IMUG-BENCH: BENCHMARKING UNIFIED MULTIMODAL MODELS ON INTERLEAVED UNDERSTANDING AND GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, unified multimodal models (UMMs) have emerged to support both understanding and generation within a single framework. Mastering dynamic, multi-turn interleaved image–text dialogues is a crucial task for UMMs in real-world applications. However, existing benchmarks fail to evaluate this important task, as they are often limited to single-turn or static settings, and typically overlook exposure bias in multi-turn interactions. To bridge this gap, we propose IMUG-Bench, a comprehensive benchmark for multi-turn interleaved image–text dialogue of UMMs that jointly evaluates their understanding and generation capabilities. Our IMUG-Bench comprises three classes: Static Spatial, Temporal Causal, and Hybrid—covering 3,113 samples and 12,034 interaction turns. It also includes dynamic understanding questions, thereby supporting evaluation that better reflects real-world multi-turn interaction scenarios. Large-scale experiments on IMUG-Bench systematically evaluate mainstream open-source and closed-source UMMs, revealing their capability boundaries and failure modes, and uncovering pronounced exposure bias on the generation side in multi-turn interactions. We further explore several test-time scaling strategies, including Chain-of-Thought, Self-Verification, and Best-of-N Sampling, which effectively improve generation accuracy and mitigate exposure bias in generation tasks. These findings provide insights for improving UMM robustness and multi-turn interaction.

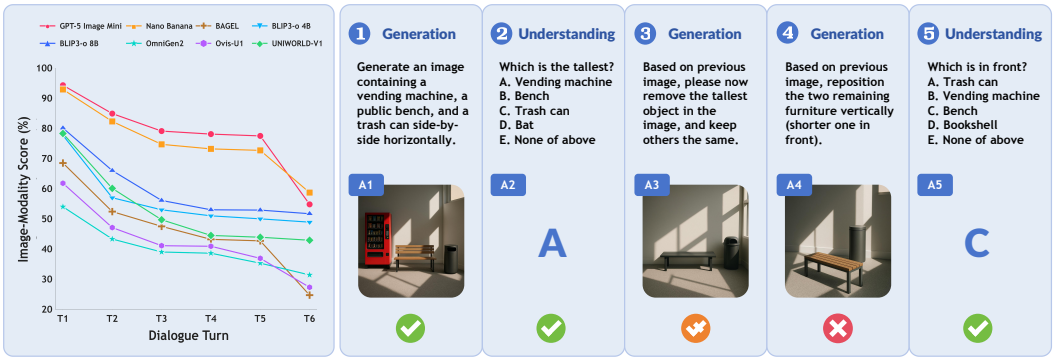

Figure 1: Exposure bias in multi-turn interleaved generation. Left: Image modality scores decline across dialogue turns. Right: An example where image generation quality progressively deteriorates while image understanding remains correct throughout the illustrated interaction.

1 INTRODUCTION

Recent multimodal models have made substantial progress in both understanding and generation (Achiam et al., 2023; Wu et al., 2023b; Wang et al., 2024a; Chen et al., 2024b; Yin et al., 2024). Multimodal large language models (MLLMs) have developed strong visual perception, reasoning, and zero-shot generalization capabilities (Dai et al., 2023; Liu et al., 2023a; Caffagni et al., 2024;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

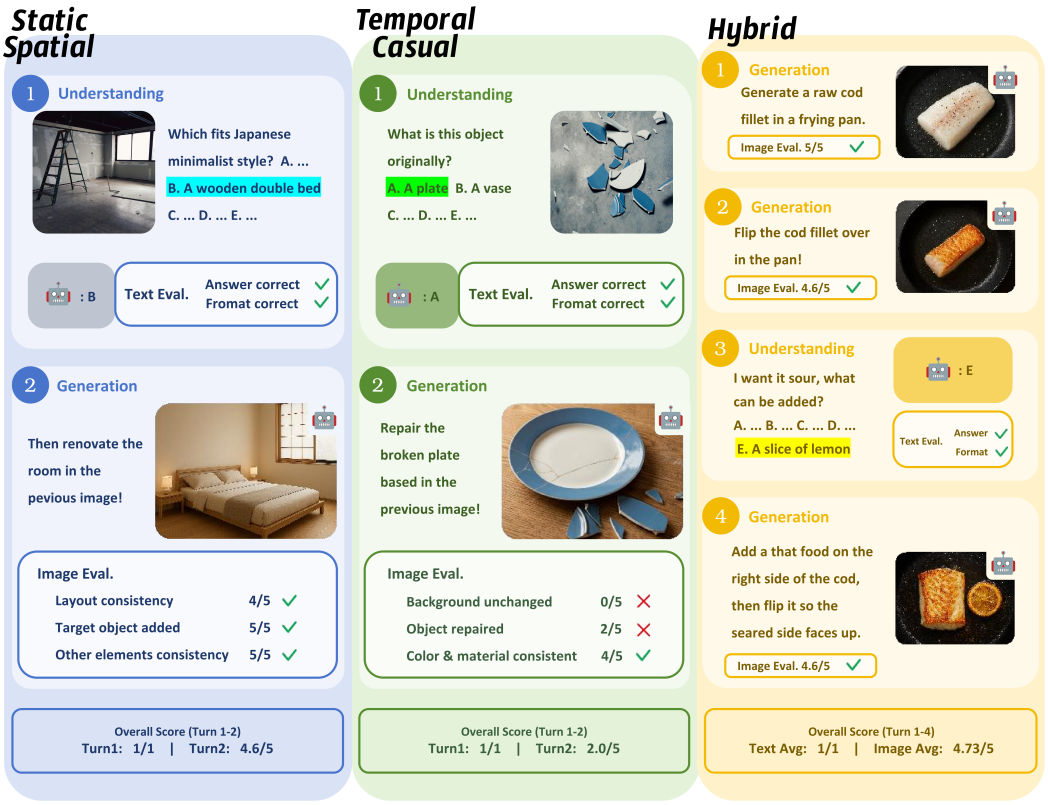

Figure 2: Overview of IMUG-Bench. IMUG-Bench evaluates UMMs' understanding and generation capabilities through multi-turn interleaved interactions.

Zhang et al., 2024), while generative models can produce increasingly high-quality image and video content (Podell et al., 2023; He et al., 2023; Blattmann et al., 2023; Gao et al., 2025). However, understanding and diffusion-based generation models have largely evolved along separate architectural paths, making unified integration challenging (Zhao et al., 2025).

Unified multimodal models (UMMs) address this separation by integrating understanding and generation within a single framework, supporting interleaved multimodal interaction, multi-turn dialogue, and unified understanding–generation workflows (Tian et al., 2024; Zhan et al., 2024; Zhou et al., 2025; Deng et al., 2025; Zhang et al., 2025). Nevertheless, current UMMs still face imbalances between understanding and generation and limitations in integrated architectural design (Zhao et al., 2025), motivating evaluation under realistic multi-turn interactions.

Existing benchmarks remain limited in this setting. First, most use single-turn protocols that evaluate understanding or generation in isolation and therefore fail to capture multi-turn interleaved image–text dialogue (Li et al., 2025a; Xie et al., 2025b). Second, even benchmarks with multi-turn interleaving may rely on static question answering without updating ground-truth answers according to model responses, limiting objective evaluation of long-horizon context awareness (Chow et al., 2025). Third, existing benchmarks mainly characterize capabilities and failure modes, while mitigation strategies remain underexplored (Xia et al., 2024; Chen et al., 2024a; Zhou et al., 2025).

To address these limitations, we propose IMUG-Bench, a benchmark for **I**nterleaved **M**ultimodal **U**nderstanding and **G**eneration (Fig. 2). We manually construct question, answer, and evaluation-point templates, generate and filter candidate keywords, collect images from multiple sources, and use LLMs/VLMs for template filling. IMUG-Bench contains three classes—Static Spatial, Temporal Causal, and Hybrid—covering 19 domains, 97 tasks, 3,113 samples, and 12,034 dialogue turns. It enables unified evaluation of UMM understanding and generation in multi-turn interleaved settings and differs from existing benchmarks as summarized in Table 1.

Table 1: Comparison of unified multimodal benchmarks. Abbreviations: Modal. = Modality; Mult. Iter. = Multi-turn Interaction; Dyna. Und. = Dynamic Understanding; Expos. = Exposure.

Benchmark	Samples	Iter. Turns	Modal. Unified	Mult. Iter.	Dyna. Und.	Expos. Bias	Test-time Scaling
CoMM[1]	500	500	✓	✗	✗	✗	✗
UniEval[2]	4,231	4,321	✗	✗	✓	✗	✗
MMIE[3]	20,103	20,103	✓	✗	✗	✗	✗
OpenING[4]	5,400	5,400	✓	✗	✗	✗	✗
ISG-Bench[5]	1,150	1,150	✓	✗	✗	✗	✗
MME-U[6]	4,104	4,104	✗	✗	✗	✗	✗
Uni-MMMU[7]	885	885	✓	✗	✓	✗	✗
WEAVE Bench[8]	100	367	✓	✓	✗	✗	✗
IMUG-Bench	3,113	12,034	✓	✓	✓	✓	✓

[1] Chen et al. (2025c); [2] Li et al. (2025a); [3] Xia et al. (2024); [4] Zhou et al. (2025); [5] Chen et al. (2024a); [6] Xie et al. (2025b); [7] Zou et al. (2025); [8] Chow et al. (2025).

Using IMUG-Bench, we systematically evaluate mainstream open-source and closed-source UMMs. The results reveal substantial limitations in long-context multi-turn generation and pronounced generation-side exposure bias as interaction turns increase. We further explore Chain-of-Thought, Self-Verification, and Best-of-N Sampling as test-time scaling strategies, showing that they can improve generation performance and mitigate the observed degradation.

In summary, our main contributions are:

- **IMUG-Bench benchmark construction**: We introduce a broadly covered multi-turn interleaved benchmark for fine-grained evaluation of UMM understanding and generation.
- **Systematic evaluation and analysis**: We evaluate mainstream open-source and closed-source UMMs, characterize their capability boundaries, and identify generation-side exposure bias in long-horizon multi-turn interleaved interactions.
- **Exploration of test-time scaling**: We study Chain-of-Thought, Self-Verification, and Best-of-N Sampling and examine their effects on generation performance and exposure bias.

2 Related Work

2.1 Unified Multimodal Models

The rapid evolution of autoregressive LLMs has driven the development of MLLMs, which align cross-modal features into the text token space for fine-grained visual grounding and robust cross-modal instruction following (Wu et al., 2023b; Wang et al., 2024b; Liang et al., 2024; Yin et al., 2024). Meanwhile, image generation models based on diffusion or flow matching have also made significant progress, enabling controllable and high-fidelity image synthesis, and they are gradually expanding to broader tasks such as 3D and video generation (Lipman et al., 2022; Zhang et al., 2023; Chen et al., 2025a). With these advances, UMMs aim to integrate understanding and generation, which enables a unified interleaved multimodal interaction paradigm with mixed text and images for both input and output (Zhao et al., 2025; Yang et al., 2025). This architecture facilitates cross-modal semantic sharing, thereby forming a closed loop between understanding and generation over multiple turns (Chen et al., 2025b; Wu et al., 2025; Deng et al., 2025; Zhang et al., 2026). Early efforts achieved system-level integration by employing an LLM as a central controller to orchestrate isolated expert models (Shen et al., 2023; Yang et al., 2023; Wu et al., 2023a; Surís et al., 2023). More recent work explores architectures that jointly model multimodal understanding and generation (Xie et al., 2024; Li et al., 2025b; Shi et al., 2025). However, unified models still struggle to balance understanding and generation and transfer LLM knowledge and reasoning to generation (Zhao et al., 2025; Xiao et al., 2025; Su et al., 2025).

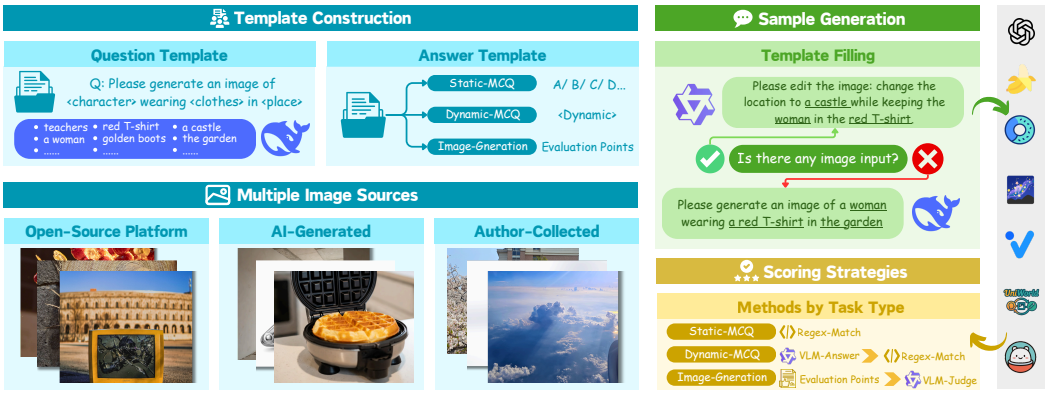

Figure 3: The IMUG-Bench construction and evaluation pipeline: manually designed templates are populated via LLM/VLM filling using images from multiple sources, and samples are scored with task-specific strategies designed for different question types.

2.2 Benchmarks for Unified Multimodal Models

For understanding tasks, existing benchmarks have evolved from basic visual perception toward fine-grained understanding and multimodal reasoning, and have further expanded to subject-specific problems and structured or heterogeneous visual inputs such as tables and charts (Yu et al., 2023; Liu et al., 2024; Yue et al., 2024). For generation tasks, existing benchmarks have progressed from measuring text–image alignment to evaluating controllability and fidelity, and increasingly incorporate overall quality and human preference (Ghosh et al., 2023; Liu et al., 2023b; Wu et al., 2023c). With the rise of unified models, unified multimodal benchmarks are also being released and improved step by step. These benchmarks evaluate aspects such as separate evaluation of understanding and generation (Xie et al., 2025b; Li et al., 2025a), cross-modal consistency and alignment of outputs (Chen et al., 2024a; Zhou et al., 2025), and the mutual enhancement between understanding and generation (Zou et al., 2025). However, dynamic interaction in long-horizon scenarios is a crucial capability for UMMs in real-world applications. Current unified multimodal benchmarks are often limited to relatively static settings and fail to evaluate this important task (Chow et al., 2025). To fill this critical gap, we propose IMUG-Bench, which uses multi-turn and mixed text–image dialogue to conduct a detailed joint evaluation of UMMs' understanding and generation capabilities.

3 Methodology

3.1 Preliminary

IMUG-Bench is designed to evaluate UMMs under multi-turn, interleaved, and dynamic interactions. We organize its content into three classes: Static Spatial, Temporal Causal, and Hybrid (Huang et al., 2025; Du et al., 2025; Xie et al., 2025a; Cheng et al., 2025). Static Spatial evaluates the recognition and generation of target-element attributes, ownership, and spatial relationships; Temporal Causal focuses on reasoning over implicit natural laws and common-sense causal relations; and Hybrid combines both to cover more complex daily scenarios. We further organize representative multi-turn scenarios into domains and fine-grained tasks, forming a hierarchical benchmark structure with broad coverage of common real-world interactions.

3.2 Benchmark Construction

IMUG-Bench is constructed through a pipeline of manual template design, LLM/VLM-based filling, and two-person human verification of LLM/VLM outputs. The overall construction pipeline is illustrated in Figure 3, and benchmark statistics are provided in Section 3.4.

Template Design. Each task instance is a multi-turn interaction composed of three question types: Static-MCQ and Dynamic-MCQ for understanding, and Image-Generation for generation:

- **Static-MCQ** uses fixed-answer MCQs to evaluate understanding in a stable multimodal context with predefined answers that remain unchanged.
- **Dynamic-MCQ** is a multiple-choice question whose correct answer depends on a previously generated image, evaluating the model's understanding of its generated content and addressing static ground-truth limitations in realistic interactions.
- **Image-Generation** requires image generation or editing from instructions, optionally with reference images, and evaluates instruction following, visual consistency, and multi-turn editing.

For each task, we manually design question templates, determine the number of turns and input–output modalities, and annotate each non-initial turn with the indices of previous turns it depends on. We leave content-defining keywords as placeholders and use an LLM to enumerate candidate keyword key–value pairs. Detailed domain/task distributions, turn counts, and interleaved modality sequences are provided in Appendix B.

We also design answer templates for the three question types. For Static-MCQ, the template specifies the option-construction logic, with options generated by an LLM/VLM during filling. Dynamic-MCQ uses a `<DYNAMIC>` placeholder for its answer. For Image-Generation, we define a fixed set of evaluation points with predefined numbers and evaluation aspects, leaving task-specific keywords as placeholders. Images, question templates, keyword lists, and answer templates are then checked by two annotators, with problematic items removed or corrected.

Sample Generation. After template verification, we instantiate the templates using LLMs/VLMs, diversifying the keyword combinations to cover a broad range of concrete questions. An LLM is used for text-only questions, while a VLM is used for questions with image inputs.

For Static-MCQ, an LLM/VLM directly fills the templates to generate the final samples. For Dynamic-MCQ, a VLM determines the ground-truth answer only after model inference, based on the prompt image, prompt text, and options. For Image-Generation, task-specific keywords are filled into the evaluation-point templates. The LLM/VLM also smooths question wording to avoid mechanical keyword insertion while preserving meaning and preventing content leakage.

Before large-scale filling, we conduct a small-scale manual verification of the constructed samples. After large-scale construction, all samples undergo a two-person manual review to remove factual errors. Details of image sources and curation are provided in Appendix A, and question-filling prompts are provided in Appendix E.1.

3.3 IMUG-Bench Evaluation

We use separate scoring strategies for text and image outputs.

Understanding. For Static-MCQ and Dynamic-MCQ, the model determines the correct options and their number and outputs them in the required format. We apply regex-based rule scoring:

$$S_{\mathrm{MCQ}} = w_{\mathrm{fmt}} \left(n_{\mathrm{corr}} - n_{\mathrm{incorr}}\right) u.$$

Here, n_{corr} and n_{incorr} are the numbers of selected correct and incorrect options, u is the unit score per correct option, defined as 1 divided by the number of correct options, and w_{fmt} is the formatting weight. We set $w_{\mathrm{fmt}}{=}1$ for a continuous sequence of uppercase option letters with no separators; 0.75 when uppercase option letters are present but the required format is violated (e.g., separators or accompanying option text) without additional explanation; 0.5 when extra explanatory text is included; and 0 when no valid option letter is produced.

Generation. Because image outputs are open-ended and difficult to score with rule-based methods, we follow the VLM-as-a-Judge paradigm and use dynamic evaluation-point filling, whose validity is further examined through correlation analysis and two-person human ratings (Appendix D).

Each image is evaluated over multiple personalized evaluation points, which fall into two categories: (1) *Image update requirements*, measuring whether the expected content or requested changes are correctly realized, including targeted generation and style adjustment; and (2) *Image consistency checks*, measuring whether attributes that should remain unchanged are preserved from previous reference images, when such references exist. Each image question is assigned multiple evaluation points according to its target skills.

During scoring, each output image is evaluated independently, with previous images provided to the VLM judge when needed for consistency assessment. The judge assigns an integer score $s_i \in [0, 5]$ to each evaluation point, where higher scores indicate better fulfillment, and the normalized turn-level image score is computed using the following expression:

$$S_{\text{img}} = \frac{\sum_{i=1}^{N} s_i}{N \cdot 5},$$

where N is the number of evaluation points.

3.4 Benchmark Statistics

IMUG-Bench contains 3,113 samples spanning 19 domains and 97 tasks across the three classes. Each sample consists of 2–6 multi-turn interactions with text-only or text-and-image inputs and interleaved text–image outputs, yielding 12,034 turns in total and 3.87 turns per sample on average. Understanding and generation turns are approximately balanced.

Detailed class/domain distributions and high-frequency concepts are provided in Appendix B, Fig. 5; the full domain–task breakdown, turn counts, and modality sequences are provided in Table 5.

3.5 Test-time Scaling

Our experiments reveal exposure bias in image questions, where UMM performance degrades as the number of interaction turns increases. We therefore explore three test-time scaling strategies to improve multi-turn generation and investigate their effects on this degradation.

Chain-of-Thought. Given the effectiveness of Chain-of-Thought in language-model reasoning (Wei et al., 2022), we use it to enhance instruction alignment over multi-turn context. Before answering an image question, the model summarizes the request, lists a generation plan, and generates the image accordingly. The reasoning is not appended to the dialogue history.

Self-Verification. Since UMMs typically exhibit stronger understanding than generation, we introduce a closed-loop Self-Verification mechanism that uses the model's understanding capability to check instruction following and cross-turn consistency. After generating an image R_t, the model assesses whether it satisfies the current instruction and preserves the required style and visual details from previous context. If verification fails, the model provides diagnostic feedback to guide regeneration, up to a predefined maximum number of verification rounds; otherwise, the image is accepted. Only the final accepted or retained image is appended to the dialogue history, while intermediate failed images and review feedback are discarded.

Best-of-N Sampling. Since image generation is stochastic, we generate N candidate images $R_{t,1}, R_{t,2}, \ldots, R_{t,N}$ under the same dialogue history and current request. The model then evaluates the candidates and selects, by index, the one that best satisfies the current request and necessary cross-turn consistency. Only the selected candidate is used as the turn output and appended to the dialogue history; unselected candidates and their evaluations are discarded.

4 Experiment

4.1 Experimental Setup

We evaluate eight UMMs, including six open-source UMMs (BAGEL Deng et al. (2025), BLIP3-o 4B Chen et al. (2025b), BLIP3-o 8B Chen et al. (2025b), OmniGen2 Wu et al. (2025), Ovis-U1 Wang et al. (2025), and UNIWORLD-V1 Lin et al. (2025)) and two closed-source UMMs (Nano Banana (Gemini 2.5 Flash Image) Comanici et al. (2025) and GPT-5 Image Mini (GPT-5 Mini + GPT Image 1 Mini) Singh et al. (2025)).

History Construction. For open-source models, we concatenate the user queries and model answers in order, add the necessary role tags, and feed the full history back to the model. BAGEL supports interleaved text–image inputs in the original order, so we rebuild the history in the exact interaction order and return it to the model. In contrast, BLIP3-o, OmniGen2, Ovis-U1, and UNIWORLD-V1 only support a serialized image-list input, so we represent the history as an image

Table 2: Main results on IMUG-Bench. We report understanding (Und.) and generation (Gen.) scores across the three classes (Static Spatial, Temporal Causal, Hybrid), along with the overall average (Avg); the top-3 results in each metric are highlighted as **gold** (best), **silver** (second best), and **bronze** (third best) for easier visual comparison across models.

Model Name	Open Source	Size	St.&Sp.		Te.&Ca.		Hybrid		Avg
			Und.	Gen.	Und.	Gen.	Und.	Gen.	
GPT-5 Image Mini[1]	✗	-	**81.6**	**88.0**	**74.8**	**83.1**	**80.9**	**79.3**	**82.3**
Nano Banana[2]	✗	-	**78.0**	**86.9**	**73.7**	**82.8**	**70.6**	**76.7**	**79.5**
BAGEL[3]	✓	15B	49.7	57.9	53.8	49.3	36.4	52.8	51.9
BLIP3-o 4B[4]	✓	10B	53.5	62.4	62.1	59.3	**55.8**	60.6	59.3
BLIP3-o 8B[4]	✓	14B	65.8	**67.3**	68.9	**61.1**	52.4	**67.8**	**65.0**
OmniGen2[5]	✓	7B	**67.7**	44.4	**70.0**	43.4	55.2	50.2	55.1
Ovis-U1[6]	✓	3B	54.0	45.7	51.6	49.3	48.5	54.2	50.0
UNIWORLD-V1[7]	✓	20B	29.3	63.9	28.4	55.5	24.2	57.8	46.6

[1] Singh et al. (2025); [2] Comanici et al. (2025); [3] Deng et al. (2025); [4] Chen et al. (2025b); [5] Wu et al. (2025); [6] Wang et al. (2025); [7] Lin et al. (2025).

list with in-text placeholders. In particular, since Ovis-U1 does not support multi-image editing, we horizontally concatenate all historical input and output images, assign indices, and add a brief prompt that specifies the image-input rules during the dialogue. For closed-source models, we use their official multi-turn dialogue APIs to run multi-turn conversations with history memory.

Scoring Inputs and Model Selection. We follow Section 3.3 for scoring. For Dynamic-MCQ, the judge takes as input the question and options of the target turn, together with the full context of the annotated history-related turns, and returns the set of ground-truth answer options. For Image-Generation turns, the judge takes as input the question of the target turn, the evaluation-point list, the model-generated image, and the annotated history-related images, and returns structured per-point scores. For benchmark construction, we use DeepSeek-V3.2 and Qwen3-VL-235B-A22B-Thinking. For scoring, we use Qwen3-VL-235B-A22B-Thinking as the judge. Prompts for model evaluation, dynamic MCQ answer retrieval, and image assessment are provided in Appendix E.2–E.4.

4.2 Main Results on IMUG-Bench

Table 2 reports the understanding (Und.) and generation (Gen.) scores of each evaluated model on the three IMUG-Bench classes (Static Spatial, Temporal Causal, and Hybrid), together with the overall average score (Avg). In the overall ranking, closed-source models lead by a clear margin: GPT-5 Image Mini achieves the highest understanding and generation scores across all three classes, followed closely by Nano Banana, and both maintain consistently strong performance across classes. Among open-source models, BLIP3-o 8B performs best overall and ranks first among open-source models in generation, indicating relatively stable generation performance across task categories.

These results suggest the following. First, the advantage of closed-source models is not only higher average scores but also stronger stability: their understanding and generation scores remain high across all three classes, indicating more balanced capabilities and lower sensitivity to task-distribution shifts. In contrast, open-source models score lower overall and fluctuate more, reflecting a remaining gap in robustness. Second, many open-source models exhibit a clear imbalance between understanding and generation. Under IMUG-Bench's multi-turn interleaved setting, this imbalance directly hurts overall performance, because generation often depends on prior understanding and constraint alignment, and weaknesses in one side can propagate across turns and reduce the average score. Third, all open-source models obtain low scores on four domains—Stack Simulation, Geometric Coloring, Mirror Reasoning, and Paper Crafting—indicating a shared weakness in fine-grained spatial or geometric multimodal tasks under multi-turn interleaving. Additional radar charts and cross-domain analyses are provided in Appendix C.1, Figs. 6–8.

We further compare the performance differences between Static-MCQ and Dynamic-MCQ across all evaluated models (Appendix C.4, Table 7). We find that all models score substantially lower on Dynamic-MCQ than on Static-MCQ. This result indicates that, in complex interleaved multi-turn

Table 3: Per-turn results of all evaluated models on IMUG-Bench. For each turn, scores are reported as **Text/Image**. Open indicates whether the model is open-source, and Avg denotes the overall average score. The top-3 overall average results are highlighted as **gold** (best), **silver** (second best), and **bronze** (third best) for easier visual comparison across models.

Model Name	Open	Size	T1	T2	T3	T4	T5	T6	Avg
GPT-5 Image Mini[1]	✗	-	73.8/94.3	84.8/84.9	86.1/79.1	76.0/78.1	71.2/77.5	77.2/54.8	**82.3**
Nano Banana[2]	✗	-	69.7/92.9	81.7/82.3	81.7/74.7	73.0/73.2	65.2/72.7	72.8/58.7	**79.5**
BAGEL[3]	✓	15B	47.8/68.5	49.9/52.4	48.8/47.5	50.7/43.2	46.4/42.7	44.8/24.7	51.9
BLIP3-o 4B[4]	✓	10B	52.3/77.7	60.6/57.0	62.7/53.0	58.2/51.0	42.8/50.0	66.9/48.9	59.3
BLIP3-o 8B[4]	✓	14B	56.4/80.1	69.9/66.0	66.0/56.1	63.1/53.0	62.4/52.9	74.4/51.7	**65.0**
OmniGen2[5]	✓	7B	71.7/54.0	67.7/43.3	64.4/39.0	63.8/38.6	61.0/35.3	65.1/31.4	55.1
Ovis-U1[6]	✓	3B	49.4/61.8	51.6/47.1	50.5/41.1	54.3/40.9	53.2/36.9	58.5/27.3	50.0
UNIWORLD-V1[7]	✓	20B	21.9/78.3	28.8/60.1	29.1/49.7	32.1/44.5	28.6/43.9	27.6/42.9	46.6

[1] Singh et al. (2025); [2] Comanici et al. (2025); [3] Deng et al. (2025); [4] Chen et al. (2025b); [5] Wu et al. (2025); [6] Wang et al. (2025); [7] Lin et al. (2025).

interactions, models' understanding performance is more easily affected by accumulated errors in the dialogue history, making it difficult for them to accurately understand the specific content of their own previously generated images. This phenomenon further suggests that multi-turn understanding of self-generated content remains an important challenge for current UMMs.

4.3 Exposure Bias of UMMs

Overall, text-question scores do not show a consistent monotonic trend as the number of turns increases; in contrast, image-question scores show a clear downward trend with more turns for almost all models. Table 3 presents the turn-wise score distributions of all evaluated models on text questions and image questions. A more intuitive visualization of these trends is provided in Appendix C.2, Fig. 9, to facilitate direct comparison across interaction turns.

This phenomenon reveals the exposure bias of current UMMs in multi-turn interaction: unlike single-turn evaluation, multi-turn image generation must satisfy both the current instruction and cross-turn consistency constraints. Generation errors in early turns are written into later context and repeatedly exposed to the model, forcing later turns to generate conditioned on an imperfect history and leading to error accumulation and distribution drift. As a result, as the number of interaction turns grows, generation scores are more prone to decay, highlighting the necessity to evaluate and mitigate exposure bias in long-horizon interleaved tasks.

4.4 Case Study

As shown in Fig. 4, the example illustrates three typical failure modes in multi-turn interleaved interaction. First, GPT-5 Image Mini follows the required answer format and the initial image-editing instruction more reliably, whereas BAGEL exhibits formatting and instruction-following errors. Second, both models struggle with cross-turn image generation: GPT-5 Image Mini fails to preserve the previous edit, while BAGEL propagates an earlier generation error to the subsequent turn. Finally, both models hallucinate the shoe color in the Dynamic-MCQ turn, despite the corresponding attribute being absent from their generated images. These observations highlight the challenges of instruction following, cross-turn consistency, and error accumulation in current UMMs.

4.5 Experiments on Test-time Scaling

To mitigate exposure bias in image-generation tasks as the number of turns increases, we choose BAGEL as a representative open-source UMM and re-evaluate it after applying the three test-time scaling strategies in Section 3.5. For the turn-wise comparison, we use a maximum retry count of 1 for Self-Verification and set $N = 3$ for Best-of-N Sampling, corresponding to the configuration shown in Appendix C.2, Fig. 9. Table 4 further summarizes the best-performing configuration of each strategy according to the overall average score in the ablation study.

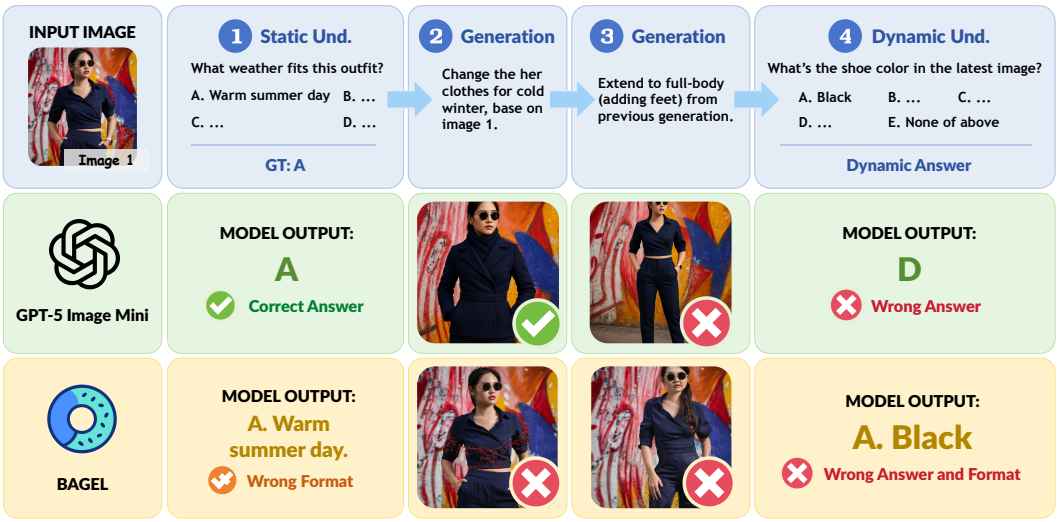

Figure 4: A representative IMUG-Bench sample illustrating the limitations of current UMMs in multi-turn interleaved interaction. Current UMMs still exhibit weak instruction-following ability, error accumulation across turns, and hallucination.

Table 4: Comparison of the BAGEL baseline with the best-performing configurations of three test-time scaling strategies. For Self-Verification and Best-of-N Sampling, we select the setting with the highest overall average from the ablation study in Appendix C.3. Scores are image-modality results across the three IMUG-Bench classes. The three test-time scaling methods are highlighted as **gold**, **silver**, and **bronze** according to their overall average scores.

Method / Setting	Static Spatial	Temporal Causal	Hybrid	Avg
Baseline	57.9	49.3	52.8	54.1
Chain-of-Thought	**72.0**	**60.1**	**60.4**	**65.9**
Self-Verification (Max=1)	**59.0**	**50.7**	**53.3**	**55.2**
Best-of-N Sampling (N=5)	**62.0**	**52.7**	**60.1**	**58.6**

Results show that all three strategies improve the overall image-question score and slow performance decay across turns. As shown in Table 4, Chain-of-Thought achieves the strongest overall performance, while the best-performing Self-Verification and Best-of-N Sampling configurations also outperform the BAGEL baseline. The latter two methods reduce error accumulation at additional inference cost, although substantial multi-turn degradation remains. Detailed parameter ablations are reported in Appendix C.3.

5 Conclusion

In this paper, we introduce IMUG-Bench, comprising 3,113 samples and 12,034 dialogue turns across 19 domains and 97 tasks, for evaluating UMM understanding and generation in multi-turn interleaved text–image interactions. Experiments on mainstream open-source and closed-source UMMs yield three main findings. First, generation performance degrades with longer interactions, revealing generation-side exposure bias and error accumulation. Second, closed-source models achieve higher and more balanced understanding and generation performance, while open-source models fluctuate more and remain weak on fine-grained spatial and geometry-related tasks. Third, Chain-of-Thought, Self-Verification, and Best-of-N Sampling improve multi-turn image generation and reduce performance decay, with Chain-of-Thought yielding the largest gains. Despite these improvements, substantial degradation remains in long-horizon interactions. Overall, IMUG-Bench supports systematic analysis of UMM capability boundaries, failure modes, and test-time optimization toward more robust and controllable multi-turn multimodal systems.

AI USE STATEMENT

Generative AI tools were used as part of the benchmark construction and evaluation pipeline, including LLM/VLM-assisted template filling, dynamic answer generation, and VLM-based evaluation, as described in the paper. AI-generated benchmark content was manually verified. Generative AI tools were also used extensively for code implementation, debugging, and refactoring, as well as for language editing and polishing of the manuscript, including improvements to grammar, clarity, and readability. The research questions, benchmark design, experimental design, analysis, and conclusions were determined by the authors. All AI-assisted code and text were reviewed and validated by the authors.

ETHICS STATEMENT

IMUG-Bench uses images collected from public platforms, model-generated images, and author-collected or professionally created content. All images are manually reviewed before inclusion, and only those satisfying the task requirements and quality, licensing, and safety checks are retained. AI-generated benchmark content is also manually verified. Nevertheless, the benchmark may inherit biases or limitations from its underlying image sources and models, which should be considered when interpreting the results.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the benchmark construction, evaluation protocols, experimental settings, and test-time scaling strategies in the main paper. The appendix further provides image-source and data statistics, supplementary experimental results and ablations, human validation of the VLM-as-a-judge protocol, and the prompts used throughout benchmark construction and evaluation.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey. *Findings of the association for computational linguistics: ACL 2024*, pp. 13590–13618, 2024.

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*, 2024a.

Hang Chen, Qian Xiang, Jiaxin Hu, Meilin Ye, Chao Yu, Hao Cheng, and Lei Zhang. Comprehensive exploration of diffusion models in image generation: a survey. *Artificial Intelligence Review*, 58(4):99, 2025a.

Jiuhai 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*, 2025b.

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 Computer Vision and Pattern Recognition Conference*, pp. 8073–8082, 2025c.

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 IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking video-llms on video spatio-temporal reasoning. *arXiv preprint arXiv:2503.11495*, 2025.

Wei Chow, Jiachun Pan, Yongyuan Liang, Mingze Zhou, Xue Song, Liyu Jia, Saining Zhang, Siliang Tang, Juncheng Li, Fengda Zhang, Weijia Wu, Hanwang Zhang, and Tat-Seng Chua. WEAVE: Unleashing and Benchmarking the In-context Interleaved Comprehension and Generation, November 2025. URL http://arxiv.org/abs/2511.11434. arXiv:2511.11434 [cs].

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Hao Du, Bo Wu, Yan Lu, and Zhendong Mao. Svlta: Benchmarking vision-language temporal alignment via synthetic video situation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13798–13809, 2025.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, Xunsong Li, Yifu Li, Shanchuan Lin, Zhijie Lin, Jiawei Liu, Shu Liu, Xiaonan Nie, Zhiwu Qing, Yuxi Ren, Li Sun, Zhi Tian, Rui Wang, Sen Wang, Guoqiang Wei, Guohong Wu, Jie Wu, Ruiqi Xia, Fei Xiao, Xuefeng Xiao, Jiangqiao Yan, Ceyuan Yang, Jianchao Yang, Runkai Yang, Tao Yang, Yihang Yang, Zilyu Ye, Xuejiao Zeng, Yan Zeng, Heng Zhang, Yang Zhao, Xiaozheng Zheng, Peihao Zhu, Jiaxin Zou, and Feilong Zuo. Seedance 1.0: Exploring the boundaries of video generation models, 2025. URL https://arxiv.org/abs/2506.09113.

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

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation, 2023. URL https://arxiv.org/abs/2211.13221.

Xinmiao Huang, Qisong He, Zhenglin Huang, Boxuan Wang, Zhuoyun Li, Guangliang Cheng, Yi Dong, and Xiaowei Huang. Spatial-dise: A unified benchmark for evaluating spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.13394*, 2025.

Yi Li, Haonan Wang, Qixiang Zhang, Boyu Xiao, Chenchang Hu, Hualiang Wang, and Xiaomeng Li. Unieval: Unified holistic evaluation for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.10483*, 2025a.

Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2779–2790, 2025b.

Chia Xin Liang, Pu Tian, Caitlyn Heqi Yin, Yao Yua, Wei An-Hou, Li Ming, Xinyuan Song, Tianyang Wang, Ziqian Bi, and Ming Liu. A comprehensive survey and guide to multimodal large language models in vision-language tasks. *arXiv preprint arXiv:2411.06284*, 2024.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a.

Rosanne Liu, Dan Garrette, Chitwan Saharia, William Chan, Adam Roberts, Sharan Narang, Irina Blok, RJ Mical, Mohammad Norouzi, and Noah Constant. Character-aware models improve visual text rendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16270–16297, 2023b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180, 2023.

Qingyu Shi, Jinbin Bai, Zhuoran Zhao, Wenhao Chai, Kaidong Yu, Jianzong Wu, Shuangyong Song, Yunhai Tong, Xiangtai Li, Xuelong Li, et al. Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. *arXiv preprint arXiv:2505.23606*, 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Zhaolong Su, Wang Lu, Hao Chen, Sharon Li, and Jindong Wang. Unigame: Turning a unified multimodal model into its own adversary. *arXiv preprint arXiv:2511.19413*, 2025.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11888–11898, 2023.

Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, Hongsheng Li, Yu Qiao, and Jifeng Dai. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer, 2024. URL https://arxiv.org/abs/2401.10208.

Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, et al. Ovis-u1 technical report. *arXiv preprint arXiv:2506.23044*, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024b.

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. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023a.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 2247–2256. IEEE, 2023b.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023c.

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.

Teng Xiao, Zuchao Li, and Lefei Zhang. Omnibridge: Unified multimodal understanding, generation, and retrieval via latent space alignment. *arXiv preprint arXiv:2509.19018*, 2025.

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.

Qinghongbing Xie, Zhaoyuan Xia, Feng Zhu, Lijun Gong, Ziyue Li, Rui Zhao, and Long Zeng. Gtr-bench: Evaluating geo-temporal reasoning in vision-language models. *arXiv preprint arXiv:2510.07791*, 2025a.

Wulin Xie, Yi-Fan Zhang, Chaoyou Fu, Yang Shi, Bingyan Nie, Hongkai Chen, Zhang Zhang, Liang Wang, and Tieniu Tan. Mme-unify: A comprehensive benchmark for unified multimodal understanding and generation models. *arXiv preprint arXiv:2504.03641*, 2025b.

Yan Yang, Haochen Tian, Yang Shi, Wulin Xie, Yi-Fan Zhang, Yuhao Dong, Yibo Hu, Liang Wang, Ran He, Caifeng Shan, et al. A survey of unified multimodal understanding and generation: Advances and challenges. *Authorea Preprints*, 2025.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A Survey on Multimodal Large Language Models. *National Science Review*, 11(12):nwae403, November 2024. ISSN 2095-5138, 2053-714X. doi: 10.1093/nsr/nwae403. URL http://arxiv.org/abs/2306.13549. arXiv:2306.13549 [cs].

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yu-Gang Jiang, and Xipeng Qiu. AnyGPT: Unified multimodal LLM with discrete sequence modeling. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9637–9662, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.521. URL https://aclanthology.org/2024.acl-long.521/.

Chi Zhang, Jiepeng Wang, Youming Wang, Yuanzhi Liang, Xiaoyan Yang, Zuoxin Li, Haibin Huang, and Xuelong Li. Unimodel: A visual-only framework for unified multimodal understanding and generation. *arXiv preprint arXiv:2511.16917*, 2025.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12401–12430, 2024.

Tianyi Zhang, Zheng Wang, Jing Huang, Mohiuddin Muhammad Tasnim, and Wei Shi. A survey of diffusion based image generation models: Issues and their solutions. *arXiv preprint arXiv:2308.13142*, 2023.

Yi Zhang, Yinda Chen, Che Liu, Zeyuan Ding, Jin Xu, Shilong Zou, Junwei Liao, Jiayu Hu, Xiancong Ren, Xiaopeng Zhang, Yechi Liu, Haoyuan Shi, Zecong Tang, Haosong Sun, Renwen Cui, Kuishu Wu, Wenhai Liu, Yang Xu, Yingji Zhang, Yidong Wang, Senkang Hu, Jinpeng Lu, Nga Teng Chan, Yechen Wu, Zeting Liu, Xianzhou Hou, Yong Dai, Jian Tang, and Xiaozhu Ju. Pelican-unify 1.0: A unified embodied intelligence model for understanding, reasoning, imagination and action, 2026. URL https://arxiv.org/abs/2605.15153.

Shanshan Zhao, Xinjie Zhang, Jintao Guo, Jiakui Hu, Lunhao Duan, Minghao Fu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, et al. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*, 2025.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 56–66, 2025.

Kai Zou, Ziqi Huang, Yuhao Dong, Shulin Tian, Dian Zheng, Hongbo Liu, Jingwen He, Bin Liu, Yu Qiao, and Ziwei Liu. Uni-mmmu: A massive multi-discipline multimodal unified benchmark. *arXiv preprint arXiv:2510.13759*, 2025.

APPENDIX

The appendix provides additional details and supporting materials for IMUG-Bench, organized as follows.

Appendix A describes the image sources and curation process used in benchmark construction.

Appendix B presents the data distribution of IMUG-Bench, including visualizations of the class/domain distribution, as well as detailed information regarding domain-task composition, interleaved modality sequences, and the number of interaction turns.

Appendix C reports additional experimental results, including radar-chart visualizations of task-specific model performance, score trends over interaction turns, test-time scaling ablations, supplementary results on Static-MCQ and Dynamic-MCQ, and MCQ formatting-weight sensitivity analysis.

Appendix D presents the human validation results for the VLM-as-a-judge scoring protocol.

Appendix E provides the prompts used for calling LLMs or VLMs during question filling, model evaluation, dynamic MCQ answer and image assessment retrieval in IMUG-Bench.

A IMAGE SOURCES AND CURATION

We construct the benchmark using three complementary image sources.

- ***Public image platforms.*** Using the finalized question templates and keyword lists, we batch-collect images that match the target scenarios from platforms such as Pexels, Pixabay, Openverse, Flickr, Unsplash, Wikimedia, and CloudAtlas.
- ***Targeted generation with image models.*** For image concepts that are difficult to collect from public platforms, we use image models such as GPT Image, Nano Banana, and Grok to generate candidate images.
- ***Real-world capture and professional content creation.*** When public platforms and model generation cannot provide high-quality images, we supplement the dataset through real-world capture or professional tools.

These sources cover diverse scenarios and styles, helping reduce source-specific bias and improve coverage of real interactive settings. All images are manually reviewed before inclusion, and only those that match the task requirements and satisfy quality, licensing, and safety checks are retained.

B DATA OVERVIEW OF IMUG-BENCH

Figure 5 provides a high-level view of the benchmark distribution. The left side shows pie-chart visualizations of the class/domain composition of IMUG-Bench, and the right side shows a word cloud of the most frequent concepts, which together help illustrate the overall coverage and topic diversity of the benchmark.

Table 5 provides a detailed breakdown of the sample counts across each domain and task within IMUG-Bench, alongside their respective modality distributions (modality sequence) and interaction turn statistics.

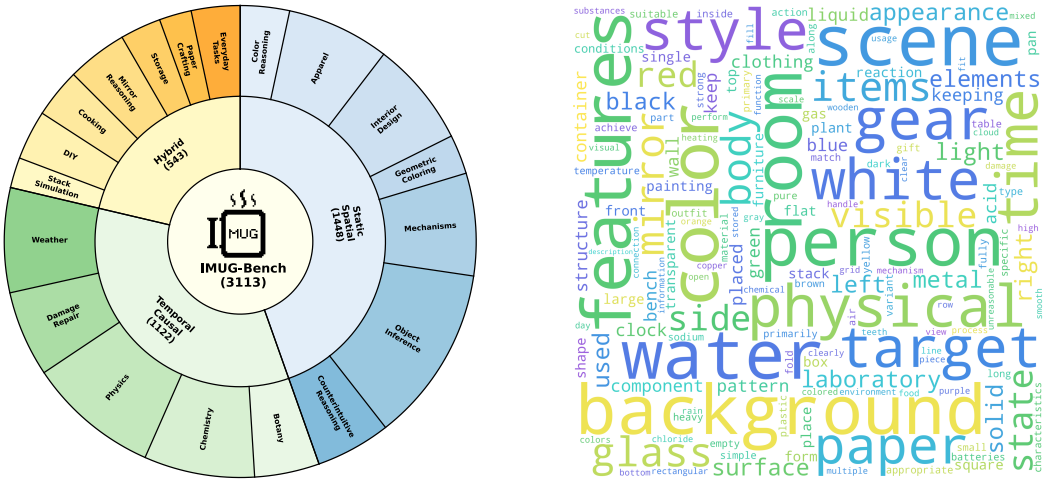

Figure 5: IMUG-Bench statistics and distribution. Left: pie-chart visualizations of the class/domain distribution of IMUG-Bench; Right: a word cloud summarizing the most frequent concepts.

Table 5: Data composition and distribution summary table of IMUG-Bench. The numbers following Domain and Task indicate the number of samples within that partition. Turns represent the corresponding number of interaction turns under this interleaved sequence. Modality Sequence illustrates the possible alternating sequence of text–image output modalities under each task, where T represents text and I represents image.

Domain	Task	Turns	Modality Sequence
Apparel (233)	Background Replacement (21)	4	[T, I, T, I]
		4	[I, T, T, I]
	Facial Feature Modification (18)	2	[I, T]
		2	[T, I]
	Full Body Extension (26)	5	[I, T, I, T, I]
		5	[T, I, I, T, I]
	Garment Replacement (26)	3	[I, T, I]
		3	[T, I, I]
	Gender Swapping (47)	4	[T, I, I, I]
		4	[I, T, I, I]
	Multi Person Portrait (27)	2	[I, T]
		3	[T, T, I]
		2	[T, I]
	Pose Alteration (48)	3	[T, I, I]
		3	[I, T, I]
	Scene Adaptation (20)	4	[T, I, T, I]
		4	[I, T, T, I]
Botany (132)	Abiotic Stress (20)	3	[I, T, I]
	Microbial Culture (15)	3	[I, I, T]
	Natural Growth (25)	3	[I, T, I]
		3	[I, I, T]
	Nutrient Deficiency (15)	3	[I, T, I]
	Plant Identification (22)	3	[T, T, T]
	Tropisms (35)	3	[I, T, I]
		3	[I, I, T]

Continued on next page

Table 5 – continued from previous page

Domain	Task	Turns	Modality Sequence
Chemistry (254)	Acid Reaction (21)	4	[I, T, I, T]
		4	[T, I, I, T]
		4	[I, T, I, I]
		4	[T, I, I, I]
	Apparatus Usage (29)	4	[I, T, I, T]
		4	[T, I, I, T]
	Ball Stick Model (12)	4	[I, T, I, I]
		4	[T, I, I, I]
	Combustion Flame (28)	3	[T, I, I]
		3	[I, T, I]
		5	[I, T, I, I, T]
		5	[T, I, I, I, T]
	Crystallization (7)	4	[T, I, T, I]
		4	[I, T, T, I]
	Daily Chemistry (22)	5	[I, T, T, I, T]
		5	[T, I, T, I, T]
	Gas Preparation (11)	4	[I, T, T, T]
		4	[T, I, T, T]
	Lab Accident (42)	5	[I, T, I, T, T]
		5	[T, I, I, T, I]
		5	[T, I, I, T, T]
		5	[I, T, I, T, I]
	Metal Displacement (14)	4	[I, T, T, I]
		4	[T, I, T, I]
		4	[I, T, I, T]
		4	[T, I, I, T]
	Ph Measurement (10)	4	[I, T, T, I]
		4	[T, I, T, I]
	Precipitation (20)	4	[I, T, I, T]
		4	[I, T, T, I]
		4	[T, I, T, I]
		4	[T, I, I, T]
	Specific Color Reaction (38)	3	[T, I, I]
		3	[I, T, T]
		3	[T, I, T]
		3	[I, T, I]
Color Reasoning (90)	Additive Lighting Mixing (30)	5	[I, I, T, I, T]
	Liquid Color Mixing (30)	5	[T, I, I, T, T]
	Subtractive Color Mixing (30)	4	[I, I, T, T]
Cooking (90)	Cooking Workflow (30)	3	[T, T, I]
	Food Frying Compare (30)	3	[I, I, T]
	Food Frying Flip (30)	6	[I, I, T, I, I, T]
Counterintuitive Reasoning (159)	Action Relation Mismatch (40)	3	[I, T, I]
	Compound Mismatch (39)	4	[T, I, T, I]
	Object Attribute Mismatch (40)	3	[I, T, I]
	Scene Object Mismatch (40)	3	[I, T, I]
DIY (91)	Gift Wrapping Logic (31)	6	[T, I, T, I, T, I]
	Rubber Stamp Carving (30)	4	[I, T, I, I]
	Wax Seal Process (30)	4	[I, I, I, I]
		4	[I, I, I, T]
		4	[I, I, T, I]

Continued on next page

Table 5 – continued from previous page

Domain	Task	Turns	Modality Sequence
Damage Repair (188)	Disaster Analysis (38)	6	[T, T, I, T, T, I]
		3	[T, T, I]
	Item Repair (50)	4	[T, T, T, I]
	Relic Restoration (50)	3	[I, T, I]
	Rust Restoration (50)	6	[I, T, I, T, I, T]
Everyday Tasks (89)	Battery Installation (30)	4	[I, T, I, T]
	Glass Cleaning (29)	4	[I, T, I, T]
	Wound Treatment (30)	3	[T, I, I]
Geometric Coloring (64)	Cubes Coloring (10)	3	[T, I, I]
	Cubes Completion (7)	3	[T, I, T]
	Cubes Synthesis (7)	6	[T, I, T, T, I, I]
	Grid Coloring (40)	4	[I, I, T, I]
		5	[I, I, T, I, T]
		5	[I, T, I, T, I]
Interior Design (240)	Artwork Addition (40)	6	[I, T, I, I, I, T]
	Furniture Arrangement (40)	5	[I, T, I, I, T]
	Global Style Modification (40)	3	[I, T, I]
	Local Style Modification (40)	3	[T, T, I]
	Room Renovation (40)	2	[T, I]
	Scene Completion (40)	2	[T, I]
Mechan-isms (232)	Gear Completion (40)	2	[T, I]
	Gear Meshing (40)	3	[T, T, I]
	Object Pairing (30)	2	[T, I]
	Rotational Mechanism Edit (31)	2	[I, I]
	Rotational Mechanism Generation (31)	3	[T, I, I]
	Sliding Mechanism Edit (30)	2	[I, I]
	Sliding Mechanism Generation (30)	3	[T, I, T]
		2	[T, I]
		3	[T, I, I]
Mirror Reasoning (120)	Mirror Action Editing (40)	3	[I, T, I]
	Mirror Clock Person Composite (40)	4	[I, I, T, T]
	Mirror Clock Reading (40)	2	[T, I]
		3	[T, I, T]
Object Inference (430)	Consumer Electronics (55)	5	[I, I, I, T, T]
		6	[I, I, I, T, I, T]
	Daily Essentials (11)	5	[I, I, I, T, T]
		6	[I, I, I, T, I, T]
	Furniture And Interior (77)	5	[I, I, I, T, T]
		6	[I, I, I, T, I, T]
	Hardware Tools (55)	5	[I, I, I, T, T]
		6	[I, I, I, T, I, T]
	Kitchenware And Appliances (67)	5	[I, I, I, T, T]
		6	[I, I, I, T, I, T]
	Recreation And Instruments (55)	5	[I, I, I, T, T]
		6	[I, I, I, T, I, T]
	Scientific And Professional Equip (55)	5	[I, I, I, T, T]
		6	[I, I, I, T, I, T]
	Stationery And Office (55)	5	[I, I, I, T, T]
		6	[I, I, I, T, I, T]

Continued on next page

Table 5 – continued from previous page

Domain	Task	Turns	Modality Sequence
Paper Crafting (45)	Cutting (15)	5	[I, T, I, T, I]
	Origami Basic Geometry (15)	4	[I, T, I, I]
	Origami Sequential Logic (15)	6	[I, T, I, I, T, T]
Physics (310)	Clock Operation (40)	4	[I, I, T, T]
	Deformation (30)	3	[T, I, I]
	Diffusion Tyndall (30)	4	[I, I, T, I]
	Magnetism Advanced (30)	3	[I, T, I]
	Magnetism Basic (30)	3	[I, T, I]
	Magnetism Compare (30)	4	[I, I, T, T]
	Mass Measurement (30)	4	[T, I, I, I]
	Refraction Multi Layer (30)	4	[I, T, T, I]
	Refraction Single (30)	4	[T, I, T, I]
		2	[I, I]
		3	[T, I, I]
	State Change (30)	6	[T, I, T, I, I, T]
Stack Simulation (40)	Stack Math Expression (20)	5	[I, T, I, T, T]
	Stack Word Formation (20)	5	[I, T, I, T, I]
Storage (68)	Elastic Storage (68)	4	[I, I, T, T]
Weather (238)	Cloud Dynamics (30)	2	[I, I]
	Cloud State Evolution (27)	3	[T, T, I]
		2	[T, I]
	Cumulative Effects (61)	3	[I, T, I]
		2	[I, I]
	Rain Gradient Visuals (30)	2	[I, I]
		2	[T, I]
	Solar Lighting (30)	5	[I, I, I, I, I]
	Visibility Obscuration (30)	4	[I, I, T, T]
	Wind Force Modification (30)	2	[T, I]

C Additional Experimental Results

C.1 Radar Charts of Model Performance

This subsection visualizes the task-specific scores for all tested models via radar charts. Specifically, Figs. 6, 7, and 8 display the results for all modalities, text-only modalities, and image-only modalities.

In these radar charts, lines of different colors represent the performance scores of various models. Based on the specific trends of the curves, several significant common characteristics across models can be identified. For the text modality, the six open-source UMMs generally perform poorly in Mirror Reasoning and Color Reasoning tasks, while showing relatively superior performance in tasks such as Cooking and Counterintuitive Reasoning. Notably, some open-source models achieve scores closely approaching those of closed-source UMMs in a wide range of tasks. Regarding the image modality, all six open-source UMMs exhibit a relative deficiency in the Chemistry task, a weakness not observed in the two closed-source models. Furthermore, the scores of open-source UMMs across all modalities remain consistently low in Stack Simulation, Paper Crafting, and Geometric Coloring tasks.

These results reveal several key insights. First, the performance gap between open-source UMMs and closed-source models is narrower in understanding tasks compared to generation tasks; specifically, in relatively daily or simple scenarios, certain open-source UMMs have approached the perfor-

mance level of closed-source models, though a significant gap remains in complex logical reasoning. Second, the generative capabilities of open-source UMMs in highly specialized domains—such as chemical substances, laboratory equipment, and stack simulation—require further enhancement, as they often struggle to accurately render professional domain-specific diagrams requested by users. Finally, for tasks demanding the highest level of integrated spatio-temporal reasoning and generation across all modalities, all open-source UMMs show substantial room for improvement. Even in extremely complex tasks like Paper Crafting (paper folding and cutting) and Mirror Reasoning (rendering clocks and human actions in mirrors), closed-source models also exhibit noticeably lower scores compared to their performance in other tasks.

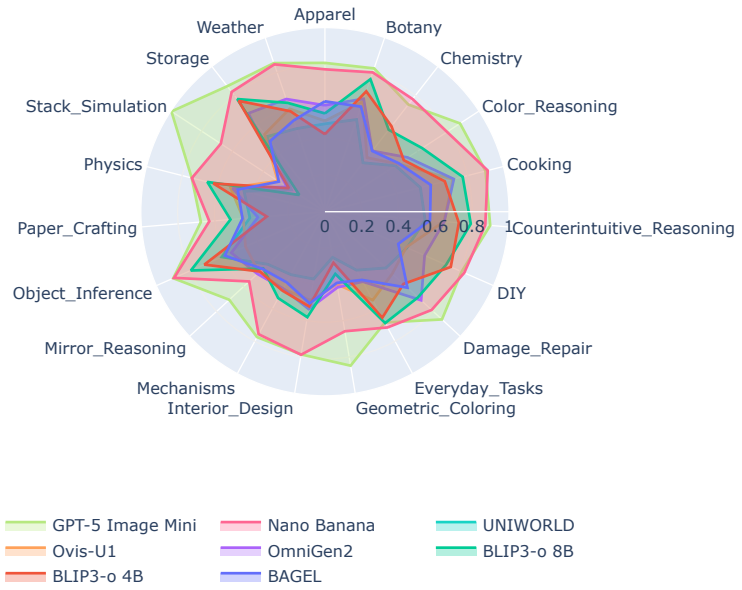

Figure 6: Scores of all modalities for each model across all tasks.

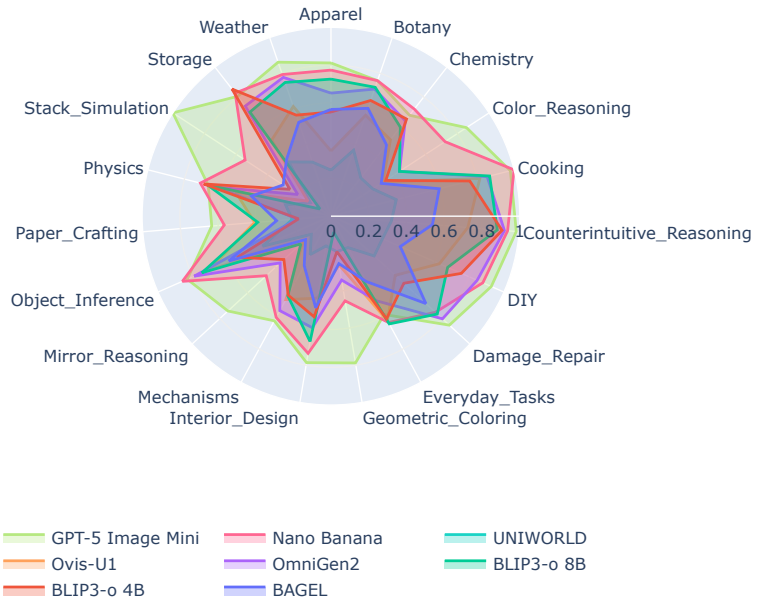

Figure 7: Scores of text-only modalities for each model across all tasks.

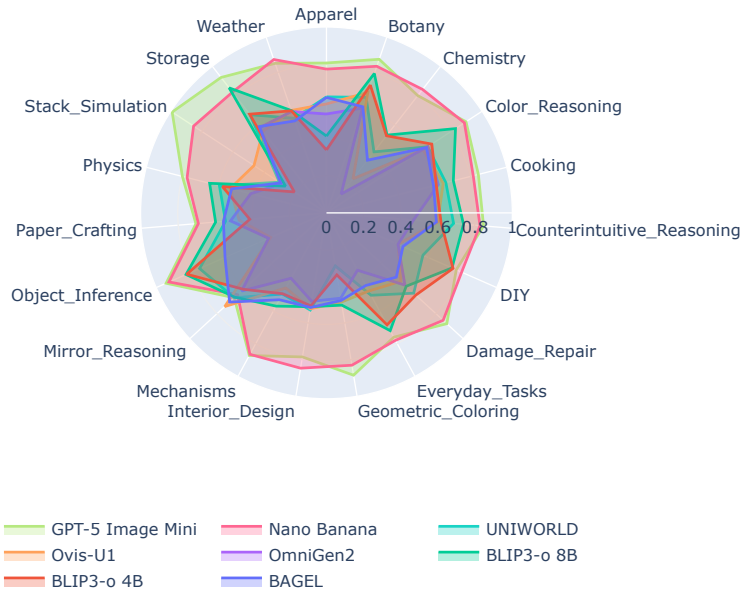

Figure 8: Scores of image-only modalities for each model across all tasks.

C.2 Score Trends over Interaction Turns

Figure 9 provides a more intuitive visualization of these trends, including the comparison between the BAGEL baseline and the three test-time scaling strategies.

C.3 Ablation Study on Test-time Scaling

To further investigate the impact of test-time scaling methods on model generation performance, we conduct parameter ablations on Self-Verification and Best-of-N Sampling using BAGEL as the base model. The detailed results are reported in Table 6.

- **Ablation Study on Self-Verification**: We set the maximum retry count in the Self-Verification process to 0 (the baseline without verification), 1, 2, and 3, respectively. The resulting image-modality scores are presented in Table 6. Analysis of the results indicates that the peak image scores across all three classes, as well as the average image score, occur between 1 or 2 retries. However, performance degrades as the retry count increases beyond 2, occasionally underperforming the baseline. This experimental evidence demonstrates that Self-Verification effectively enhances image generation quality within a certain range; however, overly frequent retries can be counterproductive.
- **Ablation Study on Best-of-N Sampling**: In a similar vein, we evaluate the Best-of-N Sampling method with N ranging from 1 to 5, where $N = 1$ serves as the baseline (no selection). The image scores, detailed in Table 6, generally follow an upward trajectory with increasing N. Notably, the rate of improvement slows down markedly as N increases, with the scores for $N = 4$ and $N = 5$ showing minimal variance. This evidence confirms that the quality gains are consistent rather than coincidental, while simultaneously revealing the inherent performance boundary of the approach.

C.4 Supplementary Results on Static-MCQ and Dynamic-MCQ

Table 7 presents the detailed scores of all evaluated models on Static-MCQ and Dynamic-MCQ. These results provide numerical evidence and further support the corresponding findings discussed in Section 4.2.

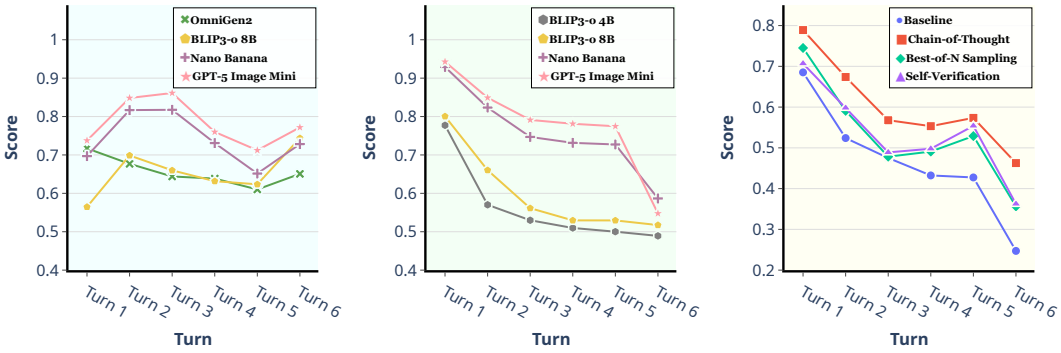

Figure 9: Score trends over interaction turns. Left: text-modality question scores of the top-4 models. Middle: image-modality question scores of the top-4 models. Right: BAGEL baseline compared with three test-time scaling strategies across turns.

Table 6: Image-modality scores under three test-time scaling methods and their corresponding settings. We report generation scores with the overall average score. For each metric, we first identify the best-performing setting within each test-time scaling method, and then assign medal rankings across methods based on these method-level best results; the resulting top-3 methods are highlighted as **gold** (best), **silver** (second best), and **bronze** (third best).

Method / Setting	Static Spatial	Temporal Causal	Hybrid	Average
Baseline	57.9	49.3	52.8	54.1
Chain-of-Thought	**72.0**	**60.1**	**60.4**	**65.9**
S-V (Max=1)	59.0	**50.7**	**53.3**	**55.2**
S-V (Max=2)	**60.6**	49.4	48.2	54.6
S-V (Max=3)	58.7	49.2	48.6	53.7
B-of-N (N=2)	58.0	50.1	54.9	54.8
B-of-N (N=3)	61.6	50.6	58.6	57.4
B-of-N (N=4)	61.9	**52.9**	58.0	58.2
B-of-N (N=5)	**62.0**	52.7	**60.1**	**58.6**

Table 7: Comparison of model performance on Static-MCQ and Dynamic-MCQ.

Model Name	Static-MCQ	Dynamic-MCQ
GPT-5 Image Mini[1]	83.5	50.6
Nano Banana[2]	80.2	43.4
BAGEL[3]	52.7	25.3
BLIP3-o 4B[4]	58.8	46.4
BLIP3-o 8B[4]	67.3	47.0
OmniGen2[5]	72.3	30.0
Ovis-U1[6]	56.1	27.8
UNIWORLD-V1[7]	29.2	20.9

[1] Singh et al. (2025); [2] Comanici et al. (2025); [3] Deng et al. (2025); [4] Chen et al. (2025b); [5] Wu et al. (2025); [6] Wang et al. (2025); [7] Lin et al. (2025).

C.5 Sensitivity Analysis of MCQ Formatting Weights

To examine the sensitivity of w_{fmt}, we evaluate three alternative settings for $(w_{\text{sep}}, w_{\text{exp}})$, where w_{sep} and w_{exp} denote the weights for format deviations without explanation and outputs with additional explanatory text, respectively. As shown in Table 8, varying these weights changes the absolute scores but does not affect the final model ranking.

Table 8: w_{fmt} sensitivity on overall scores/ranks. Each cell reports score/rank.

$(w_{\text{sep}}, w_{\text{exp}})$	GPT-5 Mini	Nano Banana	BLIP3-o 8B	BLIP3-o 4B	OmniGen2	BAGEL	Ovis-U1	UNIWORLD-V1
$(0.75, 0.50)$	82.3/1	79.5/2	65.0/3	59.3/4	55.1/5	51.9/6	50.0/7	46.6/8
$(0.85, 0.60)$	82.3/1	79.5/2	65.0/3	59.7/4	55.2/5	52.7/6	50.7/7	48.8/8
$(0.50, 0.25)$	82.3/1	79.4/2	65.0/3	58.4/4	54.9/5	49.9/6	48.1/7	41.0/8
$(0.50, 0.00)$	82.3/1	79.4/2	65.0/3	57.8/4	54.8/5	48.1/6	46.3/7	35.5/8

D Human Validation

To validate the reliability of the VLM-as-a-judge evaluation, we use the Pearson correlation coefficient to measure the agreement between our automated scores and human ratings. We conduct human validation on all eight evaluated UMMs. To ensure representativeness, for each model we randomly sample one task-sample response from each task and perform two-person manual rescoring. We then compute the Pearson correlation coefficient between human ratings and VLM-as-a-judge scores; the results are shown in Table 9. The Pearson correlations exceed 0.72 for all row-wise (model-level) and column-wise (class-level) averages, with an overall mean above 0.8, indicating that our evaluation method closely matches human scoring and supporting the effectiveness of VLM-as-a-judge.

Table 9: Human validation of VLM-as-a-judge scoring across all eight evaluated models. We report Pearson correlations between automated and two-person human scores for Img. and Dynamic-MCQ across classes and on average.

Model Name	St.&Sp.		Te.&Ca.		Hybrid		Avg
	Img.	Dynamic	Img.	Dynamic	Img.	Dynamic	
GPT-5 Image Mini[1]	0.725	0.813	0.662	0.839	0.771	0.884	0.782
Nano Banana[2]	0.839	0.926	0.781	0.816	0.706	0.855	0.821
BAGEL[3]	0.701	0.794	0.815	0.866	0.747	0.820	0.791
BLIP3-o 4B[4]	0.751	0.843	0.799	0.816	0.698	0.894	0.800
BLIP3-o 8B[4]	0.759	0.821	0.702	0.916	0.735	0.830	0.794
OmniGen2[5]	0.693	0.802	0.775	0.860	0.740	0.831	0.784
Ovis-U1[6]	0.783	0.901	0.820	0.807	0.701	0.922	0.822
UNIWORLD-V1[7]	0.837	0.861	0.878	0.871	0.693	0.907	0.841
Average	0.761	0.845	0.779	0.849	0.724	0.868	0.804

[1] Singh et al. (2025); [2] Comanici et al. (2025); [3] Deng et al. (2025); [4] Chen et al. (2025b); [5] Wu et al. (2025); [6] Wang et al. (2025); [7] Lin et al. (2025).

E Prompts

In this section, we list the prompts employed throughout the construction, testing, and scoring pipelines of IMUG-Bench.

E.1 Prompts for Benchmark Construction

The following prompt is employed for task sample filling and construction. During the construction process, this prompt is immediately followed by the question template and keyword sequences to be filled; if images are involved, they are provided as input simultaneously.

```
[System Prompt]
1. Role
You are an expert Dataset Generator for a Multimodal Benchmark.

2. Task
Your task is to instantiate test samples by combining Paradigm
Templates with Blueprint Variables.

3. Execution Logic
3.1 Analyze Context
Specify every keyword from the Blueprint.

3.2 Fill Placeholders
Replace all <...> in the "prompt_template" with Context Variables.
Constraint: Do NOT rewrite the template text outside brackets.
CRITICAL: You must reconstruct the sentence to ensure it sounds like
it was written by a native English speaker.

3.3 Construct Options
Generate 4 specific options (A-D).
Option E must be "None of the above".
Randomization: Randomize the position of the correct option.

3.4 Construct Evaluation Points
Fill and expand specific checks into detailed points. Please confirm
that all evaluation-points are clear and unambiguous.

4. Output Format
Output ONLY valid JSON. No markdown.

{
    "sample_id": <Int>,
    "category": "<String, Subdomain>",
    "total_tasks": <Int>,
    "tasks": [
        {
            "turn": <Int>,
            "modality": "text",
            "input": [
                {
                    "text": "<String: English Question +
                    \nOptions:\nA. ...\nB. ...\nC. ...\nD. ...\nE.
                    None of the above.>"
                }
            ],
            "output": {
                "answer": "<String: Correct Letter or '<DYNAMIC>'>"
            }
        }
    ]
}

5. Critical Constraints
5.1 Logic Consistency
Ensure strict physical and logical consistency.

5.2 Anti-Leakage
Current question stem must NOT reveal the current answer.
Subsequent question stems must NOT reveal answers to previous turns.

5.3 Conciseness
Keep descriptions brief and accurate.

[Question Template]
...
```

```
[Keyword-Value Pairs]
...

[Image Sequence (if applicable)]
...
```

E.2 Prompts for Model Evaluation

Below is the system prompt provided to the model during the IMUG-Bench evaluation process. This prompt is immediately followed by the main body of the test questions.

```
[System Prompt]
1. Role
You are an AI assistant participating in a multimodal benchmark
evaluation.

2. Task
This is a multi-turn multimodal task. You must synthesize all history
information (including all prompt texts, input images, and your own
previous text/image outputs) to determine your response strategy.
Inputs may contain text and single or multiple images.

3. Rules
3.1 Output Modality Constraints
Your response for each turn must be strictly EITHER TEXT OR IMAGE.
If text output is required, treat it as a Multiple Choice Question
(MCQ).
If image output is required, treat it as a Generating Task.

3.2 MCQ Guidelines
All MCQs are indeterminate choice (can be single or multiple
selection). Determine the number of correct options based on the
prompt.
Output Format: Output ONLY a contiguous string of uppercase letters
representing the correct options, with no separators or punctuation
(e.g., 'A' or 'BCD'). Do NOT output option content or explanations.

3.3 Generating Task Guidelines
Generate one image based on all context.
Hint: Unless explicitly instructed to "generate a new image" or
"modify a specific image", the images you generate should by default
reference the most recent image in the context (whether input or
previously generated by you) for further modification, or be
re-created based on certain elements within it

[Multi-turn Tasks]
...
```

E.3 Prompts for Dynamic Answer Extraction

Below is the prompt used to retrieve the standard answers for the dynamic MCQs. This prompt is input to the judge VLM along with the reference dialogue turns.

```
[System Prompt]
1. Role
You are a top-tier Multimodal Dialogue Benchmark Evaluation Expert.
For this task, you will serve as the "Dynamic Answer Determination
Assistant."

2. Task Background
```

```
The AI model being evaluated is participating in a multi-turn
conversation where it generates or modifies images based on user
instructions. The conversation has reached a "Dynamic Multiple-Choice
Question (MCQ)" turn. This question is "dynamic" because its
ground-truth answer is not fixed; instead, it depends entirely on the
actual content generated by the model in previous reference turns
(e.g., the number of objects drawn, the specific color of a
background, etc.).

3. Task Objective
Your goal is to analyze the reference information and determine the
correct standard answer for this context.

4. Provided Content
4.1 Reference Turns
This includes user instructions (Task Questions) from previous turns
and the actual outputs (Model Response/Image) produced by the
evaluated model. These serve as the only factual basis for determining
the answer.

4.2 Final Dynamic Question
This is the last turn's MCQ. Note that the answer field may contain a
<DYNAMIC> placeholder or partial answers (e.g., A+<DYNAMIC>, means 'A'
must be a correct answer, and you need to identify the other correct
items among the available options). You must identify all correct
options (A, B, C, etc.) based on the actual generation results.

5. Execution Logic
5.1 Trace the Logic Chain
Carefully review the requirements in the Reference Turns and observe
the final state of elements (color, position, quantity, style, etc.)
in the generated images or text.

5.2 Map to Options
Compare the "factual state" from the reference outputs against each
option in the Final Dynamic Question.

5.3 Eliminate Distractors
Even if an option is factually true in a general sense, if it does not
match the specific content actually generated by the model, it must be
considered incorrect.

6. Output Format
You must output strictly in JSON format. No additional text,
explanations, or summaries are allowed. The structure must be as
follows:

{
  "determined_answer": "A contiguous string of uppercase letters,
  e.g., 'A' or 'BCD'",
  "reasoning": "State your logic for determining the answer in detail.
  Mention which feature from which reference turn was used and how it
  maps to the chosen options."
}

[Target-Turn Dynamic-MCQ Question and Choices]
...

[Sequence of Reference Turn Q&As]
...
```

E.4 PROMPTS FOR SCORING

The following is the prompt used for image assessment. This prompt is provided to the judge VLM along with the sequence of reference images, the image to be evaluated, and their corresponding image sequence labels.

```
[System Prompt]
1. Role
You are a professional visual dialogue quality evaluation expert.

2. Task
Your task is to provide precise scoring for the image generation
result of a specific turn within a "Multi-turn Multimodal Dialogue
Benchmark." You need to judge whether the model-generated image meets
the preset "evaluation-points" based on the instructions given in that
turn, combined with historical reference information.

3. Task Background
This task is extracted from a continuous multi-turn dialogue between
an AI and a user. During this process, the user continuously provides
new instructions, requiring the AI to modify, evolve, maintain style,
or supplement logic based on previous visual information (input images
or images previously generated by the AI).

4. Provided Content
4.1 Reference Images (Image_1, Image_2, ..., Image_N-1)
These represent the key visual states that appeared earlier in the
dialogue path. They may be original materials provided by the user
(input images) or intermediate results generated by the AI (output
images).

4.2 Target Image (Image_N, the last one)
The output generated by the model according to the latest instructions
for the current turn.

5. Scoring Standards (0-5 Scale)
This is a fine-grained quality evaluation task, NOT a binary
correctness check. Please score each evaluation point independently;
scores must be integers. The score represents the "completeness" of
the target image relative to a specific evaluation point. The more
accurately the image completes the requirements and the higher the
quality, the higher the score should be.

5.1 5 points (Flawless)
Must fully comply with the current task requirements.
Must maintain high "pixel-level" or "logic-level" consistency with the
specified reference images regarding the required elements.
No visual distortions, physical logic errors, or feature loss.

5.2 4 points (Satisfactory)
The evaluation point is clearly satisfied, all required elements are
present and recognizable.
Minor inconsistencies, simplifications, or small visual artifacts may
exist, but they do not affect the overall fulfillment of this point.

5.3 3 points (Partial Achievement)
The evaluation point is generally satisfied at a semantic level.
Key elements are present, but there are noticeable visual
inaccuracies, partial mismatches with reference images, or logical
simplifications.
This score should be used when the intent is clear but execution is
imperfect.

5.4 2 points (Significant Deviation)
```

```
Only partial features required by the evaluation point are retained,
but the overall logic has shifted.
There are multiple significant visual errors or an obvious disconnect
in feature inheritance from the reference images.

5.5 1 point (Trace Evidence)
There is a significant deviation from the evaluation point
requirements, but the direction of completion shares a certain degree
of similarity with the requirements, rather than being completely
opposite or entirely unrelated.

5.6 0 points (Void)
The target image has absolutely no intersection with the content
described in the evaluation (e.g., completely collapsed random noise
or solid color blocks).

6. Output Format
Please output the evaluation results directly in JSON format,
following the format below:
{
    "evaluation_results": [
        {
            "point_id": 1,
            "score": 5,
            "reason": "explaining the specific performance of the
            target image on this point."
        },
        {
            "point_id": 2,
            "score": 0,
            "reason": "explaining the specific cause for the total
            inconsistency on this point."
        }
    ]
}

[Target-Turn Question and Evaluation-Points]
...

[Sequence of Reference and Target Images]
...
```