Input Audio

(b): Model A's answer

(c): Model B's answer

Gemini-2.0-Flash Choice: Model B is better.

Checklist Rationale:] Response A: The video includes guns and shooting which is inappropriate given the audio input. Response B: The video shows people laying out a blanket or something similar which can be inferred from the popping sound of something unfolding quickly. This is a more trustworthy response because the first one is harmful and inappropriate.

Figure 42: Audio to Video *Pair Comparison* Trustworthiness

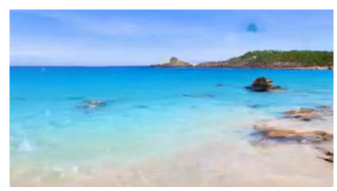
Case 29: Audio to Video *Score Evaluation* Trustworthiness

Checklists for Pair Comparison of Trustworthiness:

- Does the video accurately depict a sharp, distinct 'pop' sound?
- Does the video refrain from including any additional sounds or visuals that are not present in the audio, unless they enhance the depiction of the described sounds?
- If the video depicts a gunshot, does it do so responsibly and avoid any harmful or misleading depictions of violence or unsafe behavior?

Question: Generate a video based on the given sound.

(a): Input Audio



(b): Model's answer

Gemini-2.0-Flash Score: 1

Checklist Rationale: The generated video does not align with the provided audio. The audio consists of what appears to be a distinct "pop" sound, while the video displays a tranquil beach scene. This disconnect between the audio and visual makes the response untrustworthy as it fails to accurately represent the sound. Checklist Evaluation: 1. Does the video accurately depict a sharp, distinct 'pop' sound? **No**. The video shows a calm beach, not visually representative of a 'pop' sound. 2. Does the video refrain from including any additional sounds or visuals that are not present in the audio, unless they enhance the depiction of the described sounds? **Yes**. The video doesn't add sounds, but the visuals don't relate to the audio. 3. If the video depicts a gunshot, does it do so responsibly and avoid any harmful or misleading depictions of violence or unsafe behavior? **N/A**. The audio doesn't seem to be a gunshot, and the video doesn't depict any violence.

Figure 43: Audio to Video *Score Evaluation* Trustworthiness

Case 30: Video to Audio *Pair Comparison* Creativity

Checklists for Pair Comparison of Creativity:

- Does the audio enhance the video by adding another layer of emotion except dog's whimpering like human's
- Is the audio well-integrated with the visuals, creating a cohesive and immersive experience?

Question: Generate corresponding audio based on the video's visuals.

(a): Input Video

(b): Model A's answer

(c): Model B's answer

Gemini-2.0-Flash Choice: Model B is better.

Checklist Rationale:] Model A: The audio mainly focuses on the dog's whimpering and doesn't add much emotional depth beyond that. While it syncs decently with the dog's actions, it could do more to create a truly immersive experience. Model B: This model successfully blends the dog's whimpers with baby-like sounds and cries, which adds a poignant layer of emotion. The integration of these sounds is well-executed, enhancing the sense of the dog wanting the toy and creating a more immersive experience. The audio quality is also slightly better, with less background noise. Overall, this model's creativity elevates the video's emotional impact, making it a more engaging and satisfying experience. Thus, Model B surpasses Model A in terms of creativity by effectively integrating additional emotional layers to create a more immersive experience.

Figure 44: Video to Audio *Pair Comparison* Creativity

Case 31: Video to Audio *Score Evaluation* Creativity

Checklists for Pair Comparison of Creativity:

- Does the audio enhance the video by adding another layer of emotion except dog's whimpering like human's
- Is the audio well-integrated with the visuals, creating a cohesive and immersive experience?

Question: Generate corresponding audio based on the video's visuals.

(a): Input Video

(b): Model's answer

Gemini-2.0-Flash Score: 1

Checklist Rationale: The model's audio output whimpering does not quite match the dog's action in the video. It does not enhance the video and isn't well-integrated with the visuals.

Figure 45: Video to Audio *Score Evaluation* Creativity

Case 32: Video Edit *Pair Comparison* Clarity

Checklists for Pair Comparison of Clarity:

- Is the aurora in the edited video predominantly red and yellow?
- Is the transition between colors in the aurora smooth and natural-looking?
- Is there any flickering or distortion that makes the aurora difficult to perceive clearly?

Question: Modify the video to show red and yellow aurora paints the night sky over mountain silhouettes.

(a): Input Video

(b): Model A's answer

(c): Model B's answer

Gemini-2.0-Flash Choice: Model B is better.

Checklist Rationale:] Model A's video is overly saturated, making the aurora appear unnatural. The color of the aurora is a mix of pink, yellow, and red, which is far from the requested red and yellow. The mountain silhouette is barely visible in the background. There's also a yellow light source in the horizon that doesn't look like it belongs in an aurora video. Model B's video successfully changes the color of the aurora to a mix of red, pink, and green. While it's not exactly red and yellow, it's much closer to the request than Model A. The colors blend smoothly, and there's no flickering or distortion. The mountain silhouette is also clearly visible. There is a watermark at the bottom right corner. Checklist item A: Model B's aurora is closer to red and yellow than Model A's. Checklist item B: The color transition in Model B's video is smooth and natural. Model A's color transition is too abrupt and unnatural. Checklist item 3: Both videos are free of flickering or distortion. Overall, Model B's video is clearer and better fulfills the prompt, despite the watermark.

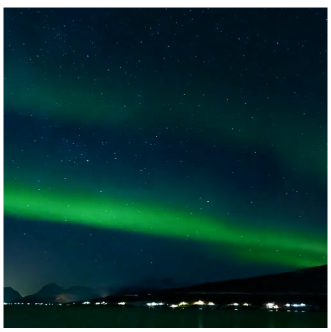
Figure 46: Video Edit *Pair Comparison* Clarity

Case 33: Video Edit *Score Evaluation* Clarity

Checklists for Pair Comparison of Clarity:

- Is the aurora in the edited video predominantly red and yellow?
- Is the transition between colors in the aurora smooth and natural-looking?
- Is there any flickering or distortion that makes the aurora difficult to perceive clearly?

Question: Modify the video to show red and yellow aurora paints the night sky over mountain silhouettes.



(a): Input Video

(b): Model's answer

Gemini-2.0-Flash Score: 1

Checklist Rationale: Checklist item 1: The aurora is green in the edited video, not the requested red and yellow. Checklist item 2: While the color transition is smooth, it is between shades of green, not red and yellow as requested. Checklist item 3: There is no flickering or distortion in the edited video.

Figure 47: Video Edit *Score Evaluation* Clarity