Unison: Benchmarking Unified Multimodal Models via Synergistic Understanding and Generation

Jinyu Liu [1 2] **Xincheng Shuai** [3] **Henghui Ding** [3] **Yu-Gang Jiang** [1 2]

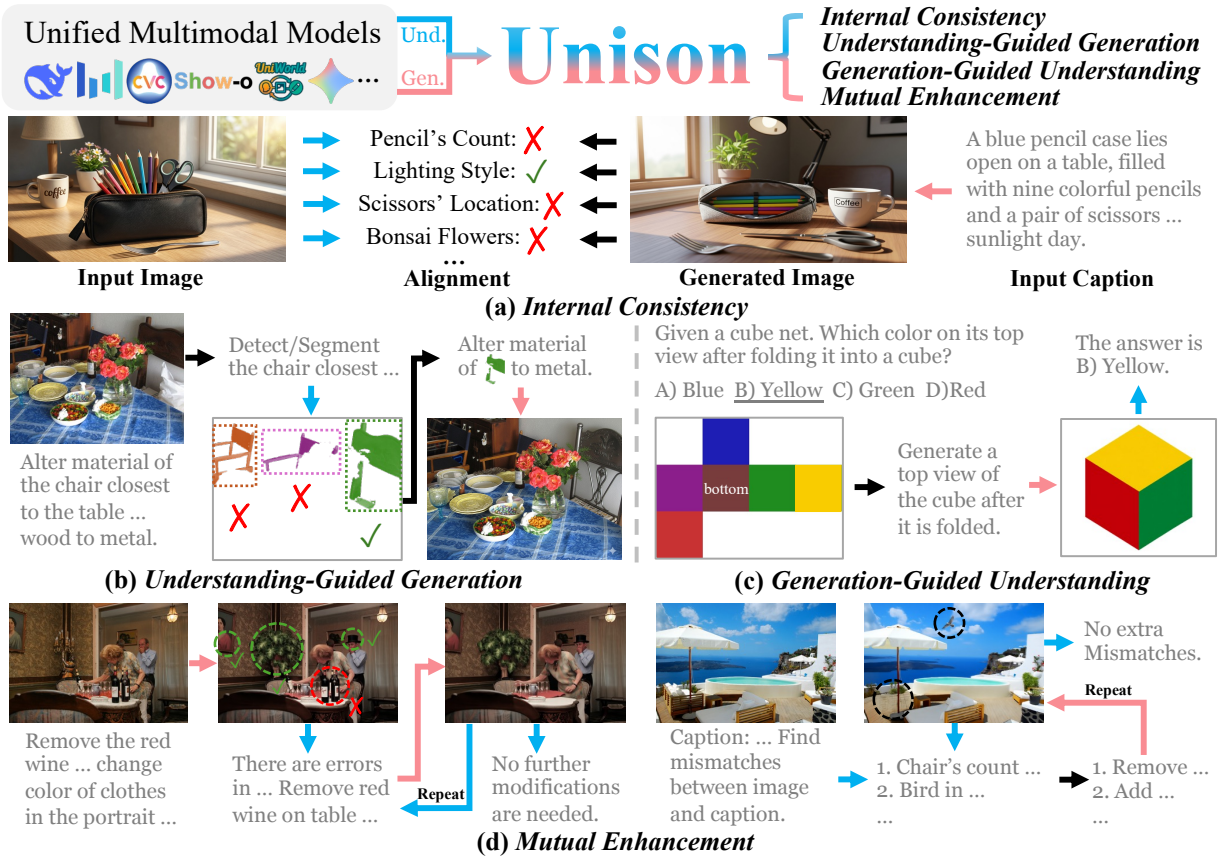


Figure 1. **Overview of Unison**. Unison decomposes unified models' capabilities into four dimensions: **(a)** ***Internal Consistency*** assesses internal alignment between understanding and generation. **(b)** ***Und.-Guided Gen.*** estimates model's ability to leverage comprehension guiding contextually generation. **(c)** ***Gen.-Guided Und.*** focuses on how synthesized outputs assist understanding tasks, and **(d)** ***Mutual Enhancement*** evaluates the iterative synergy where understanding identifies generation errors and vice versa, enabling mutual refinement.

Abstract

Unified multimodal models capable of both understanding and generation have achieved remarkable strides. However, despite their unified designs, existing evaluations typically assess understanding and generation capabilities in isolation, overlooking the synergy between comprehension and generation. To bridge this gap, we introduce **Unison**, a comprehensive benchmark comprising 2,169 high-quality unified task samples, designed to evaluate joint understanding and generation in unified multimodal models. Unison offers three key strengths: **1) Comprehensive Dimensions**: Unison encompasses *internal consistency*, *understanding-guided genera-*

[1]Institute of Trustworthy Embodied AI, Fudan University, Shanghai, China [2]Shanghai Key Laboratory of Multimodal Embodied AI, Shanghai, China [3]Institute of Big Data, College of Computer Science and Artificial Intelligence, Fudan University, Shanghai, China. Correspondence to: Yu-Gang Jiang <ygj@fudan.edu.cn>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

tion, *generation-guided understanding*, and *mutual enhancement* to enable holistic evaluation. **2) Diagnostic Evaluation**: it provides both unified and decoupled tracks for understanding and generation, allowing fine-grained attribution of failure modes and quantitative analysis of the gains from unified modeling. **3) Human Alignment**: we also introduce Unison-Judge, an evaluation model well aligned with human judgments to ensure reliable assessment. Based on systematic evaluations of state-of-the-art models on Unison, we uncover critical limitations in current unified multimodal systems and highlight promising directions for future research. Unison will be publicly available at https://github.com/FudanCVL/Unison.

1. Introduction

Multimodal large language models (MLLMs) have advanced rapidly in recent years, demonstrating impressive comprehension capabilities in complex reasoning tasks. On the other hand, visual generative models, particularly diffusion-based, have achieved unprecedented fidelity and controllability in synthesizing images and videos. Motivated by their success, unified multimodal models (UMMs) have emerged as a rapidly advancing research direction, which aim to integrate complementary understanding (textual output) and generation (visual output) capabilities within a single framework. Current UMMs are increasingly moving beyond composition of separate understanding and generative models toward natively unified designs that support shared representations across both perception and synthesis.

Nowadays, UMMs have demonstrated strong performance in multimodal understanding, exhibiting notable complex reasoning capabilities on benchmarks such as MMBench (Liu et al., 2024) and MMMU (Yue et al., 2024). They also achieve compelling results in compositional image generation, producing fine-grained, instruction-aligned outputs, as evaluated on GenEval (Ghosh et al., 2023), DPG-Bench (Hu et al., 2024) and OneIG-Bench (Chang et al., 2025). However, existing evaluations remain confined to capability-specific tasks. There is a critical need for a benchmark to holistically evaluate UMMs on unified tasks requiring both understanding and generation capabilities.

To this end, we propose **Unison**, a comprehensive unified benchmark designed to holistically evaluate UMMs through synergistic understanding and generation. Our work is motivated by three foundational questions: *1) Do current UMMs truly integrate understanding and generation capabilities?* Prior work (Li et al., 2025a) points out that most UMMs suffer from a performance trade-off between these capabilities due to conflicting visual tokenization strategies. UniGen (Tian et al., 2025) leverage understanding ability as a self-verifier to assess the quality of generated images by evaluating image-text semantic coherence, suggesting a critical insight: If a UMM correctly comprehends an input image, can it generate a new image consistent with its own semantic interpretation? We define this notion as ***internal consistency***, as shown in Figure 1 (a), to evaluate the alignment between understanding and generation. *2) For challenging tasks, what is the synergistic effect of UMMs in leveraging their own reasoning to guide generation, and vice versa?* Bagel (Deng et al., 2025) claims that reasoning before visual generation may enhance the final output in multimodal tasks. By using this approach to filter training data, Bagel achieves superior performance. Motivated by this, we design ***understanding-guided generation*** and ***generation-guided understanding***, as shown in Figure 1 (b) and (c), to evaluate the synergistic effect in UMMs. *3) Can these dual capabilities synergistically interact to enhance the performance of generation or understanding tasks?* X-Omini (Geng et al., 2025) shows that shared autoregressive modeling between image and text generation, augmented by reward models, reduces cumulative errors, demonstrating that understanding and generation can mutually reinforce each other. Thus, we propose ***mutual enhancement*** to assess UMMs' self-refinement capability through iterative cycles of alternating multimodal understanding and image generation, as shown in Figure 1 (d).

Through the Unison framework, we observe that while certain UMMs achieve state-of-the-art performance on specialized understanding or generation benchmarks, they often fall short in unified tasks requiring joint comprehension and synthesis. Subsequent experiments and analysis further reveal key limitations of current UMMs and point to promising directions for future research.

In summary, our main contributions are as follows:

- In this work, we explore a promising yet underresearched space for evaluating unified multimodal models via synergistic understanding and generation. Based on this, we introduce Unsion, consisting of 2,169 human-validated unified task samples, which encompasses *internal consistency*, *understanding-guided generation*, *generation-guided understanding*, and *mutual enhancement* to enable unified evaluation.
- Unison supports both joint and decoupled assessment of understanding and generation within the same framework, providing interpretable and fine-grained analysis.
- Extensive experiments demonstrate that Unison-Judge exhibits strong alignment with human judgments. Additionally, we provide valuable insights into current UMMs' limitations and future directions.

2. Related works

Unified Multimodal Models. Recent years have witnessed remarkable advancement in both multimodal large language models (MLLMs) and visual generative models, leading to increasing research efforts toward unifying these capabilities within a single architecture. This pursuit has given rise to unified multimodal models (UMMs) (Chen et al., 2025b; Li et al., 2025b; Wu et al., 2025d; Xie et al., 2024; 2025; Ge et al., 2024; Xiao et al., 2025; Chen et al., 2025a; Deng et al., 2025; Huang et al., 2025; Wu et al., 2025c; Qu et al., 2025; Zhou et al., 2025; Lin et al., 2025), which generally fall into three categories: auto-regressive models, diffusion-based models and hybrid models. Auto-regressive models, such as SEED-X (Ge et al., 2024) and UniWorld (Lin et al., 2025), typically tokenize both visual and text inputs into discrete sequencial tokens using transformer-based architectures, enabling seamless multimodal reasoning and generation. On the other hand, diffusion-based unified models such as D-DiT (Li et al., 2025b), which introduces a dual-branch diffusion framework that jointly generates text and images, facilitating coherent and controllable synthesis of aligned image-text pairs starting from pure noise, enabling synchronized generation across visual and textual modality. In addition, hybrid approaches (Xie et al., 2024; Zhou et al., 2025; Deng et al., 2025) integrate auto-regressive and diffusion-based components within a unified framework. These methods aim to preserve strong language reasoning while enabling high-fidelity visual generation, but they often increase training complexity and computational cost.

Multimodal Understanding Benchmarks have evolved from basic perception toward complex reasoning. Early datasets such as Flickr30k (Plummer et al., 2015) and MS COCO Captions (Chen et al., 2015) focused on image–text alignment, while VQA-style benchmarks (Antol et al., 2015; Goyal et al., 2017; Singh et al., 2019) introduced free-form question answering requiring scene-level comprehension. More recent works evaluate diverse multimodal capabilities: MMBench (Liu et al., 2024) emphasizes cross-lingual perception, MMMU (Yue et al., 2024) targets domain-specific reasoning, and SEED-Bench (Li et al., 2024) assesses fine-grained skills with external knowledge integration.

Image Generation & Editing Benchmarks. Previous image generation relied on automatic metrics like Inception Score (IS) (Salimans et al., 2016) for basic image-only quality assessment. Frechet Inception Distance (FID) (Heusel et al., 2017) and CLIPScore (Hessel et al., 2021) estimates image-text alignment in feature level. However, contemporary works (Ghosh et al., 2023; Huang et al., 2023; Hu et al., 2024) reveal significant gaps between these metrics and human judgment, particularly for complex scenes. To address this, benchmarks now emphasize compositional generation and semantic fidelity. GenEval (Ghosh et al., 2023) and T2I-CompBench (Huang et al., 2023) evaluate attributes like object color, counting, and spatial relationships. DPG-Bench (Hu et al., 2024) focuses on generation from dense prompts, and OneIG-Bench (Chang et al., 2025) supports multilingual and stylized evaluation. For image editing, where models modify images according to textual instructions, InstructPix2Pix (Brooks et al., 2023) and MagicBrush (Zhang et al., 2023) assess the correctness of local and global edits. AnyEdit (Yu et al., 2025) introduces a dedicated evaluation set covering 25 editing types for real-world applicability, while ImgEdit (Ye et al., 2025) provides a large-scale benchmark with 1.2 million high-quality edit pairs, emphasizing instruction adherence and visual realism.

3. Unison

In this work, we propose Unison, a comprehensive benchmark designed to holistically assess the unified capabilities of UMMs, focusing on the synergy between understanding and generation. To facilitate systematic evaluation, we decompose unified capability into four dimensions: *internal consistency*, *understanding-guided generation*, *generation-guided understanding*, and *mutual enhancement*.

3.1. Internal Consistency

Internal consistency, referring to the alignment of UMM's semantic encoding of understanding and generation within its internal representations, is designed by leveraging fine-grained compositional understanding of images alongside text-to-image tasks. Both two tasks fundamentally require holistic comprehension of all input information. In fine-grained compositional understanding, the model must analyze visual elements such as objects, attributes and spatial relationships, typically achieved through tasks like Visual Question Answering (VQA). On the other hand, text-to-image generation requires instruction-following and compositional reasoning to accurately synthesize images. Without this foundational consistency, the model struggles to perform complex tasks effectively, as errors in basic alignment can propagate and hinder performance.

In particular, given an image-caption pair (I, C), a set of questions $Q = \{q_i^{ic}\}_{i=1}^n$ and a set of attributes $A = \{a_i\}_{i=1}^n$, where attributes contain *object, color, texture, count, light text rendering* and *spatial relation*. The image-caption pair (I, C) is manually aligned, where each attribute a_i explicitly represented in the corresponding question q_i. For understanding, UMM perform VQA on image I to get answer $s_i^{und} = \mathcal{M}(I, q_i^{ic})$. For generation, we first prompt UMM to perform text-to-image synthesis to obtain a generated image $I_{t2i} = \mathcal{M}(C)$, then apply a dedicated judge model $\mathcal{J}$ to answer the same question on I_{t2i}, yielding $s_i^{gen} = \mathcal{J}(I_{t2i}, q_i^{ic})$. Finally, we compare answers s_i^{und} and

Table 1. Benchmark Comparison. MU denotes multi-modal understanding, IG denotes image generation, IE denotes image editing, IC denotes internal consistency, UF denotes unified understanding and generation, Det denotes detection, Seg denotes segmentation, HE denotes hybrid editing, HT denotes hybrid type, RI denotes real image, SI denotes synthetic image.

Benchmark	Evaluation					Localization		Editing		Scenario		Multi-Turn	Judge Model	Task Type
	MU	IG	IE	IC	UF	Det	Seg	HE	HT	RI	SI			
MMBench (Liu et al., 2024)	✓	✗	✗	✗	✗	✗	✗	✗	0	✓	✗	✗	✗	6
SEED-Bench (Li et al., 2024)	✓	✗	✗	✗	✗	✗	✗	✗	0	✓	✗	✗	✗	6
MMMU (Yue et al., 2024)	✓	✗	✗	✗	✗	✗	✗	✗	0	✓	✓	✗	✗	6
GenEval (Ghosh et al., 2023)	✗	✓	✗	✗	✗	✗	✗	✗	0	✗	✓	✗	✗	1
DPG-Bench (Hu et al., 2024)	✗	✓	✗	✗	✗	✗	✗	✗	0	✗	✓	✗	✗	1
OneIG-Bench-EN (Chang et al., 2025)	✗	✓	✗	✗	✗	✗	✗	✗	0	✗	✓	✗	✗	6
AnyEdit (Yu et al., 2025)	✗	✗	✓	✗	✗	✗	✗	✗	0	✓	✓	✗	✗	5
ImageEdit (Ye et al., 2025)	✗	✗	✓	✗	✗	✗	✗	✓	6	✓	✗	✓	✓	3
Unison (ours)	✓	✓	✓	✓	✓	✓	✓	✓	**11**	✓	✓	✓	✓	**14**

s_i^{gen} for each attribute-aligned question q_i calculating final internal consistency score $Score_{ic}$:

$$Score_{ic} = \sum_{i=1}^{n} f(a_i, s_i^{und}, s_i^{gen}))/n, \tag{1}$$

where $f(a_i, s_i^{und}, s_i^{gen}) \in \{0, 1\}$ returns 1 if s_i^{und} and s_i^{gen} are consistent with respect to attribute a_i, and 0 otherwise.

Notably, extreme failure cases can lead to *spurious consistency*, where a UMM may misidentify a critical attribute (e.g., an image-text pair depicts five people, but the model perceives six and subsequently generates an image showing six people). Since consistency derived from a faulty premise is meaningless, our evaluation protocol assigns a score of zero to such cases. Our strict criterion ensures that high scores only represent faithful consistency, effectively penalizing hallucinations or inherent model biases.

3.2. Understanding-Guided Generation

In this part, we leverage the reasoning capability of UMM as prior to guide contextually accurate generation for synergy's effect assessment. In short, we prompt the UMM to autonomously reason precise edit parts from complex instructions then executing generation. Based on comparison with direct image manipulation, understanding-guided generation provides a systematic analysis for causal impact of a model's comprehension on its generative fidelity.

Specifically, given an image I and a complex editing instruction E. The model first comprehends E and reason about I to output the target region $\mathcal{R}$ via bounding box/segmentation mask. Then obtaining its own localization as conditional input, the model executes the editing operation to get final edited image I_{ie}. We average the localization score and image editing score to get final score $Score_{ugg}$:

$$Score_{ugg} = (IoU(\mathcal{R}, \mathcal{R}_{gt}) + \mathcal{J}(I, I_{ie})) \,/\, 2, \tag{2}$$

where $\mathcal{R}_{gt}$ is the ground truth target region, $IoU(,)$ returns Intersection over Union score, $\mathcal{J}$ is a MLLM-based rating model to provide scores for image editing.

3.3. Generation-Guided Understanding

For this paradigm, we employ visual output as a proxy for complex reasoning, exploring the untapped potential of UMMs. We formalize this as generation-guided understanding, which requires a tight coupling between generative synthesis and the comprehension of implicit structural constraints. For example, given a cub net, determining what color of its top face after it is folded into a cube requires visual imagination, which challenges even human cognition. We reveal that UMMs can leverage their generative priors to simulate these transformations, providing an intuitive visual reference that facilitates the resolution of otherwise intractable reasoning problems.

We formalize this task as visual question answering, where an intermediate image generation step is employed to synthesize relevant visual contexts. Based on an image I (2D spatial, 3D spatial) or description D (complex relation), as illustrated in Figure 2, the task is to select the ground-truth option O_{gt} from a set of four candidates in response to a question q^{ggu}. To achieve this, UMM first performs an intermediate generative phase explicitly conditioned on the question's context. Subsequently, the model reasons over the synthesized image to derive the final output. To evaluate this process, we define an understanding score $S_u^{ggu} \in \{0, 1\}$ based on the accuracy of the predicted option. The generation score S_g^{ggu} follows the principles of internal consistency. Finally, the holistic performance is measured by the average score $Score_{ggu} = (S_u^{ggu} + S_g^{ggu})/2$.

3.4. Mutual Enhancement

In this part, we assess self-refinement ability on capability-specific tasks of UMMs by alternately motivating them alternately comprehend and regenerate their own outputs, or vice versa, referring to *mutual enhancement*. Specifically, the model first generates an initial output, justifies the coherence between the generated content and the input instruction, and then summarizes a textual operation to produce optimized outputs. This iterative loop tests whether UMM can identify and reinforce its own errors without external supervision.

For clarity, this process follows a four-step loop (up to R rounds): 1. Generation: The model produces an initial textual/visual output from the input instruction. 2. Evaluation: The model acts as an evaluator, judging if the current output satisfies the input instruction. 3. Refinement: If misalignments are found, the model needs to generate a specific refinement instruction to reproduce a new output. Then re-evaluates. 4. Termination: The loop stops when the model judges no further conflicts, or reaches the maximum round.

Self-Refining Image Editing extends standard editing by incorporating iterative feedback loops.

Formally, assuming there are at most R execution rounds. Given an initial instruction $T_1 \in \{T_r\}_{r=1}^{R}$, the model generates an edited image $I_1^{me_ie} \in \{I_r^{me_ie}\}_{r=1}^{R}$. For subsequent rounds $r \in \{2, \ldots, R\}$, the model evaluates if $I_{r-1}^{me_ie}$ fully aligns with the initial instruction T_1, If misaligned, it will be prompted to generate a refinement instruction T_r and a corresponding image $I_r^{me_ie}$, otherwise, the refinement loop terminates early. Moreover, we define self-refining gain ΔS_g to quantify the cumulative improvement. This metric represents the absolute increase achieved by final refined image I_k compared to I_1. We directly use editing score at k round as final score $Score_{me_ie}$.

Self-Refining Multimodal Understanding focuses on ability to diagnose and rectify cross-modal inconsistencies. Unlike self-refining image editing, this task begins with a pre-existing image-caption pair (I_1, C_1) that contain misalignments. The model must critically recognize these discrepancies to drive the refinement process.

In each round r, the model performs a "detect-and-correct" cycle, it first identifies semantic mismatches between the current image $I_r^{me_mu}$ and the original caption C_1, then formulates a corrective instruction T_r to produce a better-aligned image $I_{r+1}^{me_mu}$. This loop terminates when the model perceives no further conflicts or reaches the maximum R rounds. The final performance $Score_{me_mu}$ is quantified by measuring the alignment between the model-identified mismatches and human-annotated ground truth. Similar to the image editing task, we define self-refining gain ΔS_u as the absolute improvement from initial state to final round.

The overall mutual enhancement performance is defined as $Score_{me} = (Score_{me_ie} + Score_{me_mu})/2$.

4. Dataset & Evaluation

Unison supports understanding, generation, and unified assessment. Data distribution is shown in Figure 2. Section 4.1 details the data structure and compares Unison with existing benchmarks. Section 4.2 defines evaluation metrics. Section 4.3 introduces Unison-Judge, an evaluation model closely aligned with human preference for evaluation.

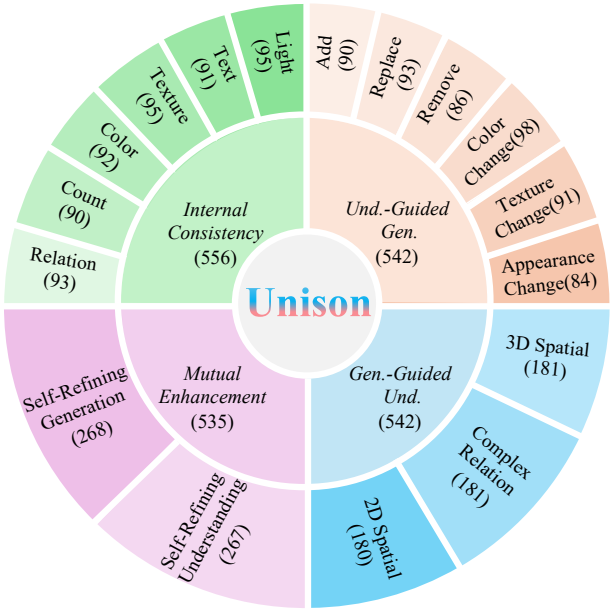


Figure 2. Data distribution of Unison. Unison comprises four core categories, each designed to evaluate a specific capability.

4.1. Benchmark Construction

To ensure a high-quality, rigorous evaluation framework, our benchmark is constructed through a collaborative pipeline involving state-of-the-art foundation models, followed by three weeks of verification and refinement by ten people.

Data Corpus and Image Generation. Our benchmark construction begins with a diverse data curation process tailored to multi-dimensional capabilities. For general vision-language tasks, we derive comprehensive captions from predefined attributes following GenEval (Ghosh et al., 2023) protocols. For complex scene understanding, we select a subset from RefCOCOg (Mao et al., 2016) characterized by multi-object clusters and significant occlusions. High-quality aesthetic samples are further sourced from LAION-Aesthetics (Schuhmann et al., 2022) by filtering for scores exceeding 5.0. To generate the visual content, we employ Qwen-Image (Wu et al., 2025a) for text-to-image synthesis, while spatial reasoning images are generated by mapping logical constraints of unfolded diagrams onto 3D structures using OpenCV. For image editing tasks, we define 11 compositional types by combining four meta-operations: *Alter*, *Remove*, *Replace*, and *Add*, ensuring each instruction contains at least two edits to maintain multi-step complexity.

Instruction Generation and Detection/Segmentation. To facilitate fine-grained interaction, we implement a multi-stage instruction generation pipeline. We leverage GPT-5.2 (OpenAI, 2025) and Gemini 3 Pro (Team et al., 2023) to produce initial image captions and high-level semantic prompts, which are then refined through customized templates to ensure information integrity. For VQA task

Table 2. Comparative results of human alignments between different models. P. denotes parameters.

Model	P.	Human Alignment (ρ)				Average
		IC	UGG	GGU	ME	
Qwen3-VL	8B	0.96	0.78	0.61	0.46	0.70
GPT-5.2	-	0.98	**0.85**	**0.79**	0.55	0.79
Gemini 3 Pro	-	**0.99**	0.83	0.67	**0.75**	0.81
Unison-Judge	8B	0.98	0.82	0.77	0.72	**0.82**

construction, we utilize the Davidsonian Scene Graph (DSG) (Cho et al., 2023) framework, prompting Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025) to generate complex questions and candidate options. In tasks requiring precise localization, target objects are manually annotated with bounding boxes, which serve as prompts for SAM 3 (Carion et al., 2025) to produce segmentation masks. These masks undergo post-processing to ensure pixel-level accuracy for subsequent editing and understanding evaluations.

Automated Filtering and Human Annotation. To guarantee the highest fidelity and logical consistency, all generated data undergoes a rigorous dual-validation process. An initial automated filtering phase is applied to eliminate structural ambiguities and samples that fail to meet aesthetic or alignment thresholds. Subsequently, we conduct an extensive expert manual review. This phase includes: 1) inspecting image-text pairs to ensure all textual attributes are accurately reflected in the visual content; 2) correcting logical inconsistencies in the VQA reasoning chains; and 3) refining instruction prompts to eliminate any potential ambiguities. This refinement process ensures that the benchmark provides high-quality ground truth for evaluating both generative and discriminative capabilities of UMMs.

4.2. Evaluation Metrics

We evaluate UMMs using a suite of task-specific metrics. For localization and segmentation, we utilize Intersection over Union (IoU) to quantify predicted regions. For UniWorld, since it generates masks directly on images, we compute pixel-wise alignment by overlaying these masks onto the original images. Image generation is benchmarked against human-annotated VQA data to ensure compositional semantic fidelity. For image editing, we follow ImgEdit (Ye et al., 2025) to employ qualitative ratings on a scale of 0 to 5, which are transformed into a normalized scale of $(0, 1)$, specifically evaluating attribute consistency of edited regions and preservation of unedited areas.

4.3. Unison-Judge

To ensure a more reasonable and fine-grained evaluation, we shift from traditional feature-level similarity metrics (Salimans et al., 2016; Heusel et al., 2017; Hessel et al., 2021) to a customized MLLM-based assessment framework, which aligns more closely with human preference. We introduce Unison-Judge, an evaluation model derived by fine-tuning Qwen3-VL-8B-Instruct on a meticulously curated corpus. This corpus consists of 4,000 manually labeled samples in total, with 1,000 allocated to each dimension to ensure a uniform distribution. Furthermore, we conduct human alignment experiments on 500 samples uniformly sampled across all tasks, as shown in Table 3. Despite Gemini 3 Pro and GPT-5.2 show competitive performance on certain subsets, Unison-Judge achieves the superior average alignment score. This result underscores both the effectiveness and computational efficiency of our proposed framework.

5. Experiment

Experimental Setups. We evaluate UMMs across a comprehensive suite of both open-source and closed-source platforms. Open-source lineup encompasses Show-o/Show-o2 (Xie et al., 2024; 2025), Janus-pro (Chen et al., 2025b), D-DiT (Li et al., 2025b), ILLUME+ (Huang et al., 2025), OmniGen2 (Wu et al., 2025b), TokenFlow (Qu et al., 2025), Bagel (Deng et al., 2025), SEED-X (Ge et al., 2024), and UniWorld (Lin et al., 2025), respectively. Closed-source models are Gemini 3 Pro (Team et al., 2023) and GPT-5.2 (OpenAI, 2025). To ensure a fair comparison, all models adhere to a standardized input/output format, with image generation and editing resolution fixed at 512×512. In mutual enhancement experiments, we cap the process at 5 rounds. Notably, early termination is triggered if a model determines that no further errors remain, thereby optimizing computational efficiency.

5.1. Evaluation Results

As shown in Table 3, we evaluate various open-source and closed-source models. We averaged the scores from the four dimensions to obtain the final score.

In internal consistency evaluation, we find that understanding performance typically exceeds generation performance. Notably, a model's high comprehension or generation capability does not guarantee good consistency. For example, D-DiT (Li et al., 2025b), despite trailing in absolute scores for both understanding and generation, maintains higher consistency compared to ILLUME+ (Huang et al., 2025). This difference may be attributed to ILLUME+'s hybrid architecture, which decouples understanding and generation, while D-DiT's unified diffusion-based approach enables better cross-capability alignment.

For the Und.-Guided Gen. task, reasoning generally improves contextual understanding. However, ILLUME+ (Huang et al., 2025) shows a downward trend, with scores dropping from 0.250 to 0.125, due to its low Internal Consistency score, indicating a semantic misalignment that

Table 3. Comparative results of open-source and closed-source models on the proposed Unison benchmark. Und., Gen., and Uni. denote understanding, generation, and unified scores, respectively. For open-source models, the bold and underlined values indicate the best and second-best performance within their respective categories. / indicates the model does not support this task.

Model	Params	Internal Consistency			Und.-Guided Gen.			Gen-Guided Und.			Mutual Enhancement			Overall
		Und.	Gen.	Uni.	Und.	Gen.	Uni.	Und.	Gen.	Uni.	Und.	Gen.	Uni.	
Open-Source Unified Multimodal Models														
Show-o	1.3B	0.759	0.617	0.566	0.000	/	/	0.013	/	/	/	/	/	-
Janus-Pro	1.5B	0.883	0.694	0.679	0.002	/	/	0.092	/	/	/	/	/	-
Show-o2	1.5B	0.882	0.654	0.557	0.001	/	/	0.010	/	/	/	/	/	-
D-DiT	2B	0.734	0.595	0.565	0.000	/	/	0.005	/	/	/	/	/	-
ILLUME+	3B	0.737	0.178	0.155	0.002	0.056	0.029	0.101	0.193	0.147	0.090	0.022	0.056	0.097
Janus-Pro	7B	0.953	0.759	0.730	0.008	/	/	0.140	/	/	/	/	/	-
Show-o2	7B	**0.971**	0.771	0.758	0.000	/	/	0.021	/	/	/	/	/	-
ILLUME+	7B	0.796	0.221	0.177	0.005	0.083	0.037	0.140	0.250	0.125	0.005	0.113	0.059	0.113
OmniGen2	7B	0.919	<u>0.803</u>	<u>0.751</u>	0.040	<u>0.427</u>	<u>0.219</u>	0.141	**0.412**	**0.206**	<u>0.393</u>	<u>0.372</u>	**0.382**	<u>0.389</u>
TokenFlow	14B	0.928	0.541	0.505	0.000	/	/	0.190	/	/	/	/	/	-
BAGEL	14B	<u>0.957</u>	**0.849**	**0.820**	0.000	**0.701**	**0.337**	**0.299**	<u>0.374</u>	<u>0.187</u>	0.040	**0.459**	0.249	**0.398**
SEED-X	17B	0.809	0.424	0.367	<u>0.042</u>	0.153	0.081	0.150	0.248	0.124	0.011	0.159	0.085	0.164
UniWorld	19B	0.921	0.702	0.664	**0.068**	0.343	0.185	<u>0.280</u>	0.315	0.160	**0.469**	0.082	<u>0.275</u>	0.318
Closed-Source Models														
Gemini 3 Pro	-	0.995	0.891	0.879	0.271	0.362	0.317	0.538	0.312	0.425	0.557	0.496	0.527	0.537
GPT-5.2	-	0.997	0.883	0.862	0.231	0.371	0.301	0.421	0.359	0.390	0.485	0.572	0.529	0.521

limits reasoning effectiveness.

In the Gen-Guided Und. task, the generated content scores remain low, revealing numerous inaccuracies, as shown in Figure 3. However, these errors paradoxically enhance the model's understanding performance, by providing contrastive boundaries that sharpen semantic comprehension.

Finally, in the Mutual Enhancement task, models like Omnigen2 and UniWorld show significant performance gains, with scores of 0.382 and 0.275, respectively. These results suggest that large-scale, integrated models benefit from multi-modal synergies, where proficiency in one modality enhances performance in others.

6. Insights and Discussions

Importance of Internal Consistency. The results in Table Table 3 highlight that internal consistency is a fundamental prerequisite for high-performing UMMs. We observe a strong correlation between a model's overall score and its performance in the internal consistency dimension. For instance, closed-source models like Gemini 3 Pro (Team et al., 2023) and GPT-5.2 (OpenAI, 2025) exhibit near perfect consistency score, which serves as the foundation for their superior reasoning capabilities. Among open-source models, UniWorld (Lin et al., 2025) and OmniGen2 (Wu et al., 2025b) achieve the highest overall scores (0.398 and 0.376 respectively) by maintaining robust consistency across both understanding and generation tasks. This emphasizes the importance of a UMM's ability to maintain self-alignment, which plays a critical role in minimizing errors.

UMM's understanding and generation capabilities can synergistic work. Contrary to the belief that understanding and generation are disparate tasks, our benchmark demonstrates a clear synergistic effect. In the Understanding-Guided Generation and Generation-Guided Understanding, models do not merely perform well in isolation, they leverage one capability to bootstrap the other. For example, BAGEL achieves a state-of-the-art open-source score of 0.337 in the UGG task. This indicates that the latent representations learned during pre-training provide rich structural priors that enhance understanding, and vice versa.

UMMs can self-reinforce autonomously. A pivotal finding from the mutual enhancement category is that UMMs possess the latent ability for autonomous self-improvement. Even when the reference information provided to the model is sub-optimal or contains noise, as implied by the gap between guided and unified scores, models like Omnigen2 and UniWorld show significant gains in the Uni. score (0.382 and 0.275) of Mutual Enhancement. This suggests that these models can perform internal cross-verification, where the generation process acts as a check on understanding, and the understanding module filters generative outputs. This mechanism allows the model to extract value even from incorrect or noisy internal states, effectively achieving a degree of self-reinforcement without external supervision.

7. Limitations

Computational Cost of Evaluation. Assessing mutual enhancement involves iterative cycles of generation followed by comprehension, and vice versa. This recursive process significantly increases computational overhead, presenting scalability challenges when extending the benchmark to high-resolution imagery or multi-frame video sequences.

Metric Sensitivity. Although we introduce Unison-Judge, demonstrating competitive human alignment compared to Gemini 3 Pro, the models under assessment may still exhibit sensitivity to prompt engineering or generative stochasticity. This can introduce a degree of variance that might not fully reflect the model's inherent reasoning limits.

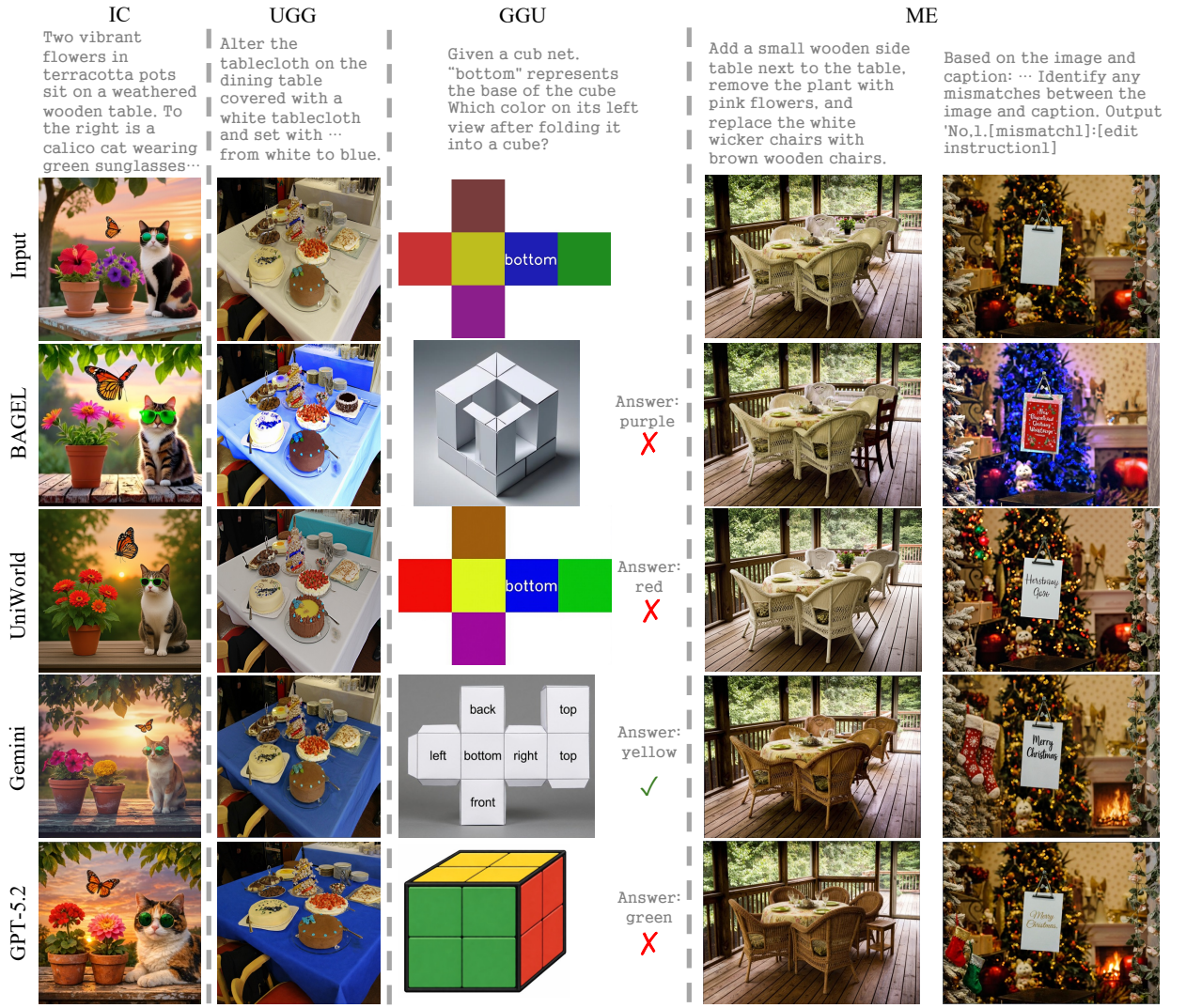


Figure 3. Qualitative Visualizations. IC, UGG, GGU and ME denotes Internal Consistency, Understanding-Guided Generation, Generation-Guided Understanding and Mutual Enhancement, respectively. For GGU, visualizations on inclue the 3D spatial task.

8. Future Directions

Inernal Consistency Improvement. Future research should move beyond standard pre-training objectives by incorporating explicit consistency objective. By penalizing discrepancies between a model's discriminative understanding and its generative output during training, we can foster more robust self-alignment and minimize the hallucinations and error accumulations inherent in current UMMs.

Self-Refinement Training via Agentic Planning. A promising direction lies in integrating UMM self-refinement with autonomous agentic planning strategies. By enabling models to autonomously design task-specific plans, they can treat understanding and generation as iterative feedback steps in a self-play loop. This allows the model to generate its own pseudo-supervision signals, facilitating continuous self-evolution and reinforcement learning

Additional Modality Support. While current open-source UMMs largely lack video synthesis capabilities, integrating the video modality is essential. Future work should bridge the gap between video understanding and generation, enabling models to simulate physical world dynamics and improve long-range causal reasoning.

9. Conclusion

We introduce Unison, a benchmark designed to evaluate the synergistic understanding and generation capabilities of unified multimodal models. Unlike prior benchmarks that assess these skills in isolation, Unison integrates four key dimensions: internal consistency, understanding-guided generation, generation-guided understanding, and mutual

enhancement, and offers both unified and decoupled evaluation tracks for fine-grained analysis. To ensure reliable scoring, we train Unison-Judge, a human-aligned evaluator. Our evaluations reveal critical gaps in current models and demonstrate performance comparable to Gemini 3 Pro.

Impact Statement

Unison provides a comprehensive framework for unified multimodal research by systematically evaluating the essential synergy between understanding and generation. This work contributes to the reliability of AI systems by identifying internal semantic conflicts and hallucinations, while enabling efficient, human-aligned automated assessment through Unison-Judge. By revealing critical performance gaps in current models, Unison offers practical observations that can help researchers design and develop more consistent and robust unified multimodal architectures in the future.

Acknowledgment

This work was supported in part by the Science and Technology Commission of Shanghai Municipality under Grant No. 25511103600.

References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *ICCV*, 2015.

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., and Zhu, K. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.

Carion, N., Gustafson, L., Hu, Y.-T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K. V., Khedr, H., Huang, A., et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Chang, J., Fang, Y., Xing, P., Wu, S., Cheng, W., Wang, R., Zeng, X., Yu, G., and Chen, H.-B. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. In *NIPS*, 2025.

Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., and Ruan, C. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

Cho, J., Hu, Y., Garg, R., Anderson, P., Krishna, R., Baldridge, J., Bansal, M., Pont-Tuset, J., and Wang, S. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. *arXiv preprint arXiv:2310.18235*, 2023.

Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., and Fan, H. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Ge, Y., Zhao, S., Zhu, J., Ge, Y., Yi, K., Song, L., Li, C., Ding, X., and Shan, Y. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

Geng, Z., Wang, Y., Ma, Y., Li, C., Rao, Y., Gu, S., Zhong, Z., Lu, Q., Hu, H., Zhang, X., et al. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025.

Ghosh, D., Hajishirzi, H., and Schmidt, L. Geneval: An object-focused framework for evaluating text-to-image alignment. In *NIPS*, 2023.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., and Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017.

Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., and Yu, G. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Huang, K., Sun, K., Xie, E., Li, Z., and Liu, X. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *NIPS*, 2023.

Huang, R., Wang, C., Yang, J., Lu, G., Yuan, Y., Han, J., Hou, L., Zhang, W., Hong, L., Zhao, H., et al. Illume+: Illuminating unified mllm with dual visual tokenization and diffusion refinement. *arXiv preprint arXiv:2504.01934*, 2025.

Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., and Shan, Y. Seed-bench: Benchmarking multimodal llms with generative comprehension. In *CVPR*, 2024.

Li, Y., Qian, R., Pan, B., Zhang, H., Huang, H., Zhang, B., Tong, J., You, H., Du, X., Gan, Z., Kim, H., Jia, C., Wang, Z., Yang, Y., Gao, M., Dou, Z.-Y., Hu, W., Gao, C., Li, D., Dufter, P., Wang, Z., Yin, G., Zhang, Z., Chen, C., Zhao, Y., Pang, R., and Chen, Z. Manzano: A simple and scalable unified multimodal model with a hybrid vision tokenizer. *arXiv preprint arXiv:2509.16197*, 2025a.

Li, Z., Li, H., Shi, Y., Farimani, A. B., Kluger, Y., Yang, L., and Wang, P. Dual diffusion for unified image generation and understanding. In *CVPR*, 2025b.

Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S., Yu, W., Wang, S., Ge, Y., et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2024.

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., and Murphy, K. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016.

OpenAI. GPT-5.2, 2025.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015.

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. In *CVPR*, 2025.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In *NIPS*, 2016.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NIPS*, 2022.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. Towards vqa models that can read. In *CVPR*, 2019.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Tian, R., Gao, M., Xu, M., Hu, J., Lu, J., Wu, Z., Yang, Y., and Dehghan, A. Unigen: Enhanced training & test-time strategies for unified multimodal understanding and generation. In *NIPS*, 2025.

Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.-m., Bai, S., Xu, X., Chen, Y., et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

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. *IJCV*, 2025c.

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. In *ICLR*, 2025d.

Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Li, C., Wang, S., Huang, T., and Liu, Z. Omnigen: Unified image generation. In *CVPR*, 2025.

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. In *ICLR*, 2024.

Xie, J., Yang, Z., and Shou, M. Z. Show-o2: Improved native unified multimodal models. In *NIPS*, 2025.

Ye, Y., He, X., Li, Z., Lin, B., Yuan, S., Yan, Z., Hou, B., and Yuan, L. Imgedit: A unified image editing dataset and benchmark. In *NIPS*, 2025.

Yu, Q., Chow, W., Yue, Z., Pan, K., Wu, Y., Wan, X., Li, J., Tang, S., Zhang, H., and Zhuang, Y. Anyedit: Mastering unified high-quality image editing for any idea. In *CVPR*, 2025.

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 *CVPR*, 2024.

Zhang, K., Mo, L., Chen, W., Sun, H., and Su, Y. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *NIPS*, 2023.

Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., and Levy, O. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *ICLR*, 2025.