DO YOU SEE WHAT YOU DRAW? A SEMANTIC CLOSED-LOOP FRAMEWORK FOR HOLISTIC EVALUATION OF UNIFIED MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

As Large Vision-Language Models increasingly aim to integrate visual generation and understanding within a single parameter space, evaluating such structural unification in a cohesive manner remains a critical challenge. Current evaluation protocols predominantly treat generative and discriminative capabilities as separate tasks, leaving a gap in system-level evaluation for unified multimodal models (UMMs). In this work, we propose Self-Generative-Understanding (SGU), a novel, annotation-free evaluation framework that probes the integrated capabilities of unified models through a semantic closed-loop challenge. Without requiring new annotations, SGU leverages the dual understanding-and-generation abilities of UMMs by asking them to first perceive an image and produce a textual description, subsequently reconstruct a visual context based on that description, and finally perform reasoning over the self-generated output. This pipeline reuses existing benchmarks to yield an integrated performance score specifically tailored for evaluating UMMs as unified systems. Extensive experiments show that even high-performing UMMs often struggle to reason over their own generated contexts, revealing limitations that are not captured by separate evaluations of understanding or generation alone. Our work provides a complementary holistic evaluation framework and offers a foundation for benchmarking the development of next-generation unified multimodal models.

1 INTRODUCTION

Multimodal models increasingly combine visual understanding with image generation Wang (2025); Durante et al. (2024). This development has led from task-specific systems to Unified Multimodal Models (UMMs) Li et al. (2025); Wang (2025); Wei et al. (2025); Xie et al. (2025a). By jointly training on generation and understanding objectives, UMMs consolidate discriminative reasoning and generative synthesis within a shared model, aiming to support both perception and creation in an integrated manner.

Despite the rapid architectural convergence toward unification, current evaluation protocols still primarily assess generation and understanding as separate capabilities Li et al. (2025); Liang et al. (2025), as illustrated in Figure 1(a). Typically, generative ability is quantified using distributional or text-image alignment metrics such as FID Heusel et al. (2017) and CLIPScore Hessel et al. (2022), while understanding ability is evaluated on visual question answering or multimodal reasoning benchmarks. These component-wise evaluations are important for diagnosing specific strengths and weaknesses. However, they do not directly provide a unified assessment of how well a UMM performs when understanding and generation are both involved within the same model. This question is difficult to answer by naively aggregating heterogeneous generation and understanding metrics. For example, two UMMs may exhibit opposite capability profiles, with one stronger in generation and the other stronger in understanding. Separate scores can reveal these differences, but they do not directly indicate whether the model can function well when both abilities must interact within a unified process. This motivates a complementary evaluation perspective that treats the UMM as an integrated system rather than only as a collection of separately measured components.

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.

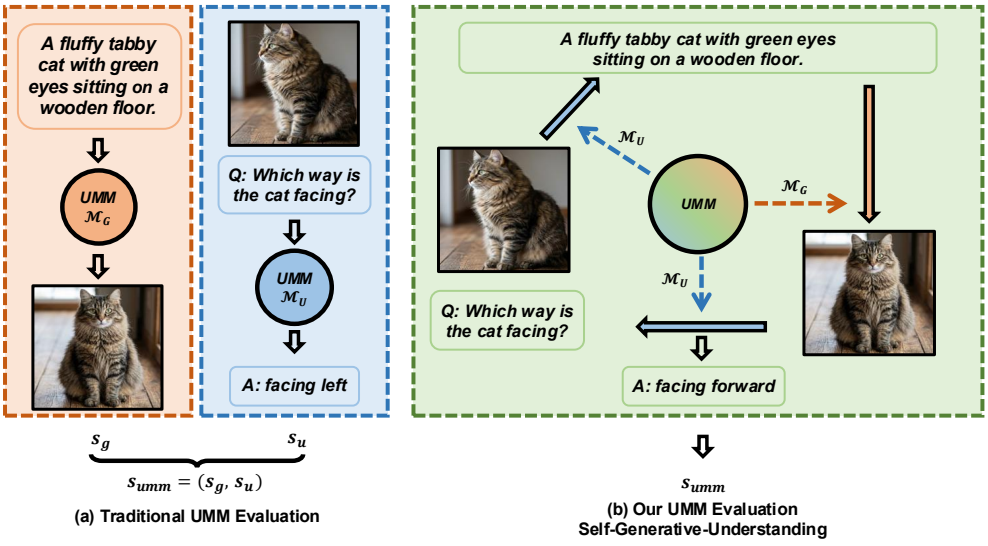

Figure 1: Illustration of traditional UMM evaluation and our Self-Generative-Understanding (SGU) framework. (a) Existing evaluations often assess generation ($\mathcal{M}_G$) and understanding ($\mathcal{M}_U$) separately, producing capability-specific scores such as s_g and s_u. These metrics remain valuable for component-wise diagnosis but do not directly provide a system-level view of how a UMM behaves when understanding and generation are jointly involved. (b) SGU provides a complementary closed-loop evaluation framework tailored to UMMs. It requires the model to understand an input image, generate a textual representation, reconstruct a visual context from that representation, and reason over the reconstructed image, yielding an outcome-based system-level score s_{umm}.

To address this gap, we propose Self-Generative-Understanding (SGU), a semantic closed-loop evaluation framework explicitly designed for UMMs. SGU uses the understanding-and-generation abilities of UMMs by asking the model to first perceive an input image and generate a textual description, then reconstruct a visual context based on that description, and finally answer image-grounded questions using the reconstructed image, as shown in Figure 1(b). Since the downstream questions are grounded in the original image, successful reasoning on the reconstructed image provides an outcome-based signal of whether the model can effectively carry out the integrated understanding-and-generation process. Weaknesses in understanding, representation, or generation can affect the final performance. The resulting SGU score is therefore a system-level performance score tailored to UMMs.

SGU offers three advantages. **(1) Integrated capability evaluation.** The closed-loop process activates multiple capabilities of UMMs, including image understanding, textual representation, visual generation, and downstream reasoning, without relying on external judge models or additional generation systems. **(2) System-level sensitivity.** The final score reflects the combined effect of these stages: information loss in understanding, intermediate representation, generation, or self-reasoning can reduce the final performance. Thus, SGU evaluates whether the UMM succeeds as an integrated system, while stage-wise analyses can be used for further diagnosis. **(3) Annotation-free scalability.** SGU can be instantiated on existing image-grounded reasoning benchmarks, such as MMBench Liu et al. (2024), by reusing their original questions and ground-truth answers, requiring no new annotations.

Extensive experiments reveal a clear gap between isolated benchmark performance and integrated multimodal behavior in current UMMs. Despite strong results on individual tasks, many models suffer substantial performance degradation under the SGU loop, exposing limitations that are not captured by separate evaluations of understanding or generation alone. Stage-wise replacement experiments further help analyze which parts of the loop contribute to the degradation, while prompt, caption-length, and cross-model answering comparisons examine how the score depends on the evaluation process. These findings show that SGU provides both a complementary system-level score and diagnostic insights for evaluating unified multimodal models.

Our contributions are summarized as follows:

- We identify the need for a complementary system-level evaluation framework for UMMs, beyond separate generation and understanding metrics.
- We propose SGU, a semantic closed-loop evaluation framework tailored for UMMs, providing an annotation-free and model-internal system-level assessment of integrated understanding-and-generation behavior.
- We evaluate six representative UMMs on four benchmarks and use stage replacements and controlled input variations to diagnose their closed-loop performance.

2 Related Work

2.1 Unified Multimodal Models

Multimodal intelligence has rapidly evolved from task-specific systems toward Unified Multimodal Models (UMMs), which aim to integrate visual understanding and generation within a single model Li et al. (2025); Liang et al. (2025); Wei et al. (2026). Existing UMMs can be broadly categorized into several representative paradigms. The first line of work adopts autoregressive modeling with unified or discrete visual tokenization, allowing text and image tokens to be processed within a shared sequence modeling framework. Representative examples include Chameleon Team (2025), Emu3 Wang et al. (2024), and Janus-Pro Chen et al. (2025b). Show-o2 Xie et al. (2025b) combines autoregressive text prediction with flow matching for visual generation. The second paradigm integrates diffusion or flow-based generation modules with multimodal reasoning backbones, aiming to improve image generation quality while preserving understanding capabilities Lipman et al. (2023); Zhou et al. (2024); Liang et al. (2025). Representative models such as BLIP3-o Chen et al. (2025a), BAGEL Deng et al. (2025), UniWorld Lin et al. (2025), and OmniGen2 Wu et al. (2025b) leverage large-scale multimodal training or interleaved data to support both visual reasoning and generation. Another line of work maintains a strong reasoning backbone while connecting it to image generation modules through learnable interfaces or routing mechanisms Pan et al. (2025). This design can preserve language and vision reasoning ability while enabling generation through specialized components. In parallel, proprietary systems such as GPT-Image OpenAI (2025b) and Gemini 2.5 Flash Image Google DeepMind (2025) further demonstrate the rapid progress of unified understanding-generation systems.

2.2 Benchmarks for Unified Multimodal Models

Evaluation of UMMs remains challenging because current methodologies typically assess multimodal understanding and generation separately. Understanding capabilities are generally evaluated through question-answering (QA) benchmarks, such as MME Fu et al. (2025) and MMMU Yue et al. (2024; 2025). Generation benchmarks such as GenEval assess whether generated images satisfy object, count, attribute, and spatial requirements in a text prompt Ghosh et al. (2023). Recent advancements, such as GIRBench Li et al. (2025), have extended this scope to visual editing, while VQAScore Lin et al. (2024) uses a VQA model's answer likelihood to assess image–text alignment. For unified evaluation, UmniBench Liu et al. (2025) assesses generation and editing performance by utilizing the model's intrinsic understanding.

These benchmarks provide valuable component-level and joint understanding–generation insights. SGU complements them by testing the original VQA task after a question-blind caption–reconstruction loop. Its score reuses the benchmark answers and the tested model's own reader. This gives a common task endpoint for examining how captioning, generation, and answering interact, alongside the component-level measures above.

3 Method

3.1 Preliminaries

Unified Multimodal Model (UMM). Formally, we define a UMM $\mathcal{M}$ as a probabilistic model designed to process and generate visual and textual modalities within a unified model. While current UMMs often share a common parameter space, their functional capabilities can be viewed through

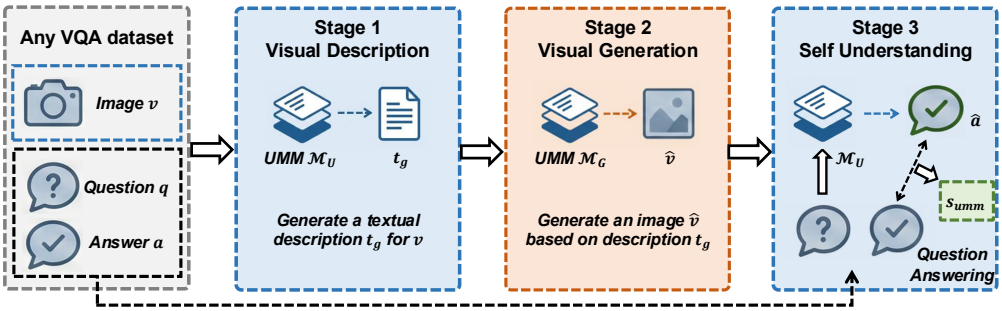

Figure 2: Overview of the Self-Generative-Understanding (SGU) protocol. Given a VQA triplet (v, q, a), the UMM first generates a textual description t_g from the original image v, reconstructs a visual context $\hat{v}$ from t_g, and then answers the original question q based on $\hat{v}$ to obtain the SGU score s_{umm}. Each stage is executed in an isolated stateless session, with no memory or intermediate features carried across stages.

conditional projections targeting different output modalities Wang (2025); Wei et al. (2025). We denote the visual generation function as $\mathcal{M}_G$ and the understanding function as $\mathcal{M}_U$.

UMM evaluation. The core challenge in UMM evaluation is to define a scoring function $\mathcal{S}$ that can assess the model as an integrated system. Given a multimodal dataset $\mathcal{D} = \{(v, t)\}$, where v denotes visual input and t denotes textual input, a system-level score can be formalized as

$$s_{\text{umm}} = \mathcal{S}(\mathcal{M}, \mathcal{D}). \tag{1}$$

Such a score should satisfy two desiderata. (1) *Unified assessment*. The score should provide a scalar signal for comparing different UMMs at the system level, complementing component-wise evaluations. (2) *Integrated sensitivity*. The score should reflect the combined effect of understanding and generation, such that weakness in either capability, or in their interaction within the model, can affect the final performance.

Current evaluation protocols. Existing evaluations commonly assess generation and understanding using separate metrics. Formally, given a generation benchmark $\mathcal{D}_G$ and an understanding benchmark $\mathcal{D}_U$, current methods can be viewed as producing a tuple of heterogeneous scores:

$$\mathbf{s}_{\text{comp}} = (s_g, s_u), \quad s_g = \mathcal{S}_G(\mathcal{M}_G, \mathcal{D}_G), \quad s_u = \mathcal{S}_U(\mathcal{M}_U, \mathcal{D}_U), \tag{2}$$

where $\mathcal{S}_G$ denotes generation-oriented metrics such as FID Heusel et al. (2017) and CLIPScore Hessel et al. (2022), and $\mathcal{S}_U$ denotes understanding-oriented metrics such as VQA accuracy Chen et al. (2024); Liu et al. (2024).

These component-wise evaluations are valuable for diagnosing specific capabilities. At the same time, their heterogeneous nature makes direct aggregation non-trivial. This motivates a complementary system-level perspective for understanding how a UMM behaves when understanding and generation are jointly involved.

3.2 Our Evaluation: Self-Generative-Understanding

We propose Self-Generative-Understanding (SGU), an annotation-free semantic closed-loop framework for system-level evaluation of UMMs. SGU is designed to evaluate whether a UMM can perform understanding and generation in an integrated manner by asking the model to operate on its own intermediate outputs. Given a VQA sample $(v, q, a) \in \mathcal{D}$, SGU enforces a three-stage process: (1) *Visual description*. The understanding function $\mathcal{M}_U$ perceives the original visual input v and generates a textual description t_g. (2) *Visual generation*. Based solely on the self-generated description t_g, the generation function $\mathcal{M}_G$ synthesizes a proxy visual context $\hat{v}$. (3) *Self understanding*. The model then uses $\mathcal{M}_U$ to resolve the original textual task on the synthesized image $\hat{v}$ and produces the final answer $\hat{a}$. The process can be formulated as

$$t_g = \mathcal{M}_U(v), \quad \hat{v} = \mathcal{M}_G(t_g; \xi), \quad \hat{a} = \mathcal{M}_U(\hat{v}, q). \tag{3}$$

3.3 SGU SCORE

To instantiate SGU, we adopt VQA-style settings, where each question is grounded in the original image and paired with a ground-truth answer. A data sample is formalized as a triplet (v, q, a), consisting of an image v, a question q, and a ground-truth answer a. Based on Eq. (3), the SGU score is calculated as

$$s_{\text{umm}} = \mathbb{E}_{(v,q,a)\sim\mathcal{D},\,\xi}\left[\mathbb{I}\left(\text{Match}\left(\mathcal{M}_U(\mathcal{M}_G(\mathcal{M}_U(v);\xi),q),a\right)\right)\right]. \tag{4}$$

The expectation averages over benchmark questions and generation randomness ξ under a fixed inference configuration. Here $\mathbb{I}(\cdot)$ is the indicator function, and Match follows each benchmark's own evaluation protocol for the downstream task, applied identically to the direct and reconstructed-image paths.

What does SGU measure? SGU measures success on the original task after the model has described and reconstructed its input. The shared task and reference answer make the direct and loop paths comparable, while their difference exposes the cost of operating through the intermediate representations. Caption-only QA and controlled stage replacements help locate changes along this path (Section 4), while the intermediate descriptions and images make the information passed between capabilities available for inspection.

Relative SGU score. Since s_{umm} is influenced by the model's base understanding ability, we introduce a relative SGU score normalized by the base understanding score s_{base}, which measures VQA performance on the original images and serves as a model-specific direct-VQA reference:

$$s_{\text{base}} = \mathbb{E}_{(v,q,a)\sim\mathcal{D}}\left[\mathbb{I}\big(\text{Match}\left(\mathcal{M}_U(v,q),a\right)\big)\right], \quad s_{\text{umm,r}} = \frac{s_{\text{umm}}}{s_{\text{base}}}. \tag{5}$$

For $s_{\text{base}} > 0$, the relative score expresses loop performance as a fraction of the model's direct VQA accuracy. Paired outcomes further describe how individual answers change along the loop (Appendix C.2). In the default protocol, captioning and generation use only the image and its description, respectively; the question enters at the final VQA stage.

Question-guided diagnostic. A task-conditioned variant supplies the question to the description stage, using an instruction p_{cap} to describe relevant visual evidence: $t_g^{(q)} = \mathcal{M}_U(v, q; p_{\text{cap}})$. Generation still receives only the description; captioning receives neither answer options nor the reference answer. This variant provides task guidance while preserving the description–generation–answering structure. Its interpretation alongside caption-only QA is discussed in Section 4.

4 EXPERIMENTS

4.1 IMPLEMENTATION

Datasets. We evaluate SGU on a diverse set of representative multimodal benchmarks covering general reasoning, perception, mathematical reasoning, and text-centric visual understanding. These datasets allow us to examine the system-level behavior of UMMs across different VQA-style tasks.

- **MMStar** Chen et al. (2024), a comprehensive benchmark designed to assess general multimodal reasoning across diverse visual and textual scenarios.
- **MMBench** Liu et al. (2024), which focuses on evaluating core vision–language understanding abilities through carefully curated multiple-choice questions.
- **MathVista** Lu et al. (2024), targeting mathematical and logical reasoning grounded in visual contexts.
- **OCR-VQA** Mishra et al. (2019), emphasizing text recognition and semantic understanding in visually rich, text-heavy images.

UMMs. We evaluate six representative unified multimodal models across diverse architectures to analyze their system-level behavior under the SGU protocol. The models are Janus-Pro-7B Chen et al. (2025b), BAGEL-7B Deng et al. (2025), UniWorld-V1 Lin et al. (2025), Show-o2-7B Xie et al. (2025b), OmniGen2 Wu et al. (2025b), and Ovis-U1-3B Wang et al. (2025). Architecture details are provided in Appendix B.1.

Table 1: Results of various unified multimodal models on diverse VQA benchmarks under the SGU protocol. For each dataset, we report direct VQA accuracy on the original image (s_{base}) and SGU score (s_{umm}). The reference row gives the best direct score for each task and their arithmetic mean. Bold and underline indicate the highest and second-highest SGU scores, with ties retained.

Model	MMStar		MMBench		MathVista		OCR-VQA		Avg	
	Original s_{base}	SGU s_{umm}	Original s_{base}	SGU s_{umm}	Original s_{base}	SGU s_{umm}	Original s_{base}	SGU s_{umm}	Original s_{base}	SGU s_{umm}
BEST DIRECT	66.47		86.13		70.60		81.33		76.13	
UniWorld-V1	60.67	38.33	86.13	71.97	67.30	37.40	81.33	28.67	73.86	44.09
Janus-Pro-7B	47.80	39.27	76.86	68.20	41.30	32.50	69.27	37.45	58.81	44.36
Show-o2-7B	55.13	<u>43.00</u>	83.92	73.95	50.70	38.90	63.43	30.34	63.30	46.55
Ovis-U1	60.60	**43.07**	83.71	<u>75.10</u>	68.90	**41.80**	79.00	34.71	73.05	48.67
BAGEL-7B	66.47	42.67	85.54	**76.38**	70.60	39.50	74.55	<u>53.35</u>	74.29	<u>52.98</u>
OmniGen2	54.93	**43.07**	82.84	73.66	63.50	<u>40.40</u>	79.39	**56.59**	70.17	**53.43**

Evaluation setup. For each sample (v, q, a), the UMM describes v, generates an image from that description, and answers q using the generated image. Captioning and VQA use deterministic greedy decoding; image generation uses three random seeds, and we average the downstream VQA scores across the three runs to reduce generation variability. Each benchmark uses its own downstream evaluation protocol. All stages run as independent inference calls, passing only the description or reconstructed image to the next stage. The tested UMM performs the complete loop, and the score reuses the benchmark's existing reference answers.

4.2 RESULTS ANALYSIS

Q1. ***Does SGU provide a system-level view beyond isolated evaluations?***

A1. Table 1 shows a consistent drop from direct VQA accuracy on the original image (s_{base}) to the SGU score (s_{umm}). The loop asks a UMM to make the visual information it understands usable through its own generation process. Models with similar direct scores can differ substantially on this task: on OCR-VQA, UniWorld-V1 achieves 81.33 direct accuracy and 28.67 SGU, whereas OmniGen2 achieves 79.39 and 56.59. A 1.94-point advantage on the original image becomes a 27.92-point disadvantage after reconstruction. SGU therefore reveals differences in integrated performance that are difficult to anticipate from direct understanding accuracy.

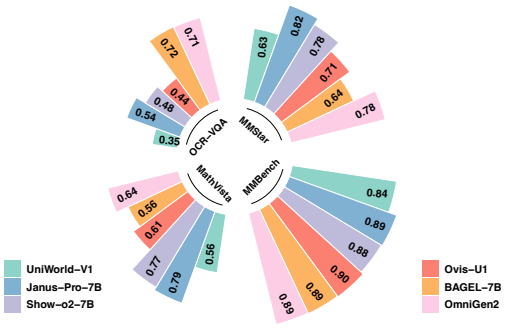

Figure 3: Relative SGU scores $s_{\text{umm,r}}$.

The relative score in Figure 3 contextualizes loop performance against each model's direct VQA reference. MathVista and OCR-VQA generally show larger losses than MMBench: preserving an exact word or mathematical relation places a more specific demand on the description and its realization than retaining broad scene information. SGU tests how successfully the model carries the information needed for the original task through this complete pathway.

Figure 5 compares SGU with an equally weighted aggregate of direct VQA, CLIP-T, and CLIP-I. BAGEL leads the component aggregate, while OmniGen2 leads mean SGU. Image–text and image–image similarities describe intermediate alignment; the final question additionally tests whether the aligned content is usable for the task. These rankings expose different strengths in component quality and integrated task completion.

Table 2: Stage replacement effects (Δs_{umm}, percentage points). One stage is replaced while the rest of each UMM's loop remains unchanged. MV: MathVista; OCR: OCR-VQA.

	Understand repl.		Generation repl.	
Model	MV	OCR	MV	OCR
UniWorld-V1	-4.16	-2.52	13.64	38.89
Janus-Pro-7B	1.88	-3.03	3.54	21.21
Show-o2-7B	-1.74	1.01	0.07	29.30
Ovis-U1-3B	1.51	2.02	6.57	34.85
BAGEL-7B	-1.60	-0.50	5.05	13.14
OmniGen2	3.93	-1.51	11.61	12.12

Table 3: Prompt sensitivity of OmniGen2 on MMBench. Each column varies only its named stage; all other prompts remain at the default. Templates are in Appendix B.3.

	SGU score (%)	
Variant	Captioning	Generation
Default	73.66	
1 (U1/G1)	73.48	75.51
2 (U2/G2)	73.99	73.48
3 (U3/G3)	73.81	73.92
4 (U4/G4)	73.57	73.61

Q2. ***How does SGU reflect and diagnose the joint effects of understanding and generation?***

A2. We replace the description model with Qwen3-VL-8B Bai et al. (2025), or the generator with Qwen-Image-2512 Wu et al. (2025a), while keeping the rest of the loop unchanged. Table 2 focuses on MathVista and OCR-VQA, where the loop gaps are most pronounced. On OCR-VQA, generator replacement improves accuracy by 12.12–38.89 points, whereas captioner replacement changes it by -3.03 to $+2.02$ points. Generation gains on MathVista range from 0.07 to 13.64 points. The larger generation gains indicate substantial room to improve how existing descriptions are rendered. Captioner replacement still relies on the native generator to realize the new description.

Caption-only QA makes these model-specific profiles visible. On OCR-VQA, UniWorld's accuracy changes from 81.33 on the original image to 63.30 on its description and 28.67 after reconstruction; OmniGen2's corresponding scores are 79.39, 60.05, and 56.59. The two caption paths differ by only 3.25 points, but the loop paths differ by 27.92 points in the opposite direction. For UniWorld, much of the information usable from its description is difficult to recover after native generation. OmniGen2 carries substantially more of its caption-level performance through reconstruction. Thus, strong upstream understanding creates an opportunity for a successful loop, and the generator strongly affects how effectively that opportunity is realized. Figure 6 shows that the stronger generator narrows the OCR-VQA gaps while leaving different amounts of headroom across models.

To examine what enables generator gains, we group UniWorld descriptions by whether the same external text-only probe can answer the question from them before replacement (Appendix G.3). The probe-correct and probe-incorrect groups gain $24.20/2.56$ points on MathVista and $53.48/7.13$ points on OCR-VQA. Although the probe-correct groups contain $51.20\%/68.52\%$ of the questions, they contribute $90.84\%/94.23\%$ of the total net improvement. Their rescue rates also remain higher after accounting for baseline errors. Most of the generator's benefit is therefore realized when the native description already supplies usable task information. The smaller gains in the other group are consistent with generation completing partially specified visual details that help the final model answer the question.

Crossed replacements further show how the two stages interact (Appendix G.2). Changing both components produces gains that differ from the sum of the individual changes by -2.21 to $+1.40$ points. Replacing the captioner changes the content and form of the generator's input, so the benefit of either component depends on the composition. The magnitude of a replacement gain therefore describes the improvement available in a particular composition. Their roles are sequential: understanding selects and expresses visual information, generation realizes it, and downstream VQA tests its use in completing the original task.

Question-guided captioning tests whether directing attention to the requested evidence substantially changes this picture (Appendix G.4). The captioner receives the question and describes the relevant visible details; generation still receives only that description. Across all four benchmarks, at the 256-token cap, guidance improves Caption-QA by 0.73–2.06 points and changes SGU by -0.40 to $+1.90$ points relative to the question-blind allocation control. Doubling the question-blind cap changes SGU by -0.40 to $+0.90$ points. Substantial loop gaps remain even when the model knows which evidence to describe and can produce a longer caption. These results point to the quality of

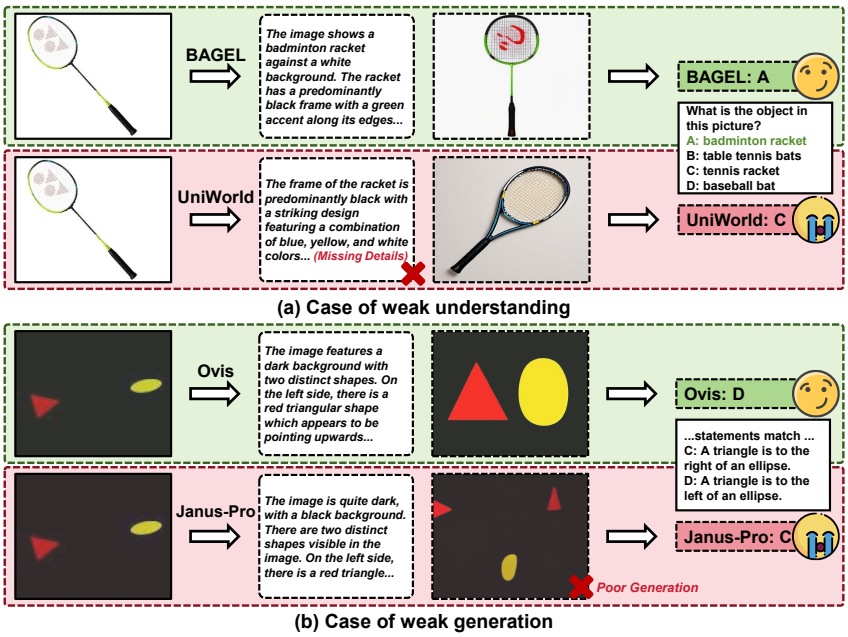

Figure 4: Qualitative case studies under the SGU protocol. The examples illustrate different loop outcomes: one model successfully performs both understanding and generation, leading to the correct answer, while another produces an incorrect answer when visual information is incompletely extracted (Case a) or the reconstructed image deviates from the intended content (Case b).

Table 4: Image-swap check on MMBench, pairing models with similar direct VQA accuracy s_{base}. Arrows denote image producer $\rightarrow$ answering model. For each target model A, $s_{\text{umm}} = \text{Acc}(A \rightarrow A)$ evaluates A on images generated by itself, while $s_{\text{cross}} = \text{Acc}(A \rightarrow B)$ evaluates the paired model B on the same images. We report $s_{\Delta} = s_{\text{cross}} - s_{\text{umm}}$.

Pair	Target	s_{Δ}	s_{base}	s_{umm}	s_{cross}
Ovis $\leftrightarrow$ UniWorld	Ovis	-0.82	83.71	75.10	74.28
	UniWorld	+0.34	86.13	71.97	72.31
Show-o2 $\leftrightarrow$ BAGEL	Show-o2	-2.77	83.92	73.95	71.18
	BAGEL	+3.04	85.54	76.38	79.42
OmniGen2 $\leftrightarrow$ Janus	OmniGen2	-3.06	82.84	73.66	70.60
	Janus	+5.83	76.86	68.20	74.03

the extracted and reconstructed content as the central challenge in these settings. VQA focuses on concrete visual facts and relations; selecting, expressing, and reproducing those details is precisely the integrated capability that SGU exercises. Figure 4 illustrates both stages: Case (a) hinges on object-specific details in the description, whereas Case (b) hinges on preserving spatial relations during generation.

Q3. ***Is SGU sensitive to understanding and generation prompts?***

A3. Table 3 compares four captioning and four generation variants, changing one instruction at a time with all other settings unchanged. Complete templates are in Appendix B.3. Scores range from 73.48 to 75.51, versus 73.66 for the default. Captioning variants stay within 0.33 points of the default, and three generation variants stay within 0.26 points. Directly using the caption as the generation input (G1) gives the largest change, +1.85 points, suggesting that the instruction wrapped around a description can affect its realization. The modest effects of the other variants support a simple, consistent prompt choice and place these wording effects in context alongside the larger component-replacement effects.

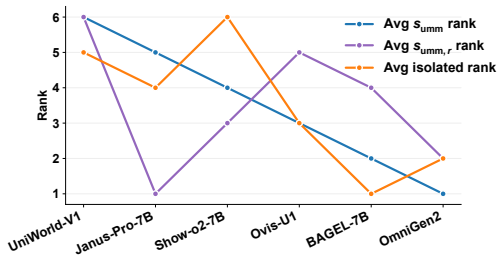

Figure 5: Model rankings by SGU, relative SGU, and the component average (Appendix C.1).

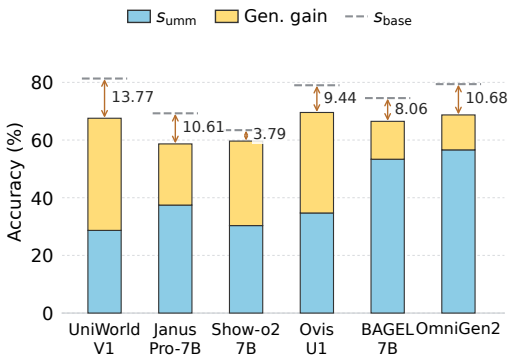

Figure 6: Generator replacement on OCR-VQA. Blue: SGU; yellow: gain; dashed lines: direct VQA. Labels give the remaining gap in points.

Q4. ***Does SGU reflect task-relevant visual information?***

A4. Image controls test how the reconstructed content contributes to the answer. Across the 24 model–dataset combinations, SGU exceeds question-only and mismatched-image accuracy by 13.54–44.42 and 14.55–46.07 points (Appendix G.1). For example, OmniGen2 reaches 56.59 on OCR-VQA, compared with 12.17 without an image and 10.52 with a mismatched reconstruction. Keeping the question and answering model the same while changing image correspondence sharply reduces performance, linking the SGU outcome to the reconstructed visual content. The image-swap comparison asks whether that content is preferentially readable by its producer. Cross-minus-self changes range from -3.06 to $+5.83$ points on MMBench (Table 4); both positive and negative changes occur. The reconstructions can therefore support other UMMs as well, without a consistent advantage for self-reading.

Independent human judgments connect these model outcomes to the visual evidence in the reconstructed images (Appendix G.6). SGU has the highest association with the judged evidence among the compared signals, reaching 0.551 on OCR-VQA and 0.314 on MathVista. Adding SGU to a predictor based on Direct, Caption-QA, CLIP-T, and CLIP-I reduces OCR-VQA test Brier score from 0.207 to 0.188 and log-loss from 0.598 to 0.557. MathVista shows a smaller, mixed change: Brier score decreases from 0.221 to 0.219, while log-loss increases from 0.634 to 0.636. This contrast is consistent with the tasks: readable visual text directly constrains OCR-VQA answers, whereas a reconstructed mathematical diagram still requires further reasoning. SGU evaluates the completion of that entire task, and its added information is clearest where the preserved visual evidence closely determines the answer. Together, the input controls, cross-model answering, and independent judgments support the use of SGU as a task-grounded assessment of the integrated loop.

5 Conclusion

We presented Self-Generative-Understanding (SGU), a semantic closed-loop framework for evaluating the integrated capabilities of unified multimodal models. By asking a UMM to describe an image, reconstruct its content, and answer the original visual question, SGU evaluates understanding and generation through a shared task outcome. The framework reuses existing benchmarks and reference answers without requiring new annotations. Across six models and four benchmarks, SGU reveals differences in integrated performance that separate capability scores leave unresolved. The analyses show how extracting useful information and realizing it through generation jointly support task completion. This makes SGU a practical tool for comparing UMMs as complete systems and studying how their capabilities work together.

Limitations and future work. The present evaluation covers image-based VQA and a single description–reconstruction cycle. Extending SGU to video, interactive tasks, and richer intermediate representations would broaden its scope. Its explicit intermediate outputs also offer a basis for training models to preserve task-relevant information across modalities.

AI USE STATEMENT

In this work, we used generative AI tools solely to improve the language and readability of the manuscript and to assist with implementing our evaluation protocol in code. All AI-assisted text and code were carefully reviewed by the authors to ensure correctness and preserve the intended technical meaning. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the assistance of generative AI.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset, 2025a. URL https://arxiv.org/abs/2505.09568.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024. URL https://arxiv.org/abs/2403.20330.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025b. URL https://arxiv.org/abs/2501.17811.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025. URL https://arxiv.org/abs/2505.14683.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, Katsushi Ikeuchi, Hoi Vo, Li Fei-Fei, and Jianfeng Gao. Agent ai: Surveying the horizons of multimodal interaction, 2024. URL https://arxiv.org/abs/2401.03568.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2025. URL https://arxiv.org/abs/2306.13394.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment, 2023. URL https://arxiv.org/abs/2310.11513.

Google DeepMind, 2025. URL https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. URL https://arxiv.org/abs/2104.08718.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Hongxiang Li, Yaowei Li, Bin Lin, Yuwei Niu, Yuhang Yang, Xiaoshuang Huang, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Long Chen. Gir-bench: Versatile benchmark for generating images with reasoning, 2025. URL https://arxiv.org/abs/2510.11026.

Yongyuan Liang, Wei Chow, Feng Li, Ziqiao Ma, Xiyao Wang, Jiageng Mao, Jiuhai Chen, Jiatao Gu, Yue Wang, and Furong Huang. Rover: Benchmarking reciprocal cross-modal reasoning for omnimodal generation, 2025. URL https://arxiv.org/abs/2511.01163.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, Yatian Pang, and Li Yuan. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation, 2025. URL https://arxiv.org/abs/2506.03147.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Kai Liu, Leyang Chen, Wenbo Li, Zhikai Chen, Zhixin Wang, Renjing Pei, Linghe Kong, and Yulun Zhang. Umnibench: Unified understand and generation model oriented omni-dimensional benchmark, 2025. URL https://arxiv.org/abs/2512.17196.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024. URL https://arxiv.org/abs/2307.06281.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024. URL https://arxiv.org/abs/2310.02255.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 947–952, 2019. doi: 10.1109/ICDAR.2019.00156.

OpenAI, 2025b. URL https://openai.com/index/image-generation-api/.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with metaqueries, 2025. URL https://arxiv.org/abs/2504.06256.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2025. URL https://arxiv.org/abs/2405.09818.

Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, Yang Li, and Qing-Guo Chen. Ovis-u1 technical report, 2025. URL https://arxiv.org/abs/2506.23044.

Maijunxian Wang. From understanding the world to intervening in it: A unified multi-scale framework for embodied cognition, 2025. URL https://arxiv.org/abs/2503.00727.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024. URL https://arxiv.org/abs/2409.18869.

Hongyang Wei, Baixin Xu, Hongbo Liu, Cyrus Wu, Jie Liu, Yi Peng, Peiyu Wang, Zexiang Liu, Jingwen He, Yidan Xietian, Chuanxin Tang, Zidong Wang, Yichen Wei, Liang Hu, Boyi Jiang, William Li, Ying He, Yang Liu, Xuchen Song, Eric Li, and Yahui Zhou. Skywork unipic 2.0: Building kontext model with online rl for unified multimodal model, 2025. URL https://arxiv.org/abs/2509.04548.

Jingxuan Wei, Caijun Jia, Xi Bai, Xinglong Xu, Siyuan Li, Linzhuang Sun, Bihui Yu, Conghui He, Lijun Wu, and Cheng Tan. Ggbench: A geometric generative reasoning benchmark for unified multimodal models, 2026. URL https://arxiv.org/abs/2511.11134.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Exploration to advanced multimodal generation, 2025b. URL https://arxiv.org/abs/2506.18871.

Ji Xie, Trevor Darrell, Luke Zettlemoyer, and XuDong Wang. Reconstruction alignment improves unified multimodal models, 2025a. URL https://arxiv.org/abs/2509.07295.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models, 2025b. URL https://arxiv.org/abs/2506.15564.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. URL https://arxiv.org/abs/2311.16502.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model, 2024. URL https://arxiv.org/abs/2408.11039.

A Datasets and Evaluation Benchmarks

We evaluate SGU on four representative multimodal benchmarks covering general perception, visual reasoning, mathematical reasoning, and text-centric visual understanding. These datasets provide diverse VQA-style tasks for examining the system-level behavior of UMMs under the SGU protocol.

A.1 Dataset Descriptions

MMStar Chen et al. (2024). MMStar is a high-quality vision-language benchmark curated to emphasize *vision-dependent* questions. It is designed to reduce the influence of language priors by filtering samples that can be answered reliably without the image, and covers diverse perception-centric skills such as fine-grained recognition, spatial relations, and multi-object reasoning.

MMBench Liu et al. (2024). MMBench is a broad multimodal evaluation suite targeting general visual perception and reasoning. Its questions span recognition and attributes, spatial understanding, commonsense and social reasoning, and multi-step visual reasoning, providing a comprehensive testbed for overall VQA competence under a unified format.

MathVista Lu et al. (2024). MathVista focuses on visually grounded mathematical reasoning, where solving requires jointly interpreting visual evidence, such as geometry diagrams, plots, tables, and charts, and performing mathematical inference. This benchmark is sensitive to errors in visual reconstruction, since small distortions in the reconstructed image may affect the evidence needed for correct reasoning.

OCR-VQA Mishra et al. (2019). OCR-VQA pairs book-cover images with questions about attributes such as the title and author. Answering these questions requires reading text and associating it with the requested book attribute, making the benchmark useful for examining text extraction and reconstruction.

A.2 Summary of Data Usage

Table 5 summarizes the datasets used in SGU and the primary evaluation focus of each.

Table 5: Dataset and primary evaluation focus used in this work.

Dataset	Type	Primary focus
MMStar	General	Vision-dependent reasoning; perception
MMBench	General	Visual perception; general reasoning
MathVista	Math & charts	Geometric structure; mathematical reasoning
OCR-VQA	Text-in-image	OCR; visual text understanding

B Implementation Details

This section provides the inference configurations and prompts used to execute the SGU protocol. All evaluated UMMs are used without training or fine-tuning. We run the SGU pipeline on NVIDIA A800 GPUs with 80GB memory; the exact runtime varies across models and datasets because each sample requires captioning, image generation, and final VQA inference.

B.1 Evaluated Model Descriptions

- **Janus-Pro-7B** Chen et al. (2025b) adopts an autoregressive framework that decouples visual encoding to mitigate conflicts between multimodal understanding and generation.
- **BAGEL-7B** Deng et al. (2025) is a decoder-only foundation model trained on massive interleaved image-text data, subsequently exhibiting strong emergent multimodal reasoning capabilities across benchmarks.
- **UniWorld-V1** Lin et al. (2025) uses a high-resolution semantic encoder to connect visual understanding with image generation and editing.

- **Show-o2-7B** Xie et al. (2025b) combines autoregressive text modeling with flow matching for visual generation, using fused semantic and low-level visual features.
- **OmniGen2** Wu et al. (2025b) supports arbitrary multimodal input-output sequences with a focus on generalized instruction following in complex scenarios.
- **Ovis-U1-3B** Wang et al. (2025) is a lightweight unified model that enables evaluation at a smaller parameter count with lower inference cost.

B.2 Model Inference Configurations

To ensure fair comparison across UMMs, we use a consistent inference protocol whenever possible, while following each model's native generation interface.

- **Understanding phase ($I \rightarrow T$).** For the image-to-text stage, we use greedy decoding by setting `do_sample` to **False**. The output length is capped with `max_new_tokens` set to **256**. Token counts follow each model's own tokenizer; the length comparison additionally uses a 512-token cap.
- **Generation phase ($T \rightarrow I$).** Generation configurations differ across model architectures. We use the three random seeds **42**, **2026**, and **123**. When applicable, we use a Classifier-Free Guidance (CFG) scale of **5.0**. Specific settings follow each model's native architecture:
  - *Autoregressive architecture (Janus-Pro-7B):* The generation process uses **576 image tokens**, decoded with the model's native image decoder, with multinomial sampling and a temperature of 1.0.
  - *Diffusion and flow-based architectures (BAGEL, OmniGen2, Ovis, UniWorld, Show-o2):* We use **30 inference steps** and generate images at $\mathbf{512} \times \mathbf{512}$ resolution. When supported by the model, the CFG scale is set to **5.0**.
- **VQA phase.** For the final question-answering stage, we use greedy decoding by setting `do_sample` to **False**. The output length is restricted with `max_new_tokens` set to **64**, since VQA answers are typically concise.

B.3 Prompt Engineering

The prompts used in each phase of the SGU cycle are listed below. The placeholders `{description}`, `{Question}`, and `{Options}` denote the generated caption, the input question, and the candidate answer options, respectively.

Standard protocol. For the main experiments, we use the following default prompts:

- **Phase 1: Understanding ($I \rightarrow T$)**

  ```
  <image> Describe this image in detail.
  ```

- **Phase 2: Generation ($T \rightarrow I$)**

  ```
  Generate an image based on the following description:
  {description}.
  ```

- **Phase 3: Evaluation (VQA)**
  - *Multiple-choice questions:*

    ```
    {Question} {Options} Answer with the option letter
    directly.
    ```

  - *Open-form questions:*

    ```
    {Question} Answer the question directly.
    ```

Prompt variants for sensitivity analysis. Table 3 changes one stage at a time. U1–U4 replace only the image-to-text instruction; G1–G4 replace only the text-to-image instruction. The other stages retain their default prompts, and all captioning variants remain question-blind.

U1 *Generate a detailed description of the visual content in this image.*

U2 *Analyze the image and provide a thorough description capturing all key elements.*

U3 *Describe the objects, attributes, and spatial relationships visible in this image.*

U4 *Provide a detailed visual account of the image, including its main elements and their arrangement.*

G1 *{description}*

G2 *Generate a high-quality, detailed image of {description}.*

G3 *Create an image depicting the following description: {description}.*

G4 *Render the scene described here: {description}.*

B.4 Stateless Pipeline Execution

SGU is executed in a *stateless* manner across stages. The captioning, generation, and final VQA steps are run as independent inference calls. Only the explicit intermediate artifact, i.e., the generated caption or reconstructed image, is passed to the next stage; no hidden states, KV-cache, or conversation history are carried over.

Stage isolation. For each stage, we create a fresh model invocation with a clean input payload. We do not reuse the same chat session or inference context across stages. In question-blind runs, we cache a deterministic caption per source image and a reconstruction per image and generation seed, and reuse them for its associated questions. Question-guided runs and their allocation control instead generate separately for each question (Appendix G.4).

Inputs to each stage in the default protocol. (i) **Captioning** takes the original image v and the captioning prompt, and outputs a description t_g. (ii) **Generation** takes only the text prompt formed from t_g and produces a reconstructed image $\hat{v}$. (iii) **VQA scoring** takes only $\hat{v}$ and the question q, and returns the answer used to compute s_{umm}; the original image v and the intermediate caption t_g are not provided to the final VQA stage.

C Additional Metrics and Theoretical Details

C.1 Auxiliary Metrics for Isolated Assessment

The following definitions cover FID, discussed as a conventional generation metric, and the CLIP-based similarities used in our auxiliary results.

FID (Fréchet Inception Distance) Heusel et al. (2017). FID measures the distributional distance between two image sets by comparing their feature statistics in a pretrained Inception network. Given feature activations with empirical means and covariances (μ_r, Σ_r) for real images and (μ_g, Σ_g) for generated images, FID is defined as

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\Big(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\Big), \tag{6}$$

where $\text{Tr}(\cdot)$ denotes the matrix trace.

CLIP-based similarities Hessel et al. (2022). We use CLIP embeddings to quantify (i) visual similarity between the original and reconstructed images, and (ii) text–image similarity between the model-generated caption and the original image. Let $f_I(\cdot)$ and $f_T(\cdot)$ denote the CLIP image and text encoders, and let $\langle \cdot, \cdot \rangle$ denote cosine similarity. We compute

$$\text{CLIP-I} = \left\langle \frac{f_I(I)}{\|f_I(I)\|}, \frac{f_I(I')}{\|f_I(I')\|} \right\rangle, \tag{7}$$

$$\text{CLIP-T} = \left\langle \frac{f_I(I)}{\|f_I(I)\|}, \frac{f_T(C)}{\|f_T(C)\|} \right\rangle, \tag{8}$$

where I is the original image, I' is the reconstructed image, and C is the intermediate caption produced at the understanding stage. We report raw cosine similarities.

Component average and model rankings. For model m, the component average in Figure 5 is

$$a_m = \frac{1}{12} \sum_{d \in \mathcal{B}} \left(s_{\text{base},m,d} + \text{CLIP-T}_{m,d} + \text{CLIP-I}_{m,d} \right), \tag{9}$$

where $\mathcal{B}$ contains the four benchmarks and direct accuracy is expressed as a fraction. The three components and four benchmarks receive equal weight. The other curves rank the four-benchmark means of SGU accuracy and of the per-dataset SGU/direct ratio, respectively. The latter averages ratios before ranking, rather than dividing the two mean accuracies.

C.2 Direct-VQA Reference for SGU

We use direct VQA accuracy on the original image as a model-specific direct-VQA reference for interpreting the SGU score:

$$s_{\text{base}} = \mathbb{E}_{(v,q,a)\sim\mathcal{D}} \left[\mathbb{I} \left(\text{Match} \left(\mathcal{M}_U(v,q), a \right) \right) \right]. \tag{10}$$

Compared with direct VQA, SGU introduces an additional understanding–generation loop before the final answer is produced. The model must first convert the original image into an intermediate textual representation, reconstruct a visual context from it, and then answer the original question using the reconstructed image. Direct VQA provides the original-image reference for this comparison. A reconstructed input can also make a previously failed question easier to answer, so the paired outcomes describe both preserved and newly correct answers.

Correctness sets. For a fixed generation seed ξ, we define the sets of correctly answered samples under direct VQA and SGU as

$$\mathcal{G}_{\text{base}} \coloneqq \left\{ (v,q,a) \in \mathcal{D} \;\middle|\; \text{Match} \left(\mathcal{M}_U(v,q), a \right) = 1 \right\}, \tag{11}$$

$$\mathcal{G}_{\text{SGU}} \coloneqq \left\{ (v,q,a) \in \mathcal{D} \;\middle|\; \text{Match} \left(\mathcal{M}_U(\mathcal{M}_G(\mathcal{M}_U(v);\xi),q), a \right) = 1 \right\}. \tag{12}$$

To describe the paired outcomes, let p_{bl} denote the fraction with direct correctness b and loop correctness l, on identical questions and a declared generation run. Then

$$s_{\text{base}} = p_{11} + p_{10}, \qquad s_{\text{umm}} = p_{11} + p_{01}, \qquad s_{\text{base}} - s_{\text{umm}} = p_{10} - p_{01}. \tag{13}$$

For $s_{\text{base}} > 0$, direct-correct retention is p_{11}/s_{base}, whereas

$$s_{\text{umm,r}} = \frac{p_{11}}{s_{\text{base}}} + \frac{p_{01}}{s_{\text{base}}}. \tag{14}$$

The relative score includes contributions from both retained and newly correct answers. The joint correctness counts quantify these two contributions separately. Averaging the joint cells over generation seeds preserves these identities because direct correctness is fixed.

In Figure 7, $\mathcal{H}_v$ denotes a schematic space of visual semantic content and $\mathcal{B}_d$ the content needed for the example question. The regions represent described and reconstructed content, whereas $\mathcal{G}_{\text{base}}$ and $\mathcal{G}_{\text{SGU}}$ above are sets of correct answers.

Figure 7 visualizes this abstraction. In case (a), the understanding–generation composition retains the information needed to answer the question, so the model remains correct after the SGU loop. In case (b), part of the relevant visual information is lost during the loop, leading to an incorrect final answer. This illustrates how SGU reflects the integrated behavior of understanding, intermediate representation, and generation.

D More SGU Case Examples

Figure 8 provides additional qualitative examples of the full SGU loop. These cases trace how visual information is described, reconstructed, and used for final reasoning. Comparing the intermediate outputs with the final answer helps locate successful information transfer and the stage at which a useful detail is lost.

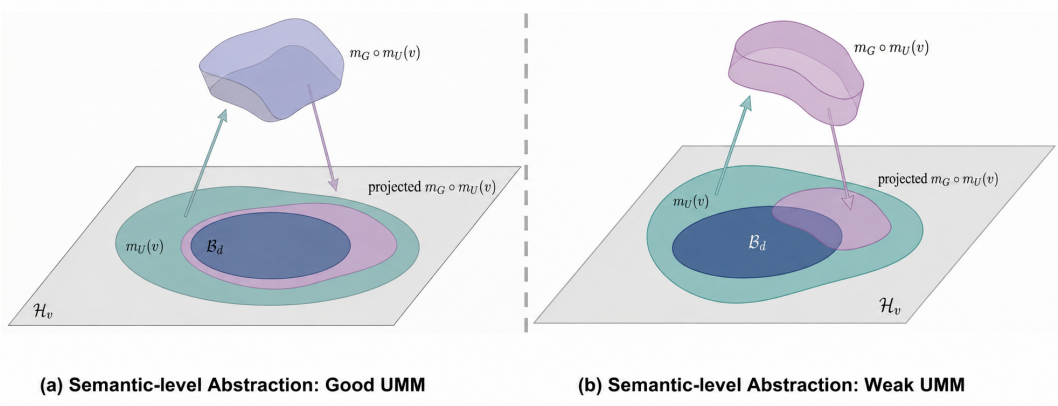

Figure 7: Abstraction-level illustration of SGU success and failure. (a) The understanding–generation composition retains the information needed for the target question. (b) Part of the information is lost during the loop, causing the reconstructed image to no longer support the correct answer.

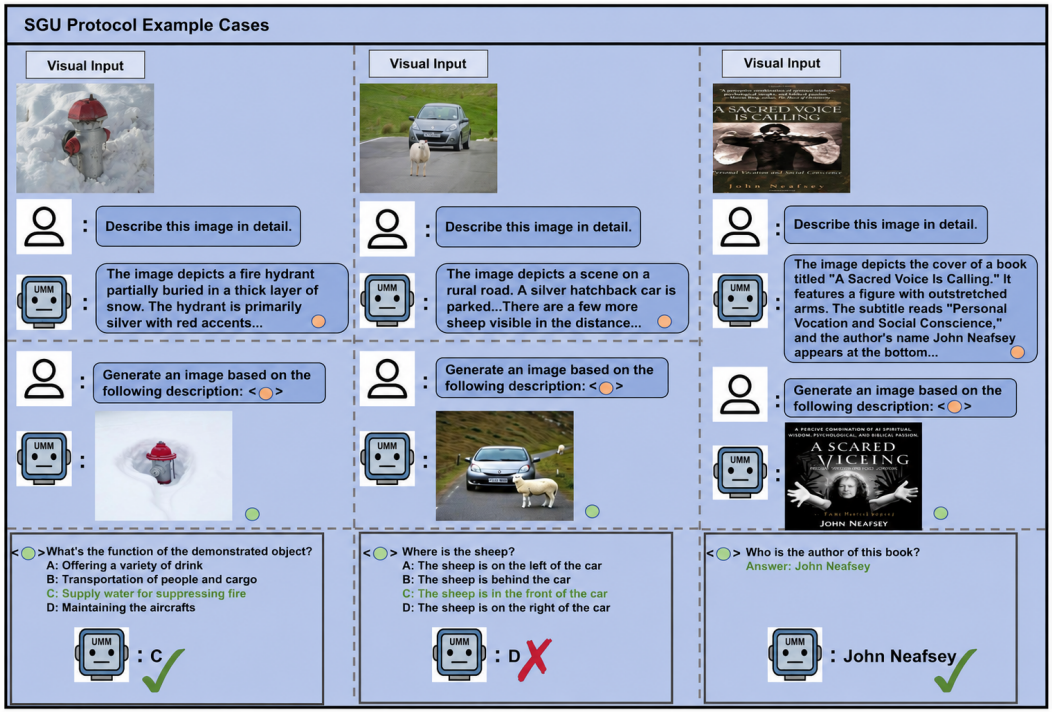

Figure 8: Additional SGU case examples showing the complete loop from the input VQA sample to the intermediate caption, reconstructed image, and final answer.

E Intermediate Similarity Results

This section reports intermediate similarity measures that complement the task-based SGU score.

E.1 Intermediate CLIP-T and CLIP-I Signals

To connect SGU with conventional component-level evaluation signals, we additionally report intermediate similarity metrics. Specifically, we compute **CLIP-T** between the model-generated caption and the original image, and **CLIP-I** between the reconstructed image and the original image.

Table 6: CLIP-T similarity between each model-generated caption and the original image. Higher is better.

Model	MMStar	MMBench	MathVista	OCR-VQA
Janus-Pro-7B	0.2841	0.2765	0.3102	0.3499
BAGEL-7B	0.2817	0.2703	0.3081	0.3158
UniWorld-V1	0.2820	0.2749	0.3026	0.3555
Show-o2-7B	0.2790	0.2704	0.3050	0.3189
OmniGen2	0.2794	0.2710	0.3038	0.3515
Ovis-U1-3B	0.2822	0.2778	0.3040	0.3497

Table 7: CLIP-I similarity between the reconstructed image and the original image. Higher is better.

Model	MMStar	MMBench	MathVista	OCR-VQA
Janus-Pro-7B	0.7126	0.7736	0.6494	0.6373
BAGEL-7B	0.7213	0.7577	0.6873	0.6881
UniWorld-V1	0.5790	0.6849	0.5058	0.3454
Show-o2-7B	0.6712	0.7316	0.6340	0.4752
OmniGen2	0.6946	0.7430	0.6860	0.6894
Ovis-U1-3B	0.7042	0.7515	0.6831	0.4909

These metrics are not used to compute SGU, but serve as auxiliary references for interpreting the captioning and reconstruction stages.

F Analysis of Information Transfer through the Loop

This section develops the connections between prompt sensitivity, intermediate descriptions, and the joint contribution of understanding and generation to SGU.

Prompt sensitivity. The comparison in Table 3 changes one instruction at a time while holding the model and task fixed. Across eight variants, the score range is 2.03 points. Captioning changes stay within 0.33 points of the default; directly using the caption as the generation input produces the largest change, $+1.85$ points.

Table 8: Text-interface analyses on OCR-VQA. Left: caption-only QA with Janus-Pro-7B at two caption-length limits. Right: OmniGen2 answers from the original image, its description, or its reconstruction.

Caption Length		Answering Path	
Setting	**Acc.**	**Setting**	**Acc.**
Janus-Pro-7B, 256 tokens	49.50	Original image QA (s_{base})	79.39
Janus-Pro-7B, 512 tokens	49.53	Caption-only QA	60.05
–	–	Full SGU (s_{umm})	56.59

Intermediate text bottleneck. The caption expresses the visual information selected by the model. Producing it requires selecting and expressing visual information in a form usable by the next stage. This makes the text interface part of the SGU task. A generic description may omit a fact that becomes relevant only after a particular question is supplied; question guidance helps distinguish this selection effect from difficulty extracting the requested evidence.

Question-guided captioning changes which information the description prioritizes. The instruction asks for visible evidence relevant to the question, rather than an answer or option identifier; the reference answer is never supplied. Generation receives only the resulting description, and final answering remains conditioned on the generated image and question. At a fixed caption budget, the

contrast with generic captioning probes sensitivity to task guidance. Caption-only QA and generated-image QA characterize its effects at two points in the process: the usability of the description and the usability of its visual reconstruction.

Caption length. Table 8 compares caption-only QA with maximum caption lengths of 256 and 512 tokens. On OCR-VQA with Janus-Pro-7B, accuracy changes from 49.50 to 49.53, a difference of 0.03 percentage points. Increasing the cap has little effect in this setting. Actual caption length and the text accepted by the generator determine how much of the larger budget reaches the next stage.

Understanding and generation contributions. Conditional generator comparisons assess how replacement gains depend on the native description. Let P_i indicate whether the fixed text-only probe correctly answers question i from that description; this partition is fixed before generator replacement. For a paired generation run, let ℓ_i^0 and ℓ_i^1 indicate final-answer correctness with the native and replacement generators, respectively. With $N_p = |\{i : P_i = p\}|$ and $N = N_0 + N_1$, the within-group and overall accuracy changes satisfy

$$\Delta_p = \frac{1}{N_p} \sum_{i:P_i=p} (\ell_i^1 - \ell_i^0), \qquad \Delta = \sum_{p\in\{0,1\}} \frac{N_p}{N} \Delta_p. \tag{15}$$

Multiplying by 100 gives percentage-point changes. Within each nonempty group, net gain equals rescues ($0 \to 1$) minus regressions ($1 \to 0$), divided by N_p. Weighting these changes by group size gives each group's contribution to the overall improvement. The rescue rate among baseline failures additionally measures how often the replacement succeeds on an available improvement opportunity.

Larger gains on probe-correct descriptions connect a stronger generator's benefit to task-relevant information supplied upstream. The generator can render that information in a form more useful for subsequent VQA. Smaller gains for the other group are consistent with completing partially specified visual content during generation. This conditional view explains how a large generation-replacement effect can coexist with an important understanding contribution.

The OmniGen2 comparison in Table 8 makes the intermediate text path explicit: original-image QA reaches 79.39, caption-only QA 60.05, and SGU 56.59. These input paths place the larger change before reconstruction for OmniGen2. Together with the complementary UniWorld profile, they show how caption-only results and component replacements locate different constraints on the integrated process.

Interpretation of the SGU score. SGU evaluates completion of the original task through the model's own outputs. Direct VQA, caption-only QA, image similarity, and component replacement each describe a different aspect of that process. Their joint interpretation makes the final score useful for diagnosis while retaining the task and protocol on which it is defined.

G More Experimental Results

G.1 Answering paths and image controls

Direct, Caption-QA, and SGU use the native model as the final reader. A fixed Qwen3-VL-8B probe answers separately from the native caption or reconstructed image. Comparing the native reader with this fixed probe characterizes how much the answer depends on the reader given the same intermediate content. Both probes receive the question and any answer options; the image probe receives neither the caption nor the original image.

Question-only inputs retain the question and, where available, its answer options. Mismatched inputs pair each question with a generated image from a different source, using a fixed derangement shared across models. All paths use the same question manifest and the benchmark's own evaluation protocol. Captions and generated images are cached, so reader comparisons use identical intermediate outputs. Tables 9 and 10 organize these comparisons by dataset and model.

Table 9: Path and input-control accuracy (%). Probe-T/I denote a fixed external reader on native captions/reconstructed images. Caption denotes the native model answering from its own caption.

Dataset	Model	SGU	Caption	Probe-T	Probe-I	No image	Wrong image
MMStar	UniWorld	38.33	44.08	47.29	40.11	24.79	23.73
MMStar	Janus	39.27	43.68	45.53	42.46	25.38	24.72
MMStar	Show-o2	43.00	49.16	48.44	44.17	26.23	24.94
MMStar	Ovis	43.07	50.92	55.35	45.71	27.10	26.26
MMStar	BAGEL	42.67	56.79	59.04	41.63	28.33	27.18
MMStar	OmniGen2	43.07	46.68	50.19	41.52	26.64	26.11
MMBench	UniWorld	71.97	76.13	79.76	73.47	38.45	37.99
MMBench	Janus	68.20	70.66	72.63	70.92	35.91	36.00
MMBench	Show-o2	73.95	77.54	77.01	74.45	38.01	37.81
MMBench	Ovis	75.10	78.23	82.87	77.34	39.06	39.19
MMBench	BAGEL	76.38	79.12	81.76	74.95	40.20	39.83
MMBench	OmniGen2	73.66	77.83	81.77	71.63	38.78	39.13

SGU exceeds question-only accuracy by 13.54 to 16.77 points on MMStar, 32.29 to 36.18 points on MMBench, 15.05 to 19.26 points on MathVista, and 20.31 to 44.42 points on OCR-VQA. Mismatched-image scores are close to the question-only scores, whereas the corresponding SGU advantages range from 14.55 to 46.07 points. Successful answering thus depends strongly on the supplied image corresponding to the question.

Table 10: Path and input-control accuracy (%) on MathVista and OCR-VQA. Probe-T/I use the same external reader as Table 9.

Dataset	Model	SGU	Caption	Probe-T	Probe-I	No image	Wrong image
MathVista	UniWorld	37.40	49.15	51.20	40.40	19.12	17.44
MathVista	Janus	32.50	38.83	39.59	36.87	17.45	16.42
MathVista	Show-o2	38.90	45.77	43.78	40.99	20.65	19.07
MathVista	Ovis	41.80	56.27	59.57	45.45	22.54	21.29
MathVista	BAGEL	39.50	60.65	61.53	39.75	23.65	22.15
MathVista	OmniGen2	40.40	53.70	55.92	40.03	21.62	20.27
OCR-VQA	UniWorld	28.67	63.30	68.52	33.75	8.36	6.08
OCR-VQA	Janus	37.45	49.50	53.35	44.09	9.01	7.55
OCR-VQA	Show-o2	30.34	47.18	48.56	34.86	7.57	5.35
OCR-VQA	Ovis	34.71	60.35	66.90	40.56	8.69	7.16
OCR-VQA	BAGEL	53.35	63.64	68.06	55.97	11.43	9.17
OCR-VQA	OmniGen2	56.59	60.05	65.50	58.42	12.17	10.52

The fixed image reader changes accuracy by -2.03 to $+6.64$ points relative to native SGU. Its gains are largest on OCR-VQA, where Janus rises from 37.45 to 44.09; on MMBench, it reduces OmniGen2's score from 73.66 to 71.63. The native reader therefore has neither a uniform advantage nor a uniform disadvantage on the same intermediate images.

Caption-QA exceeds SGU in every model–dataset combination, but the gap ranges from 2.46 to 34.63 points. For example, UniWorld's OCR-VQA description supports 63.30 accuracy, compared with 28.67 after reconstruction, while OmniGen2's corresponding scores are 60.05 and 56.59. Together with the replacement results, these paths locate model-specific losses along the loop. Independent evidence judgments in Appendix G.6 examine whether the generated image supports the reference answer.

G.2 Crossed component replacements

Let A denote the native captioner and generator, B an external captioner with the native generator, C the native captioner with an external generator, and D both external components. For each row, the corresponding UMM performs final VQA in all four conditions. A and C share the same native caption; B and D share the same external caption. Qwen3-VL-8B supplies external captions, and

Qwen-Image-2512 supplies external generations. Each generator receives the saved caption without a further rewrite.

The contrast $D - B - C + A$ measures whether the joint replacement differs from the sum of the two individual changes. A positive value indicates a larger joint gain; a negative value indicates a smaller joint gain. This comparison tests how component combinations affect task completion using each UMM's own VQA stage (Tables 11 and 12).

Table 11: Crossed component replacements. A: native; B: captioner replaced; C: generator replaced; D: both replaced. Each row retains the corresponding UMM for final VQA. Accuracy is in %; interaction $D - B - C + A$ is in percentage points.

Dataset	Model	A	B	C	D	Interaction
MMStar	UniWorld	38.33	39.53	44.86	44.82	-1.24
MMStar	Janus	39.27	39.87	42.03	43.70	1.07
MMStar	Show-o2	43.00	41.88	43.95	44.23	1.40
MMStar	Ovis	43.07	45.11	47.18	48.90	-0.32
MMStar	BAGEL	42.67	42.01	44.73	42.65	-1.42
MMStar	OmniGen2	43.07	44.57	46.79	48.95	0.66
MMBench	UniWorld	71.97	71.36	76.14	73.86	-1.67
MMBench	Janus	68.20	69.66	69.77	71.57	0.34
MMBench	Show-o2	73.95	74.28	73.64	74.60	0.63
MMBench	Ovis	75.10	74.55	77.59	76.16	-0.88
MMBench	BAGEL	76.38	77.18	77.50	76.39	-1.91
MMBench	OmniGen2	73.66	74.86	75.57	76.79	0.02

Table 12: Crossed component replacements on MathVista and OCR-VQA. A–D follow Table 11. A, B, and C are obtained from Tables 1 and 2.

Dataset	Model	A	B	C	D	Interaction
MathVista	UniWorld	37.40	33.24	51.04	45.98	-0.90
MathVista	Janus	32.50	34.38	36.04	39.14	1.22
MathVista	Show-o2	38.90	37.16	38.97	38.62	1.39
MathVista	Ovis	41.80	43.31	48.37	49.56	-0.32
MathVista	BAGEL	39.50	37.90	44.55	41.50	-1.45
MathVista	OmniGen2	40.40	44.33	52.01	56.64	0.70
OCR-VQA	UniWorld	28.67	26.15	67.56	63.30	-1.74
OCR-VQA	Janus	37.45	34.42	58.66	55.88	0.25
OCR-VQA	Show-o2	30.34	31.35	59.64	61.26	0.61
OCR-VQA	Ovis	34.71	36.73	69.56	70.57	-1.01
OCR-VQA	BAGEL	53.35	52.85	66.49	63.78	-2.21
OCR-VQA	OmniGen2	56.59	55.08	68.71	67.10	-0.10

The joint replacement does not consistently outperform generator replacement alone. On OCR-VQA, UniWorld decreases from 67.56 in C to 63.30 in D, whereas Show-o2 increases from 59.64 to 61.26. On MathVista, OmniGen2 increases from 52.01 to 56.64, while BAGEL decreases from 44.55 to 41.50. Across all four datasets, the interaction ranges from -2.21 to $+1.40$ points. These differences show that replacing a captioner also changes the input distribution encountered by the generator; component gains need not add independently.

The modular pipeline uses Qwen3-VL-8B for both captioning and answering, with Qwen-Image-2512 for generation (Table 13). Its generated images are also used in cell D, allowing the reader to change while image content is held fixed. Differences between the modular pipeline and native UMMs characterize the evaluated systems; their parameter counts, training data, and inference budgets also differ.

The modular pipeline reaches loop accuracies of 48.43, 79.05, 52.97, and 72.47 on MMStar, MMBench, MathVista, and OCR-VQA. These exceed the six native loop scores, while remaining 22.86, 9.44, 20.52, and 14.22 points below the modular reader's direct accuracies. The closed-loop chal-

Table 13: Modular pipeline under the SGU protocol (accuracy, %). Qwen3-VL-8B captions and answers; Qwen-Image-2512 generates.

Dataset	Direct	Caption-QA	Loop
MMStar	71.29	58.98	48.43
MMBench	88.49	80.31	79.05
MathVista	73.49	63.19	52.97
OCR-VQA	86.69	76.02	72.47

lenge therefore also exposes losses in a composition of specialist models. This comparison describes complete pipelines with their respective components and resources.

G.3 Which descriptions support generator gains?

A fixed text-only probe answers each question from the native description. Its correctness defines two groups before either generated image is evaluated. Both generation paths receive exactly the same caption. Group shares count questions; conditional accuracies and transitions average paired outcomes over the three generation seeds. Table 14 reports these quantities.

For group g, let r_g and f_g be the fractions changing from incorrect to correct and from correct to incorrect, respectively. Its net gain is $r_g - f_g$. If π_g is the group share, the aggregate gain is $\sum_g \pi_g(r_g - f_g)$. This identity checks the conditional analysis against the overall replacement effect. The contribution of group g to all rescued answers is $\pi_g r_g / \sum_h \pi_h r_h$. Dividing r_g by the group's baseline error rate additionally accounts for the available rescue opportunities.

Table 14: Conditional generator-replacement effects for UniWorld-V1. Share: group fraction; A/C: native/replaced-generator accuracy; Rescue/Regress: paired transitions within the group; R/fail: rescues among baseline errors. All entries are percentages except Gain (percentage points).

Dataset	Probe	Share	A	C	Gain	Rescue	Regress	R/fail
MathVista	Correct	51.20	54.87	79.07	24.20	27.93	3.73	61.89
MathVista	Incorrect	48.80	19.07	21.63	2.56	7.92	5.36	9.79
OCR-VQA	Correct	68.52	35.06	88.54	53.48	57.12	3.64	87.96
OCR-VQA	Incorrect	31.48	14.76	21.89	7.13	8.46	1.33	9.93

On MathVista, probe-correct descriptions constitute 51.20% of the questions and gain 24.20 points after generator replacement, compared with 2.56 points for probe-incorrect descriptions. They account for 78.72% of all rescued answers. On OCR-VQA, the corresponding group share is 68.52%, the net gains are 53.48 and 7.13 points, and the probe-correct group contributes 93.63% of rescues.

The contrast remains after accounting for baseline errors. Among initially incorrect answers, replacement rescues 61.89% versus 9.79% on MathVista, and 87.96% versus 9.93% on OCR-VQA. Most of the improvement is therefore concentrated where the native description already provides information accessible to the fixed probe. The smaller positive gains in the other group are consistent with the generator completing partially specified visual details that become useful for final VQA.

Weighting the group gains by their shares recovers the overall measured changes of 13.64 points on MathVista and 38.89 points on OCR-VQA to the reported precision. The probe-correct groups contribute 90.84% and 94.23% of these net gains, respectively. Both the rescued answers and the net improvement are concentrated in descriptions already usable by the probe, connecting the generator benefit to the information extracted upstream.

G.4 Caption length and question guidance

The comparison crosses a 256/512-token output cap with question-blind/question-guided captioning. The guided prompt is: *Describe the visible details in this image that are relevant to the following question. Do not answer the question. Question: {question}.* The captioner receives the image and question but no answer options or reference answer. The generator receives only the resulting caption.

For images associated with multiple questions, question-guided captioning creates a separate generation for each question. A question-blind per-question control matches this allocation. At 256 tokens, SGU guidance effects are compared against that control; length effects compare the two caps with question access and generation allocation held fixed. Caption-QA uses the same deterministic question-blind captions under either allocation. Caption-QA and SGU are reported together in Tables 15 and 16.

Table 15: Caption interventions (accuracy, %). B512: question-blind 512-token captions; Alloc: question-blind per-question allocation control; Q256/Q512: question-guided captions. C/L denote Caption-QA/loop accuracy.

Dataset	Model	B512-C	B512-L	Alloc-L	Q256-C	Q256-L	Q512-C	Q512-L
MMStar	UniWorld	44.97	38.84	38.43	45.34	39.02	46.17	39.22
MMStar	Janus	43.93	39.15	39.17	44.63	40.26	44.99	39.96
MMStar	Show-o2	50.49	43.80	43.15	50.74	42.92	51.88	43.42
MMStar	Ovis	51.50	43.26	43.00	52.25	43.80	52.79	43.73
MMStar	BAGEL	57.90	42.48	42.77	57.72	43.15	58.76	43.52
MMStar	OmniGen2	47.49	43.66	43.05	47.86	44.44	48.59	44.27
MMBench	UniWorld	76.72	72.40	72.09	77.19	72.51	77.79	72.74
MMBench	Janus	70.61	67.99	68.11	71.48	68.98	71.52	68.68
MMBench	Show-o2	78.51	74.64	74.13	78.91	73.73	79.84	74.23
MMBench	Ovis	78.52	75.18	75.04	79.43	75.67	79.70	75.60
MMBench	BAGEL	79.92	76.09	76.52	79.85	76.75	80.65	77.15
MMBench	OmniGen2	78.32	74.19	73.63	78.81	74.82	79.35	74.62

Table 16: Caption interventions on MathVista and OCR-VQA (accuracy, %). Notation follows Table 15. The Janus B512 caption result is also reported in Table 8.

Dataset	Model	B512-C	B512-L	Alloc-L	Q256-C	Q256-L	Q512-C	Q512-L
MathVista	UniWorld	50.26	38.02	37.49	50.85	38.59	51.84	38.76
MathVista	Janus	39.28	32.49	32.39	40.30	33.97	40.75	33.67
MathVista	Show-o2	47.26	39.80	39.05	47.83	39.36	49.10	39.86
MathVista	Ovis	57.10	42.10	41.71	58.09	43.00	58.85	42.91
MathVista	BAGEL	61.92	39.43	39.57	62.03	40.50	63.21	40.90
MathVista	OmniGen2	54.69	41.13	40.33	55.36	42.23	56.20	42.02
OCR-VQA	UniWorld	64.01	28.97	28.79	64.86	29.72	65.47	29.93
OCR-VQA	Janus	49.53	37.15	37.36	50.84	38.75	50.87	38.45
OCR-VQA	Show-o2	48.27	30.95	30.53	49.06	30.61	50.04	31.11
OCR-VQA	Ovis	60.75	34.68	34.66	62.04	35.76	62.40	35.65
OCR-VQA	BAGEL	64.53	52.95	53.46	64.88	54.17	65.71	54.59
OCR-VQA	OmniGen2	60.62	56.98	56.58	61.50	58.31	62.07	58.08

At the 256-token cap, question guidance improves Caption-QA by 0.73–2.06 points across the 24 model–dataset combinations. Its SGU effect relative to the allocation-matched control ranges from -0.40 to $+1.90$ points. For Show-o2 on MathVista, Caption-QA increases from 45.77 to 47.83, while the loop changes from 39.05 to 39.36. On MMBench, its guided caption raises Caption-QA from 77.54 to 78.91 while loop accuracy decreases from 74.13 to 73.73. Guidance can improve textual answerability, but the size and direction of the downstream change depend on the model and task.

Doubling the question-blind caption cap changes SGU by -0.40 to $+0.90$ points; under guidance, the corresponding range is -0.30 to $+0.50$. Janus on MathVista moves from 33.97 with Q256 to 33.67 with Q512, whereas Show-o2 moves from 39.36 to 39.86. A longer description can add useful evidence, but its extra details may be redundant or harder to render faithfully. Length alone therefore does not ensure a better reconstruction. These modest, mixed changes leave the main loop gaps in place under both caption caps.

Table 17 reports the caption length response. Mean lengths range from 117.93 to 165.66 tokens at the 256-token cap and from 123.84 to 190.27 tokens at the 512-token cap, increasing by 5.91–24.61

tokens across datasets. Cap-hit rates fall from 5.20–16.77% to 0.55–3.52%. With the same prompt and greedy decoding, only captions reaching the original cap can extend; most captions already terminate below it. The larger budget therefore reduces truncation for a minority of descriptions, while the mixed SGU changes indicate that the utility of the added details depends on their relevance and visual realization.

Table 17: Question-blind caption lengths and rates of reaching the token cap, macro-averaged over the six models. Token counts use each model's tokenizer and describe within-model length changes.

Dataset	Mean (256)	Cap hit (%)	Mean (512)	Cap hit (%)
MMStar	131.68	7.78	140.27	1.01
MMBench	117.93	5.20	123.84	0.55
MathVista	151.30	12.17	168.43	2.73
OCR-VQA	165.66	16.77	190.27	3.52

G.5 Paired correctness and answer retention

Let p_{bl} denote the fraction of paired outcomes with Direct correctness $b \in \{0, 1\}$ and loop correctness $l \in \{0, 1\}$, averaged over the three generation seeds. The joint cells sum to one, with $s_{\text{base}} = p_{11}+p_{10}$ and $s_{\text{umm}} = p_{11} + p_{01}$ when accuracies are expressed as fractions. Retention among directly correct answers is $p_{11}/(p_{11} + p_{10})$. Its difference from $s_{\text{umm}}/s_{\text{base}}$ is p_{01}/s_{base}: the aggregate ratio also includes questions that change from incorrect to correct. Table 18 separates these outcomes using paired predictions.

Table 18: Joint Direct/loop correctness (%). p_{bl} indexes Direct/loop correctness. Retention is $100p_{11}/s_{\text{base}}$ when both quantities are expressed in percent; Ratio is $100s_{\text{umm}}/s_{\text{base}}$.

Dataset	Model	p_{11}	p_{10}	p_{01}	p_{00}	Retention	Ratio
MathVista	UniWorld	34.20	33.10	3.20	29.50	50.82	55.57
MathVista	Janus	28.30	13.00	4.20	54.50	68.52	78.69
MathVista	Show-o2	35.26	15.44	3.64	45.66	69.55	76.73
MathVista	Ovis	39.06	29.84	2.74	28.36	56.69	60.67
MathVista	BAGEL	35.02	35.58	4.48	24.92	49.60	55.95
MathVista	OmniGen2	36.55	26.95	3.85	32.65	57.56	63.62
OCR-VQA	UniWorld	24.21	57.12	4.46	14.21	29.77	35.25
OCR-VQA	Janus	32.04	37.23	5.41	25.32	46.25	54.06
OCR-VQA	Show-o2	25.41	38.02	4.93	31.64	40.06	47.83
OCR-VQA	Ovis	30.72	48.28	3.99	17.01	38.89	43.94
OCR-VQA	BAGEL	47.78	26.77	5.57	19.88	64.09	71.56
OCR-VQA	OmniGen2	51.46	27.93	5.13	15.48	64.82	71.28

Wrong-to-correct transitions occur for 2.74–5.57% of the questions in these comparisons. These newly correct answers contribute to loop accuracy alongside answers preserved from the direct path. For UniWorld on OCR-VQA, the relative SGU score is 35.25% and retention among directly correct answers is 29.77%; for Janus on MathVista, the values are 78.69% and 68.52%. The paired analysis separates these contributions and traces how the original-image performance changes through reconstruction.

G.6 Independent evidence assessment

The assessment uses randomly selected source images from MathVista and OCR-VQA, one fixed question per source, and the six native model outputs at seed 42. Three raters independently inspect every output, with model identities hidden. The first pass presents only the generated image and question. Each rater supplies an answer, identifies visible supporting evidence, and marks the question answerable, insufficient, or uncertain. After these responses are locked, the second pass reveals the original image and reference answer. The rater then labels the reconstructed image as supporting, contradicting, omitting, or leaving uncertain the reference evidence. Overall fidelity and task-relevant fidelity receive separate 1–5 ratings. Model answers and captions remain hidden throughout.

For the binary endpoint H, at least two support judgments define $H = 1$; at least two judgments in the combined contradicting and omitting categories define $H = 0$. Remaining outputs are unresolved. Fleiss' κ uses the four second-pass categories before majority aggregation. All metric comparisons use the same resolved-output mask. Table 19 separates correct answers without supporting evidence from missed answers despite supporting evidence, while Table 20 compares associations with the human endpoint. The joint outcomes relate task completion under SGU to independently judged visual evidence.

Table 19: Human evidence assessment. L denotes SGU correctness; H denotes evidence support. Coverage is the percentage of audited outputs with a resolved label. The four joint columns are percentages of resolved outputs. Agreement is Fleiss' κ; fidelity is the mean task-relevant rating.

Task	Coverage	H_1L_1	H_0L_1	H_1L_0	H_0L_0	κ	Fidelity
MathVista	94.8	30.49	10.44	25.27	33.79	0.48	3.00
OCR-VQA	96.9	37.90	6.18	17.20	38.71	0.67	3.30

Table 20: Spearman associations with human-judged evidence support on the same resolved outputs. For two binary variables, this equals the phi coefficient.

Task	SGU	CLIP-I	CLIP-T	Direct	Caption-QA
MathVista	0.314	0.288	0.119	0.241	0.301
OCR-VQA	0.551	0.482	0.173	0.108	0.461

Evidence judgments resolve 94.8% of MathVista outputs and 96.9% of OCR-VQA outputs. Rater agreement is 0.48 and 0.67, respectively, with mean task-relevant fidelity ratings of 3.00 and 3.30. Among correct loop answers, 25.51% on MathVista and 14.02% on OCR-VQA lack majority-judged supporting evidence. Conversely, among images judged to contain supporting evidence, 45.32% and 31.22%, respectively, receive an incorrect model answer. These joint outcomes show how visual evidence is used in downstream task completion.

Their Spearman associations are 0.314 on MathVista and 0.551 on OCR-VQA. The corresponding CLIP-I values are 0.288 and 0.482, and Caption-QA values are 0.301 and 0.461. SGU has the largest point estimate in each task, with a smaller separation on MathVista. This task dependence matches the joint outcomes: a usable mathematical diagram still requires successful reasoning, while missing or corrupted text is a more direct obstacle in OCR-VQA.

Added predictive information. Source images are divided equally into development and test partitions within each dataset before the human labels are inspected, keeping all outputs from each source together. A base predictor uses Direct correctness, Caption-QA correctness, CLIP-T, and CLIP-I; an augmented predictor adds SGU correctness. Both use L2 logistic regression with $C = 1$, an unpenalized intercept, and features standardized on the development data. Model fitting uses development labels only. Held-out Brier score and log-loss compare the predictors in Table 21; negative augmented-minus-base differences indicate improved prediction of independent evidence support.

Table 21: Held-out prediction of human-judged evidence support. Lower Brier score and log-loss are better. Differences are augmented minus base.

Task	Brier base	+SGU	Δ	Log-loss base	+SGU	Δ
MathVista	0.221	0.219	-0.002	0.634	0.636	0.002
OCR-VQA	0.207	0.188	-0.019	0.598	0.557	-0.041

Adding SGU reduces OCR-VQA test Brier score from 0.207 to 0.188 and log-loss from 0.598 to 0.557. MathVista shows a smaller Brier change, 0.221 to 0.219, together with a slight increase in log-loss, 0.634 to 0.636. The added predictive information is clear on OCR-VQA; MathVista shows a weaker, mixed change across the two criteria, consistent with its additional reasoning demands.

G.7 GENERATION-ONLY COMPARISON

GenEval evaluates generation directly from its official prompts, without captioning an input image. Table 22 compares the six UMMs under a common prompt list, evaluator revision, and number of outputs per prompt. The generation-only endpoint and SGU test different inputs and success criteria: one follows a supplied text specification, while the other reconstructs a model's own description for a downstream question.

Table 22: Generation-only evaluation with the same GenEval protocol, reported on a 0–1 scale.

Generator	GenEval
UniWorld-V1	0.78
Janus-Pro-7B	0.80
Show-o2-7B	0.71
Ovis-U1	0.87
BAGEL-7B	0.81
OmniGen2	0.79

The six native generators score 0.71–0.87 on GenEval. Ovis scores 0.87 and BAGEL 0.81, compared with 0.79 for OmniGen2; their OCR-VQA loop scores are 34.71, 53.35, and 56.59, respectively. The generation-only ordering thus differs from the loop ordering. The object-focused text-prompt task and the reconstruction task emphasize different capabilities, providing complementary views of these models.

Aggregation and pairing. Benchmark accuracy assigns equal weight to questions and averages generated-image outcomes over the three seeds. Direct and caption-only answers use deterministic decoding. Component and prompt contrasts pair the same question and seed across conditions; repeated questions sharing a source image also share the cached reconstruction in the default protocol. Human comparisons instead use seed 42 and the common resolved-output mask described above. Outputs from one source image are kept together when defining the development and test partitions.