

# UNIM: A Unified Any-to-Any Interleaved Multimodal Benchmark

Yanlin Li<sup>1</sup>, Minghui Guo<sup>1</sup>, Kaiwen Zhang<sup>1</sup>, Shize Zhang<sup>1</sup>, Yiran Zhao<sup>1</sup>,  
Haodong Li<sup>2</sup>, Congyue Zhou<sup>2</sup>, Weijie Zheng<sup>3</sup>, Yushen Yan<sup>2</sup>, Shengqiong Wu<sup>1</sup>,  
Wei Ji<sup>4</sup>, Lei Cui<sup>5</sup>, Furu Wei<sup>5</sup>, Hao Fei<sup>1\*</sup>, Mong-Li Lee<sup>1</sup>, Wynne Hsu<sup>1</sup>

<sup>1</sup>NUS <sup>2</sup>SCUT <sup>3</sup>NTU <sup>4</sup>NJU <sup>5</sup>Microsoft Research

yanlin.li@u.nus.edu, haofei7419@gmail.com

## Abstract

*In real-world multimodal applications, systems usually need to comprehend arbitrarily combined and interleaved multimodal inputs from users, while also generating outputs in any interleaved multimedia form. This capability defines the goal of any-to-any interleaved multimodal learning under a unified paradigm of understanding and generation, posing new challenges and opportunities for advancing Multimodal Large Language Models (MLLMs). To foster and benchmark this capability, this paper introduces the UNIM benchmark, the first Unified Any-to-Any Interleaved Multimodal dataset. UNIM contains 31K high-quality instances across 30 domains and 7 representative modalities: text, image, audio, video, document, code, and 3D, each requiring multiple intertwined reasoning and generation capabilities. We further introduce the UNIM EVALUATION SUITE, which assesses models along three dimensions: Semantic Correctness & Generation Quality, Response Structure Integrity, and Interleaved Coherence. In addition, we propose UNIMA, an agentic baseline model equipped with traceable reasoning for structured interleaved generation. Comprehensive experiments demonstrate the difficulty of UNIM and highlight key challenges and directions for advancing unified any-to-any multimodal intelligence. The project page is <https://any2any-mlm.github.io/unim>.*

## 1. Introduction

MLLMs have rapidly progressed from early architectures centered on visual-language understanding to recent unified frameworks that jointly support both understanding and generation within a single model [1, 10, 12, 17, 19, 37, 41, 43]. This integration has substantially broadened the functional scope of MLLMs, enabling more comprehensive multimodal reasoning and content creation. Yet, studies indicate that simple unification

remains inadequate for achieving true general-purpose multimodal intelligence [11, 14, 24, 36]. A more practical and flexible paradigm, termed *interleaved multimodal learning* [2, 3, 15, 16, 20, 29–31], is needed, where inputs and outputs consist of arbitrarily ordered sequences of modalities (e.g., text and images). Such interleaving better reflects real-world multimodal interactions and is key to building systems that can seamlessly perceive, reason, and respond across diverse modalities. To advance this paradigm, several interleaved multimodal benchmarks have been introduced, including ITLVD-BENCH [22], MMIE [39], CoMM [7], ISG-Bench [5], OpenING [46], which assess models on interleaved text–image understanding and generation tasks. However, these benchmarks still exhibit notable limitations that constrain further development in the field.

The primary limitation lies in their narrow focus on only two modalities, i.e., **text** and **image**, thus failing to capture the full potential of multimodal learning. In contrast, the MLLM landscape has rapidly advanced, with many recent models capable of understanding and generating across a broader spectrum of modalities. Representative examples such as Unified-IO [23], NEXT-GPT [37], and AnyGPT [44], etc., illustrate this growing paradigm of *any-to-any multimodal learning*. In practical applications such as AI assistants, programming copilots, and engineering design systems, users typically interact through complex interleaved multimodal inputs and expect correspondingly diverse multimodal outputs. As shown in Fig. 1, an AI assistant may process intertwined textual instructions, sketches, and images, and respond with a combination of textual reasoning, annotated visuals, or synthesized audio–visual content. Such capabilities represent the next generation of MLLMs, emphasizing a unified treatment of both modality and functionality. However, current interleaved multimodal benchmarks (e.g., MMIE [39]) fall short of capturing this essential *any-to-any interleaved multimodal learning* paradigm within a unified framework.

Beyond the limitation in modality coverage and

\*Corresponding author: Hao Fei.



Figure 1. Illustration of the any-to-any interleaved multimodal paradigm with different real-world application scenarios. Solving any-to-any interleaved multimodal learning requires complex and combined capabilities.







paradigm, existing interleaved multimodal benchmarks suffer from several additional shortcomings. **First**, their evaluation dimensions are neither sufficiently universal nor diverse, in which each data instance typically targets a single, isolated capability, failing to reflect the composite and intertwined nature of real multimodal reasoning. In contrast, authentic any-to-any multimodal learning usually involves multiple competencies within a single instance. For example as in Fig. 1, the AI assistant may simultaneously require comprehension of audio, image, and video inputs, precise image localization, and multimodal generation tasks such as producing images or audio, all demanding temporal understanding, spatial reasoning, and multi-step cognitive planning, etc. Thus, a genuinely unified any-to-any interleaved paradigm should not only expand modality coverage but also emphasize complex reasoning and multi-stage generation. **Second**, current benchmarks mainly concentrate on general-domain scenarios and overlook the diversity of real-world contexts. To accurately model practical applications and rigorously evaluate MLLMs generalization, any-to-any multimodal learning should encompass a broader range of domains and tasks.

To bridge all these gaps, we propose **UNIM**, the first *Unified Any-to-Any Interleaved Multimodal* Benchmark. We curate data from real-world sources, such as publicly open data, online social platforms, and large-scale knowledge bases such as YouTube and Wikipedia. After rigorous manual filtering, annotation, and quality verification, we construct a dataset of 31,026 high-quality instances spanning 30 diverse domains. UNIM

covers 7 representative modalities, i.e., text, image, audio, video, document, code, and 3D, with each instance intentionally designed to involve multiple intertwined tasks and reasoning skills. The benchmark faithfully mirrors real-world any-to-any interleaved input-output patterns within an open-form framework. To establish a structured evaluation protocol, all instances are divided into three difficulty levels ranging from basic to advanced. Existing evaluation metrics (e.g., accuracy) can be largely insufficient for assessing such flexible any-to-any multimodal generation, i.e., often failing to provide objective or consistent measurement. To remedy this, we thus develop the **UNIM EVALUATION SUITE**, which evaluates model performance along three complementary dimensions: (1) *Semantic Correctness & Generation Quality*, (2) *Response Structure Integrity*, and (3) *Interleaved Coherence*. Together, these criteria provide a more comprehensive and accurate assessment of a model’s capability to understand and generate within a unified any-to-any interleaved paradigm.

Benchmarking any-to-any interleaved multimodal learning requires models capable of reasoning over complex multimodal contexts and performing structured, goal-oriented generation rather than simple content synthesis. To this end, we introduce **UNIMA**, a *Unified Any-to-Any Interleaved Multimodal Agentic* model that serves as the baseline system for UNIM. Built upon an agentic framework, UNIMA integrates specialized multimodal encoders and decoders to enable coherent comprehension and generation across heterogeneous modalities. At its core, it devises a *Traceable Evidence Rea-*

Table 1. Comparison with existing interleaved multimodal benchmarks. Inter. Comb.: Interleaved combinations of modalities. Cap. per Instance: Capability per instance. Difficulty Tax.: Difficulty taxonomy.

Benchmarks	Domains	Num.	Inter. Comb.	Cap. per Instance	Eval. Metric	Difficulty Tax.	Any-to	to-Any	Modalities
ITLVD-BENCH [22]	10	815	2	Single	5	✗	✗	✗	
OpenING [46]	8	5,400	4	Single	7	✗	✗	✗	
ISG-Bench [5]	8	1,150	3	Single	4	✗	✗	✗	
CoMM [7]	3	/	4	Single	3	✗	✗	✗	
MMIE [39]	10	20,103	3	Single	7	✗	✗	✗	
<b>UNIM (Ours)</b>	<b>30</b>	<b>31,026</b>	<b>41</b>	<b>Multiple</b>	<b>13</b>	✓	✓	✓	

soning Module that plans, validates, and refines intermediate reasoning steps before producing the final interleaved outputs, thereby improving response, generation quality, and overall interleaved coherence.

Extensive experiments show that current any-to-any MLLMs still struggle considerably on UNIM, highlighting the intrinsic difficulty and challenge of this setting. In-depth analyses further reveal the specific weaknesses of existing models and offer actionable insights for advancing unified interleaved multimodal learning. Meanwhile, UNIMA establishes a strong and interpretable baseline, achieving consistent improvements over prior methods. Overall, this work presents the first comprehensive benchmark for unified any-to-any interleaved multimodal learning, offering a large-scale, high-quality dataset, a principled evaluation suite, and a robust baseline to catalyze future research in this emerging frontier.

## 2. Related Work

Multimodal learning has witnessed rapid advances with the emergence of MLLMs that can integrate text and image modalities for tasks such as captioning and visual question answering [18, 21, 26, 28, 38, 42]. Recent developments [4, 6, 8, 9, 32, 40, 41] have extended these models toward interleaved comprehension and generation, where text and images appear in alternating sequences rather than isolated pairs. Such interleaved modeling is increasingly regarded as a critical capability for next-generation multimodal systems, since it better reflects the natural patterns of human communication. However, existing studies predominantly concentrate on the image-text scenario, and thus fall short of addressing the broader requirement of *any-to-any interleaving* that involves modalities beyond vision and language, such as audio, video, documents, code, and 3D. This paradigm represents a more realistic setting, supported by a growing number of advanced MLLMs, such as NEXT-GPT [37], AnyGPT [44], MIO [34], Spider [13], Codi-2 [27] and ModaVerse [33].

Meanwhile, a variety of datasets and benchmarks have been proposed to support research in interleaved multimodal learning. Relevant resources, such as MMIE [39] and CoMM [7], provide structured benchmarks for interleaved multimodal comprehension and generation. Although these benchmarks represent an important step forward, they remain restricted to image-

text scope and fail to support the evaluation of arbitrary modality combinations. Moreover, these benchmarks only simulate relatively simple language-vision interleaving scenarios, whereas realistic any-to-any interactions are far more complex, especially when involving multiple heterogeneous modalities. This leaves clear gaps for a unified, high-quality benchmark that can systematically assess *any-to-any interleaved multimodal comprehension and generation*, which motivates the development of UNIM. Table 1 illustrates the comparison between UNIM and existing interleaved benchmarks.

## 3. UNIM: Unified Any-to-Any Interleaved Multimodal Benchmark

UNIM takes an open-formed QA format, where input or output is a sequence of interleaved information pieces of any combined modalities, with non-textual pieces represented by placeholder tags (e.g., ‘<image1>’, ‘<video2>’). Also, the sequences may encompass multiple modalities, and each modality contains multiple items (e.g., multiple images). Fig. 1 exemplifies the paradigm. Appendix §B gives a formal task definition.

### 3.1. Data Construction

**Process Pipeline.** We first collect a wide range of multimodal data mainly from three sources: curated samples from public datasets, real-world multimedia content from social media (vlogs, posts), and open resources (forums, websites). Then, we manually design interleaved combinations tailored to different modalities, providing templates for the later construction. During the construction of QA pairs, we design task types and template instances to ensure task diversity and semantic validity. Based on these, GPT-5-mini [25] is employed to generate additional candidate instances for data expansion, so as to simulate more any-to-any multimodal scenarios that are hard to directly retrieve from the Internet.

**Quality Control.** We adopt a two-phase quality control process. All QA pairs are manually reviewed and revised as needed to ensure that the modality placeholder tags comply with task specifications and that the content remains logically consistent. Then, an independent checking process is conducted, where reviewers carefully examine each completed sample to further ensure the reliability and high quality of our dataset. Details re-

Table 2. General statistics of UNIM dataset.

Modality						
	Image	Audio	Video	Document	Code	3D
<b>Num.</b>	22,936	24,963	2,336	3,858	807	420
	(73.9%)	(80.5%)	(7.5%)	(12.4%)	(2.6%)	(1.4%)
Field			Difficulty			
	NS	SS	GA	Easy	Medium	Hard
<b>Num.</b>	10,624	11,574	8,828	10,678	14,088	6,260
	(34.2%)	(37.3%)	(28.5%)	(34.4%)	(45.4%)	(20.2%)

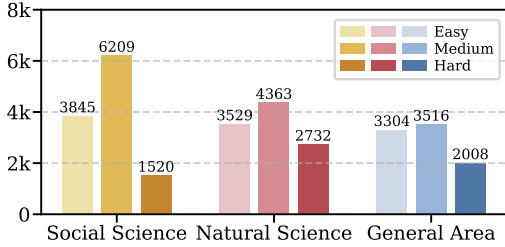


Figure 2. Distribution of different difficulty levels.

garding dataset construction and quality control are provided in Appendix §C.1 and Appendix §C.2.

**Data Statistics.** The resulting UNIM data contains a total of 31,026 instances, covering 30 real-world domains across natural science, social science and general area. Table 2 presents statistics of UNIM. We further define a rule-based progressive difficulty taxonomy, which categorizes instances into three levels: *Easy*, *Medium*, and *Hard*. The distribution of difficulty levels across different fields is shown in Fig. 2. Details regarding the classification criteria and grading process are provided in Appendix §C.3.

### 3.2. Data Core Characteristics

UNIM features the following key aspects: **Any-to-Any Interleaved Modalities.** UNIM covers 7 modalities, supports any-to-any interleaved combinations, faithfully simulating real-world scenarios.

**Universal and Diverse Capabilities.** UNIM evaluates the comprehensive and diverse capabilities of MLLMs. Appendix §C.4 details definitions of each capability.

**Multi-domain Coverage.** UNIM encompasses 30 real-world domains across different fields.

**Multiple Tasks per Instance.** Each instance in UNIM encompasses multiple task objectives ranging from understanding to generation. Detailed descriptions and definitions of tasks can be found in Appendix §C.5.

**Progressive Difficulty.** Instances in UNIM are graded into 3-scale difficulty levels, providing comprehensive evaluation support for both simple and complex interleaved scenarios.

**Large Scale and High Quality.** UNIM contains 31,026 instances, constructed through a rigorous pipeline for high semantic validity and logical coherence.

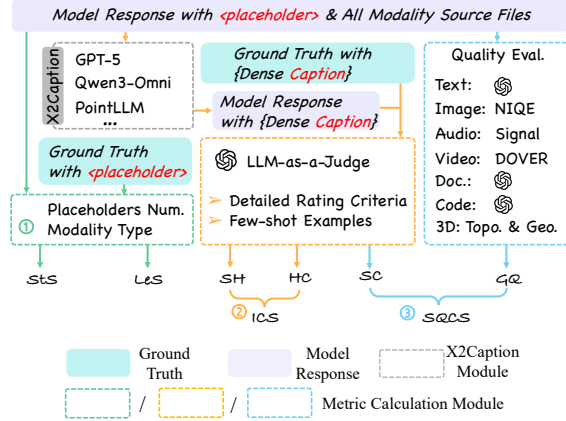


Figure 3. Illustration of the UNIM evaluation suite. ① refers to the calculation process of the StS and LeS (§4.2). ② represents the calculation process of the ICS in Eq. (2). ③ refers to the calculation process of the SQCS; please refer to Eq. (1).

## 4. Evaluation Suite for UNIM

Traditional metrics (e.g., accuracy) focus on single-modality or closed-form matching, which limits their applicability in complex any-to-any interleaved scenarios. We thus rethink evaluation methods and design a systematic evaluation suite for UNIM, with three dimensions: *Semantic Correctness & Generation Quality*, *Response Structure Integrity* and *Interleaved Coherence* (cf., Fig. 3). All these metrics and algorithms details are further provided in Appendix §D.

### 4.1. Semantic Correctness & Generation Quality

**Semantic Correctness (SC)** measures how well the generated output semantically aligns with the reference answer. To ensure fair evaluation across modalities with varying instruction-following capabilities, we convert all modality outputs into comparable caption-like textual representations and employ the LLM-as-a-Judge [45] strategy for measurement. **Generation Quality (GQ)** evaluates the perceptual quality and structural coherence of generated content. Accordingly, we design modality-specific no-reference quality assessment methods to ensure unified and comparable quality metrics across multimodal scenarios. Then, we compose both SC and GQ into **Semantic-Quality Coupled Score (SQCS)** to reflect the overall performance:

$$SQCS = SC \cdot (\eta^{SQCS} + (1 - \eta^{SQCS}) \cdot GQ) . \quad (1)$$

### 4.2. Response Structure Integrity

We devise **Response Structure Integrity** to evaluate whether a model adheres to task-defined structural requirements regarding modality types and item quantities, regardless of semantic or logical correctness. Technically, we break it down into two branches:

**Strict Structure Score (StS)** evaluates the strict structural consistency of a model’s output. StS requires

that the types and quantities of modalities generated in model’s response precisely correspond to those in the ground truth. Any missing or redundant modalities, or discrepancies in the number of modality placeholder tags, are explicitly penalized.

**Lenient Structure Score (LeS)** evaluates the degree of coverage at the modality level. LeS assesses whether the types of modalities generated in model’s response are consistent with those in the ground truth.

### 4.3. Interleaved Coherence

**Interleaved Coherence** is designed to evaluate a model’s ability to maintain logical connectivity and expressive coordination during multimodal integration, measured by **Holistic Coherence (HC)**, which focuses on cross-modal semantic and structural consistency, and **Stylistic Harmony (SH)**, which evaluates consistency in writing style, tone, and visual aesthetics. We adopt the LLM-as-a-Judge [45] to quantify HC and SH, and ultimately use a composite metric over them: **Interleaved Coherence Score (ICS)**:

$$ICS = \eta^{ICS} \cdot HC + (1 - \eta^{ICS}) \cdot SH. \quad (2)$$

### 4.4. Supporting Rate

There might be the common case in UNIM, where an MLLM may not support certain modalities and thus fail on a portion of samples (resulting in overall low metrics), yet it still can achieve top performance on those samples it supports well. To objectively assess a model’s performance, we further introduce the **Supporting Rate** ( $\tau$ ) as a conditional modifier on top of the above three evaluation dimensions, i.e., by distinguishing the model performance between two conditions:  $\mathcal{X}^{abs}$  and  $\mathcal{X}^{rel}$ .  $\mathcal{X}^{abs}$  represents the model’s original capability (i.e., absolute), while  $\mathcal{X}^{rel}$  considers the model’s supporting rate on the entire UNIM:

$$\mathcal{X}^{rel} = \tau \cdot \mathcal{X}^{abs}. \quad (3)$$

## 5. UNIMA: An Agentic Model for UNIM

**Overview.** We build UNIMA to bridge the gap in existing MLLMs that lack strong enough any-to-any multimodal transformation across interleaved modalities. Technically, the overall agentic pipeline operates through three coordinated modules (cf. Fig. 4). The **Receiving Module** converts non-text modalities into *task-conditioned dense caption (TCDC)*, forming a unified text space for subsequent reasoning. The **Traceable Evidence Reasoning (TER) Module** performs structured reasoning by generating, verifying, and refining traceable evidence to construct a logically consistent and verifiable final report that guides subsequent generation. Finally, the **Generating Module**, driven by the verified *final report*, produces interleaved multimodal outputs, completing a loop from understanding to generation.

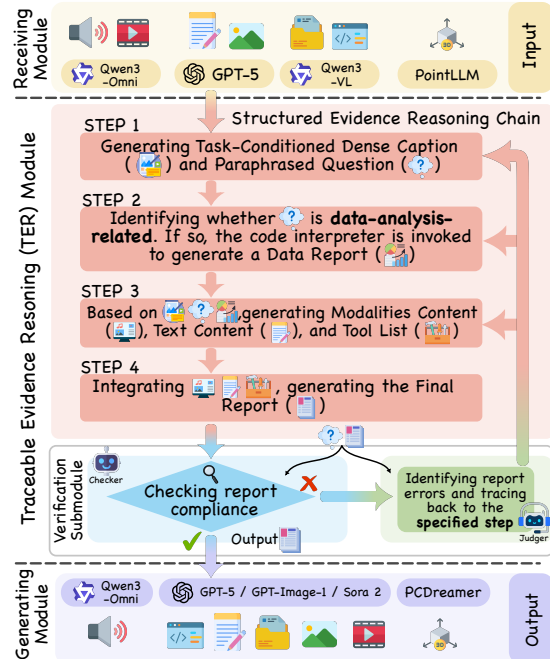


Figure 4. Overview of the UNIMA architecture.

### 5.1. Traceable Evidence Reasoning Module

TER serves as the core inference module of UNIMA, responsible for transforming multimodal inputs passed from the front-end into a structured and verifiable final report. Instead of relying on an implicit Chain-of-Thought [35], TER establishes an explicit evidence-grounded and traceable reasoning chain, enabling each inference step to be justified, verifiable, and revisable within the overall reasoning process. As illustrated in Fig. 4, there are four key steps in the Structured Evidence Reasoning Chain (SERC) within TER.

- **Step 1:** TER leverages the Receiving Module to produce TCDC and a *paraphrased question* from multimodal inputs, primarily enhancing the SC of the content.
- **Step 2:** TER determines whether the task involves data analysis; if so, the code interpreter is invoked to generate a *data report*, improving SC through factual grounding.
- **Step 3:** The extracted information is organized into *modalities content*, *text content*, and a *tool list*: modalities content mainly improves SQCS, text content enhances ICS, and the tool list increases StS/LeS.
- **Step 4:** TER integrates all evidence to produce a draft of the *final report*, enhancing all evaluation dimensions to support reliable multimodal generation.

Notably, in TER, the *Checker* detects factual and logical errors in the report, while the *Judge* backtracks for corrective reasoning. Through iterative *generation-checking-backtracking-regeneration* cycles, TER achieves traceable and reliable multimodal reasoning. More details of UNIMA are provided in Appendix §E.

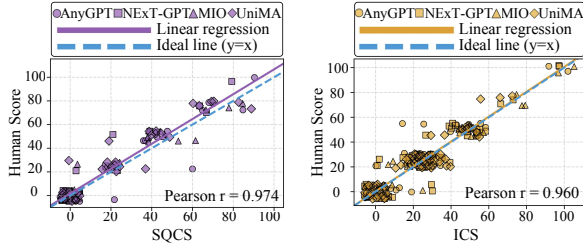


Figure 5. Results for rationality verification of SQCS and ICS.

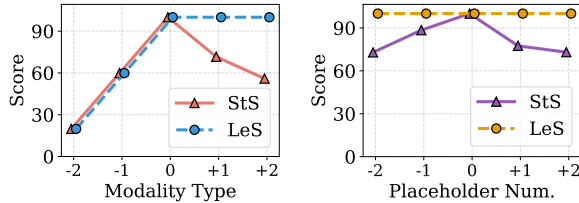


Figure 6. Results for rationality verification of StS and LeS.

## 6. Experiments

We select **AnyGPT** [44], **NExT-GPT** [37], and **MIO** [34] as the representative any-to-any MLLMs.<sup>1</sup>

In our experiments,  $\eta^{\text{SQCS}} = 0.7$ ,  $\eta^{\text{ICS}} = 0.8$ , which help to achieve the best alignment with human evaluations. More settings are detailed in Appendix §F.1.

### 6.1. Rationality of Evaluation Suite

Before we analyze the model performance, here, we study the rationality of our proposed evaluation suite.

► **SQCS and ICS.** We assess the rationality of SQCS and ICS through Pearson correlation analysis between automatic and human evaluations. As shown in Fig. 5, both metrics exhibit a remarkably high degree of linear correlation with human evaluation results, with the Pearson correlation coefficient reaching  $r = 0.974$  for SQCS and  $r = 0.960$  for ICS. These results confirm that the proposed automatic metrics can accurately reflect the performance assessed by human evaluation.

► **StS and LeS.** We introduce two types of controlled perturbations: one varying the modality types and the other varying the number of modality placeholder tags. The perturbation set is constructed based on the ground truth. As shown in Fig. 6, StS decreases when modality types or placeholder tags are added or removed, whereas LeS decreases only when modality types are removed. These results confirm that both types of perturbations induce the anticipated score deviations, validating the soundness and discriminative effectiveness of the proposed metrics.

Detailed procedures and experimental settings of rationality experiments are provided in Appendix §F.3.

<sup>1</sup>We experiment with more MLLMs, where results and analyses are shown in Appendix §F.4.

## 6.2. Main Results and Observations

► **Baseline models exhibit poor performance in terms of absolute metrics on UNIM.** First, according to the results in Table 4, baseline models achieve low SQCS (mostly below 20%), indicating substantial semantic deviations between their responses and ground truth. Then, NExT-GPT and MIO perform quite poorly on StS and LeS (mostly below 5%, cf. Table 3), representing severe limitations of the baselines in matching the required modality coverage and quantity. In addition, baselines perform slightly better on ICS than on the other metrics, but their overall performance remains low (mostly below 50%; cf. Table 5).

► **Notably lower relative scores owing to baselines’ limited support for diverse interleaved input modalities.** All baseline models suffer notably reduced relative scores. As shown in Table 3, the most significant decline appears in AnyGPT in general area, where  $\text{StS}^{\text{abs}}$  drops to 12.5%,  $\text{StS}^{\text{rel}}$  to 9.8%,  $\text{LeS}^{\text{abs}}$  to 16.4%, and  $\text{LeS}^{\text{rel}}$  to 13.6%. In Table 4, the largest degradation is observed for MIO in natural science, with  $\text{SQCS}^{\text{abs}}$  decreasing to 15.9% and  $\text{SQCS}^{\text{rel}}$  to 10.0%. Similarly, in Table 5, the most substantial drop occurs for MIO in natural science, where  $\text{ICS}^{\text{abs}}$  reaches 52.1% and  $\text{ICS}^{\text{rel}}$  31.8%. Their restricted support for diverse interleaved input modalities (cf.  $\tau$  column of Table 4) further indicates that these MLLMs exhibit overall inferior relative performance across tasks.

► **Model performance varies significantly across different fields and domains.** For SQCS, models perform best in social science ( $\text{SQCS}^{\text{abs}}$ : 13.3%-72.7% &  $\text{SQCS}^{\text{rel}}$ : 10.8%-72.7%, cf. Table 4), likely due to the prevalence of common concepts and descriptive reasoning patterns in pretraining data. In contrast, natural science tasks require precise terminology and structured logic rarely seen during training, leading to weaker semantic alignment. For ICS, the general area performs best ( $\text{ICS}^{\text{abs}}$ : 28.1%-69.8%,  $\text{ICS}^{\text{rel}}$ : 20.0%-69.8%, cf. Table 5), as open-domain data better aligns with the model’s training style for smoother coherence and tone. Meanwhile, the higher modality and stylistic complexity of natural and social science tasks make consistent expression harder, lowering ICS performance.

► **UNIMA shows clearly superior performance on UNIM, serving as a reasonable baseline.** UNIMA surpasses the baselines by a large margin, with StS/LeS 2–6 $\times$  higher than AnyGPT and 15–40 $\times$  higher than NExT-GPT and MIO (cf. Table 3). UNIMA achieves an SQCS of around 60% and an ICS approaching 70%, both exceeding those of the best-performing baseline model (cf. Table 4 and Table 5).

Overall, existing any-to-any MLLMs still face significant challenges under unified any-to-any interleaved

Table 3. Assessment results on Response Structure Integrity.

Model	Natural Science				Social Science				General Area			
	StS <sup>abs</sup>	LeS <sup>abs</sup>	StS <sup>rel</sup>	LeS <sup>rel</sup>	StS <sup>abs</sup>	LeS <sup>abs</sup>	StS <sup>rel</sup>	LeS <sup>rel</sup>	StS <sup>abs</sup>	LeS <sup>abs</sup>	StS <sup>rel</sup>	LeS <sup>rel</sup>
AnyGPT [44]	12.9	27.8	12.2	21.4	14.9	16.6	14.5	16.2	12.5	16.4	9.8	13.6
NExT-GPT [37]	2.0	2.2	1.2	1.3	1.3	1.7	1.2	1.5	2.2	2.5	1.4	1.7
MIO [34]	1.3	1.9	0.9	1.3	4.1	5.2	4.0	5.1	3.3	3.8	2.4	2.9
UNI <sup>MA</sup>	50.8	62.2	50.8	62.2	58.1	72.9	58.1	72.9	71.3	84.3	71.3	84.3

Table 4. Performance on Semantic Correctness & Generation Quality and Supporting Rate ( $\tau$ ).

Field	Models	SC	GQ	SQCS <sup>abs</sup>	$\tau$	SQCS <sup>rel</sup>
Natural Science	AnyGPT [44]	13.7	37.9	11.1	90.4	10.7
	NExT-GPT [37]	8.4	23.4	6.2	62.0	2.9
	MIO [34]	19.7	29.1	15.9	59.2	10.0
	UNI <sup>MA</sup>	59.8	79.7	57.3	100	57.3
Social Science	AnyGPT [44]	18.0	23.8	15.5	94.7	14.7
	NExT-GPT [37]	16.8	31.9	13.3	89.0	10.8
	MIO [34]	25.2	32.8	21.4	80.8	16.1
	UNI <sup>MA</sup>	76.2	81.0	72.7	100	72.7
General Area	AnyGPT [44]	19.0	30.1	17.9	90.3	17.2
	NExT-GPT [37]	5.4	30.0	4.4	76.0	3.4
	MIO [34]	24.8	37.5	21.2	71.7	15.2
	UNI <sup>MA</sup>	64.7	83.6	62.2	100	62.2

Table 5. Evaluation results on Interleaved Coherence.

Field	Models	HC	SH	ICS <sup>abs</sup>	ICS <sup>rel</sup>
Natural Science	AnyGPT [44]	39.9	46.3	41.8	38.5
	NExT-GPT [37]	23.5	26.1	24.9	16.3
	MIO [34]	49.4	63.7	52.1	31.8
	UNI <sup>MA</sup>	68.4	71.9	69.1	69.1
Social Science	AnyGPT [44]	31.3	35.3	32.1	29.2
	NExT-GPT [37]	24.5	27.1	21.4	19.0
	MIO [34]	46.3	55.0	51.6	42.0
	UNI <sup>MA</sup>	73.1	76.5	73.8	73.8
General Area	AnyGPT [44]	36.5	41.9	43.6	31.3
	NExT-GPT [37]	27.9	31.1	28.1	20.0
	MIO [34]	68.3	77.7	60.0	45.7
	UNI <sup>MA</sup>	68.7	74.3	69.8	69.8

paradigm. UNI<sup>MA</sup> achieves acceptable performance on UNI<sup>MA</sup>, yet there is still room for further enhancement.

## 7. In-depth Analysis and Discussion

### 7.1. Performance across Various Capabilities

By evaluating all metrics across different capability dimensions, we identify the following trends.

**RQ-1: How do MLLMs perform across diverse capability dimensions?** As shown in Fig. 7, UNI<sup>MA</sup> consistently achieves the highest and most balanced performance across all 10 capability dimensions, while AnyGPT, MIO, and NExT-GPT show uneven capability distribution with sharp declines in tasks requiring complex multimodal interactions such as content editing. This indicates that UNI<sup>MA</sup> attains more stable multimodal fusion and structural control, whereas the baseline models remain fragmented across capabilities, failing to sustain performance once tasks demand higher levels of cross-modal coordination.

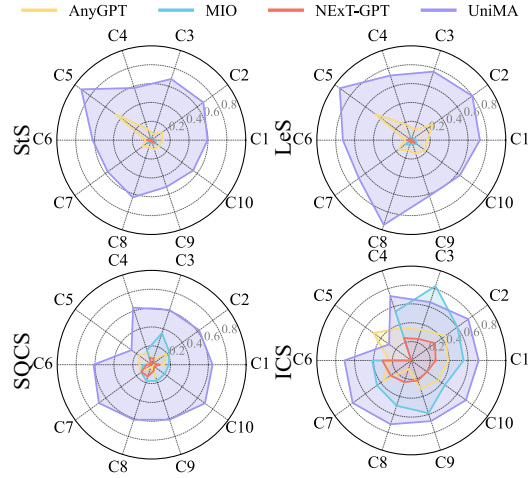


Figure 7. Results across 10 capabilities on UNI<sup>MA</sup>. C1: Perceptual Understanding, C2: Spatial Understanding, C3: Temporal Understanding, C4: Semantic Generation, C5: Content Editing, C6: Creative Expression, C7: Reasoning Capability, C8: Emotional Analysis, C9: Structural Analysis, and C10: Planning Capability. Refer to Appendix §C.4 for details.

**RQ-2: What trends emerge in metric behavior across different capability types?** Across all models, SQCS and ICS remain relatively stable in perception, spatial, and reasoning capabilities, while StS and LeS vary significantly across capabilities. Structurally demanding tasks such as temporal understanding, content editing, and multimodal planning show the lowest scores, with MIO and NExT-GPT both below 10%. These results indicate that current MLLMs can maintain semantic alignment when input semantics are concrete but struggle to preserve modal precision and temporal synchronization in tasks with compositional or sequential dependencies, leading to failures in maintaining structural integrity, semantic correctness, and coherence during generation.

### 7.2. Analysis across Varying Levels of Difficulty

We analyze evaluated MLLMs performance across instances of varying difficulty and find the following.

**RQ-1: Do MLLMs exhibit performance variation consistent with task difficulty?** Only UNI<sup>MA</sup> shows a clear performance gradient aligned with increasing difficulty, whereas AnyGPT, NExT-GPT, and MIO consistently fail, even at the easiest level (cf. Fig. 9). Notably, while UNI<sup>MA</sup>’s StS and LeS decline as difficulty increases, while its SQCS and ICS remain stable, suggesting robustness in semantic reasoning but limitations

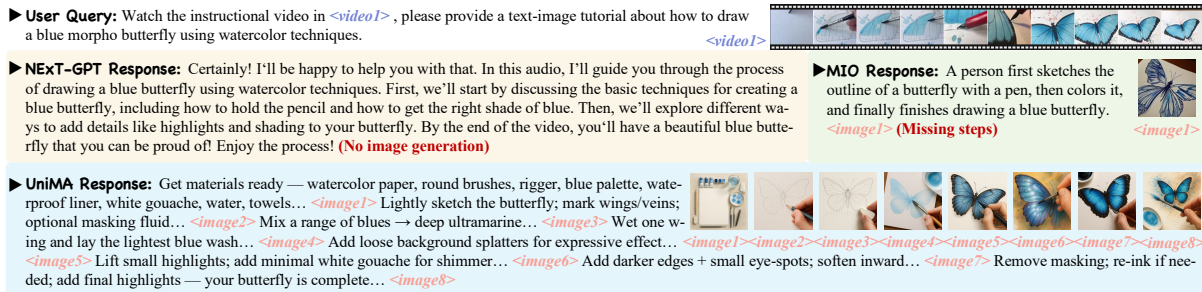


Figure 8. Comparison of model responses in the cross-modal painting instruction task. NEXT-GPT [37] outputs only generic text without image generation, lacking procedural structure. MIO [34] produces a single sketch image but omits key intermediate steps. UNIMA generates a coherent step-by-step text-image workflow aligned with the video, covering the entire painting process.

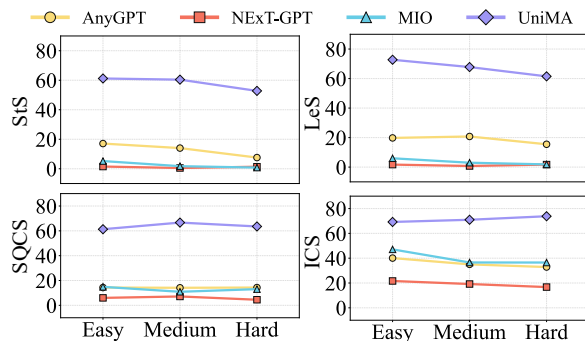


Figure 9. Performance comparison of MLLMs under different difficulty levels across evaluation metrics. Each curve illustrates the performance trend of a model as task complexity increases, where higher values indicate better performance.

in high-order modality interleaving.

**RQ-2: Why do baselines perform equally poorly across all difficulty levels?** Most current MLLMs are insensitive to task difficulty because even easy interleaved tasks already exceed their compositional and alignment capacities. AnyGPT, NEXT-GPT, and MIO show uniformly low scores (mostly below 20%), regardless of difficulty, indicating that even the simplest interleaved tasks in UNIM surpass their multimodal compositional and alignment capacities. Therefore, these models fail to distinguish task complexity.

### 7.3. Ablation Study on UNIMA

Following the previous results where UNIMA consistently outperforms baselines, we further conduct a controlled ablation to quantify each component's contribution. We compare the full model with three variants: removing the TER module, replacing TCDC with a vanilla dense caption, and removing only the verification submodule. As shown in Table 6, removing TER causes the largest StS/LeS drops, confirming its central role in structural adherence and instruction-following. Replacing TCDC leads to moderate SQCS and ICS declines, indicating that task-conditioned evidence mainly enhances semantic grounding and cross-modal coher-

Table 6. Ablation results on UNIMA.

	SQCS	ICS	StS	LeS
<b>UNIMA</b>	<b>85.1</b>	<b>63.4</b>	<b>52.7</b>	<b>82.6</b>
w/o TER	72.9 (-12.2)	56.6 (-6.8)	16.4 (-36.3)	21.8 (-60.8)
w/o TCDC	78.4 (-6.7)	57.7 (-5.7)	46.2 (-6.5)	82.1 (-0.5)
w/o Ver.	72.9 (-12.2)	54.7 (-8.7)	38.3 (-14.4)	66.8 (-15.8)

ence. Removing the verification submodule produces notable degradation across all metrics, demonstrating that checking-backtracking-regeneration design is critical for reliable interleaved outputs. More model ablation details are provided in Appendix §E.4.

### 7.4. Qualitative Case Study

Finally, to better illustrate the performance gap between existing baselines and our UNIMA, we present a visualization of case study in Fig. 8, comparing the outputs of UNIMA with representative models (MIO, NEXT-GPT). Compared with existing models that struggle to achieve synchronization between text and image modalities, UNIMA constructs a complete text-image instructional workflow. Its generated results maintain consistency in temporal order, modality coverage and stylistic coordination, enabling the model to produce semantically grounded and structurally clear instructional content, fully demonstrating UNIMA's advances in multimodal reasoning and coordinated generation.

## 8. Conclusion

This paper presents UNIM, for the first time benchmarking the *Unified Any-to-Any Interleaved Multimodal Learning*. UNIM provides a large-scale, high-quality dataset covering 7 modalities and 30 real-world domains, involving complex task-solving capabilities, together with a principled evaluation suite for assessing comprehension and generation in complex interleaved scenarios. Furthermore, an agentic baseline model UNIMA is deliberately designed to better benchmark the task. Our experiments show that existing MLLMs still face significant challenges under this unified any-to-any Interleaved paradigm, revealing key limitations and opportunities for future progress.

## Acknowledgement

This work is supported by the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001).

## References

- [1] Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, et al. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*, 2025. 1
- [2] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. Minigt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*, 2024. 1
- [3] Roman Bachmann, Oğuzhan F Kar, David Mizrahi, Ali Garjani, Mingfei Gao, David Griffiths, Jiaming Hu, Afshin Dehghan, and Amir Zamir. 4m-21: An any-to-any vision model for tens of tasks and modalities. In *Proceedings of the NIPS*, pages 61872–61911, 2024. 1
- [4] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 3
- [5] Dongping Chen, Ruoxi Chen, Shu Pu, Zhaoyi Liu, Yanru Wu, Caixi Chen, Benlin Liu, Yue Huang, Yao Wan, Pan Zhou, et al. Interleaved scene graphs for interleaved text-and-image generation assessment. *arXiv preprint arXiv:2411.17188*, 2024. 1, 3
- [6] Jiu hai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models — architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025. 3
- [7] Wei Chen, Lin Li, Yongqi Yang, Bin Wen, Fan Yang, Tingting Gao, Yu Wu, and Long Chen. Comm: A coherent interleaved image-text dataset for multimodal understanding and generation. In *Proceedings of the CVPR*, pages 8073–8082, 2025. 1, 3
- [8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the CVPR*, pages 24185–24198, 2024. 3
- [9] Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jingsheng Wang, Wenxuan Wang, et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025. 3
- [10] Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *Proceedings of the NIPS*, pages 57207–57239, 2024. 1
- [11] Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, et al. On path to multimodal generalist: General-level and general-bench. In *Proceedings of the ICML*, 2025. 1
- [12] Qingpei Guo, Kaiyou Song, Zipeng Feng, Ziping Ma, Qinglong Zhang, Sirui Gao, Xuzheng Yu, Yunxiao Sun, Tai-Wei Chang, Jingdong Chen, et al. M2-omni: Advancing omni-mlm for comprehensive modality support with competitive performance. *arXiv preprint arXiv:2502.18778*, 2025. 1
- [13] Jinxiang Lai, Jie Zhang, Jun Liu, Jian Li, Xiaocheng Lu, and Song Guo. Spider: Any-to-many multimodal llm. *arXiv preprint arXiv:2411.09439*, 2024. 3
- [14] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214, 2024. 1
- [15] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 1
- [16] Huayang Li, Siheng Li, Deng Cai, Longyue Wang, Lemao Liu, Taro Watanabe, Yujiu Yang, and Shuming Shi. Textbind: Multi-turn interleaved multimodal instruction-following in the wild. In *Findings of the ACL*, pages 9053–9076, 2024. 1
- [17] Han Li, Xinyu Peng, Yaoming Wang, Zelin Peng, Xin Chen, Rongxiang Weng, Jingang Wang, Xunliang Cai, Wenrui Dai, and Hongkai Xiong. Onecat: Decoder-only auto-regressive model for unified understanding and generation. *arXiv preprint arXiv:2509.03498*, 2025. 1
- [18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the ICML*, 2023. 3
- [19] Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Zichun Liao, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Omniflow: Any-to-any generation with multimodal rectified flows. In *Proceedings of the CVPR*, pages 13178–13188, 2025. 1
- [20] Kevin Qinghong Lin, Pengchuan Zhang, Difei Gao, Xide Xia, Joya Chen, Ziteng Gao, Jinheng Xie, Xuhong Xiao, and Mike Zheng Shou. Learning video context as interleaved multimodal sequences. In *Proceedings of the ECCV*, pages 375–396, 2024. 1
- [21] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the NIPS*, 2023. 3
- [22] Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiabin Zhang, and Lifu Huang. Holistic evaluation for interleaved text-and-image generation. *arXiv preprint arXiv:2406.14643*, 2024. 1, 3
- [23] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 1

- [24] Alhassan Mumuni and Fuseini Mumuni. Large language models for artificial general intelligence (agi): A survey of foundational principles and approaches. *arXiv preprint arXiv:2501.03151*, 2025. 1
- [25] OpenAI. Introducing gpt-5, 2025. 3
- [26] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 3
- [27] Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. Codi-2: In-context interleaved and interactive any-to-any generation. In *Proceedings of the CVPR*, pages 27425–27434, 2024. 3
- [28] Qwen Team. Qwen3-vl-30b-a3b-instruct model card, 2025. 3
- [29] Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhui Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, et al. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *arXiv preprint arXiv:2401.10208*, 2024. 1
- [30] Alex Jinpeng Wang, Linjie Li, Kevin Qinghong Lin, Jianfeng Wang, Kevin Lin, Zhengyuan Yang, Lijuan Wang, and Mike Zheng Shou. Cosmo: Contrastive streamlined multimodal model with interleaved pre-training. *arXiv preprint arXiv:2401.00849*, 2024.
- [31] Hanlin Wang, Zhan Tong, Kecheng Zheng, Yujun Shen, and Limin Wang. Contextual ad narration with interleaved multimodal sequence. In *Proceedings of the CVPR*, pages 8372–8383, 2025. 1
- [32] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yuezhe Wang, Zhen Li, Qiyang Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 3
- [33] Xinyu Wang, Bohan Zhuang, and Qi Wu. Modaverse: Efficiently transforming modalities with llms. In *Proceedings of the CVPR*, pages 26606–26616, 2024. 3
- [34] Zekun Wang, King Zhu, Chunpu Xu, Wangchunshu Zhou, Jiaheng Liu, Yibo Zhang, Jiashuo Wang, Ning Shi, Siyu Li, Yizhi Li, et al. Mio: A foundation model on multimodal tokens. *arXiv preprint arXiv:2409.17692*, 2024. 3, 6, 7, 8
- [35] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the NIPS*, pages 24824–24837, 2022. 5
- [36] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the CVPR*, pages 12966–12977, 2025. 1
- [37] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NEXT-GPT: Any-to-any multimodal LLM. In *Proceedings of the ICML*, pages 53366–53397, 2024. 1, 3, 6, 7, 8
- [38] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 3
- [39] Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, et al. Mmie: Massive multimodal interleaved comprehension benchmark for large vision-language models. *arXiv preprint arXiv:2410.10139*, 2024. 1, 3
- [40] Shitao Xiao, Yuezhe Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the CVPR*, pages 13294–13304, 2025. 3
- [41] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 1, 3
- [42] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 3
- [43] Hanrong Ye, De-An Huang, Yao Lu, Zhiding Yu, Wei Ping, Andrew Tao, Jan Kautz, Song Han, Dan Xu, Pavlo Molchanov, et al. X-vila: Cross-modality alignment for large language model. *arXiv preprint arXiv:2405.19335*, 2024. 1
- [44] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. In *Proceedings of the ACL*, pages 9637–9662, 2024. 1, 3, 6, 7
- [45] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the NIPS*, pages 46595–46623, 2023. 4, 5
- [46] Pengfei Zhou, Xiaopeng Peng, Jiajun Song, Chuanhao Li, Zhaopan Xu, Yue Yang, Ziyao Guo, Hao Zhang, Yuqi Lin, Yefei He, et al. Opening: A comprehensive benchmark for judging open-ended interleaved image-text generation. In *Proceedings of the CVPR*, pages 56–66, 2025. 1, 3