YANAMI: A LARGE-SCALE MULTIMODAL DATASET AND BENCHMARK FOR MULTITURN STATEFUL INTERACTION

Anonymous authors
Paper under double-blind review



Figure 1: **YANAMI:** the first large-scale multimodal dialogue dataset and benchmark for stateful interaction.

ABSTRACT

Unified Multimodal Models (UMMs) integrate visual understanding, generation, and editing within a single unified framework. However, existing unified multimodal datasets and benchmarks remain dominated by single-turn tasks. Even in recent multi-turn datasets, many turns can be solved independently without tracking information across turns. Consequently, they provide limited supervision and evaluation for multimodal *stateful interaction*, which requires models to understand, maintain, update, and faithfully realize evolving dialogue states across turns. To address this gap, we introduce **YANAMI** (**Y**our **A**ssistant **N**eeds to **A**ssimilate **M**ultiturn **I**nteraction), the first large-scale multimodal dialogue dataset and benchmark for stateful interaction, comprising **YANAMI-400K** and **YANAMI-Bench**. YANAMI-400K contains 400K multi-turn multimodal dialogues, with more than 2M images and 2.6M dialogue turns. The dialogues span four representative scenarios, Multi-Image Fusion, Image Editing, World Modeling, and Visual Reasoning, and require models to continuously track, preserve, and revise evolving multimodal dialogue states across turns. YANAMI-Bench comprises 400 multi-turn dialogues for systematically evaluating multimodal stateful interaction. It jointly evaluates **Dialogue State Understanding** and **Visual Realization**, characterizing the gap between models' ability to understand evolving dialogue states and their ability to faithfully realize them in visual outputs. Experiments show that fine-tuning BAGEL on YANAMI-400K improves image editing, generation, and multimodal stateful interaction by 10.3%, 2.0%, and 13.23%, respectively, while preserving its general visual understanding capabilities. Evaluations on YANAMI-Bench further reveal that existing UMMs still struggle to maintain, update, and visually realize evolving multimodal dialogue states. Together, YANAMI provides a unified foundation for training and evaluating UMMs toward robust multimodal stateful interaction.

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.

1 INTRODUCTION

Unified Multimodal Models (UMMs) (Deng et al., 2025; Wu et al., 2025c; Xie et al., 2026a) represent an emerging shift from task-specific systems toward unified visual modeling. Unlike conventional models designed for individual tasks (Liu et al., 2023; Rombach et al., 2022), UMMs integrate image understanding, generation, and editing within a single framework. This unified modeling enables UMMs to interleave processing across textual and visual modalities, highlighting their potential for complex multi-turn multimodal interaction.

However, most publicly available datasets focus on single-turn tasks (Chen et al., 2025a; Wang et al., 2025b). Only a few support multi-turn interaction (Chow et al., 2026b; Feng et al., 2026), and even these often contain turns that can be solved *independently* without tracking information across turns, falling short of complex multi-turn multimodal interaction. In contrast, real-world user-assistant image creation and editing are typically not completed in a single step, but instead involve continuous multi-turn *stateful interaction* for gradual refinement.

Here, *stateful interaction* refers to a multi-turn dialogue in which the user and assistant discuss alternatives, make decisions, establish cross-turn constraints and dependencies, plan subsequent steps, and revise previous decisions or requirements in response to user corrections. Consequently, datasets without such interactions provide limited supervision for learning to understand, maintain, update, and faithfully realize evolving multimodal *dialogue states* across turns. It further limits the exploration of UMMs for complex multi-turn multimodal interaction. For example, as illustrated in Fig. 1, existing multi-turn datasets typically alternate between user instructions and assistant edits, with each turn treated independently and lacking cross-turn contextual dependencies, whereas our dataset captures stateful user–assistant interactions in which dialogue states are established, maintained, and updated through ongoing context-dependent exchanges.

To address these limitations, we introduce **YANAMI**, the first large-scale multimodal dialogue dataset and benchmark for *stateful interaction*, comprising **YANAMI-400K** and **YANAMI-Bench** to capture evolving multimodal dialogue states and evaluate models' performance across understanding, generation, and editing. YANAMI-400K contains **400K** multi-turn multimodal dialogues, comprising more than **2M** images and **2.6M** dialogue turns in total. To construct genuinely stateful interactions, we explicitly incorporate six key capabilities into each dialogue: *1) Reference Resolution*; *2) Context Retrieval*; *3) Consistency Maintenance*; *4) Spatial and Physical Consistency*; *5) Constraint Keeping*; and *6) Correction-Based State Updating*. Together, these capabilities make stateful interaction explicit in the data, with each dialogue requiring multimodal states to be continuously tracked, preserved, and revised across turns. To instantiate such stateful interaction in diverse settings, we construct dialogues under four representative scenarios: *1) Multi-Image Fusion*; *2) Image Editing*; *3) World Modeling*; and *4) Visual Reasoning*. These scenarios jointly exercise diverse capabilities, spanning visual-state composition, modification, transition, and reasoning.

YANAMI-Bench comprises 400 multi-turn dialogues designed to systematically evaluate models' ability to track, maintain, and update multimodal dialogue states throughout interactions. Specifically, YANAMI-Bench covers the same four scenarios, including complex understanding, generation, and editing tasks, with each dialogue jointly involving all six key capabilities and requiring their coordinated use throughout the interaction. YANAMI-Bench evaluates models along two complementary dimensions: **Dialogue State Understanding (DSU)** and **Visual Realization (VR)**. **DSU** evaluates whether a model can infer the currently valid multimodal state from the full interaction history. We assess it through capability-oriented dialogue-based visual question answering along four dimensions: *1) Reference Resolution*; *2) Context Retrieval*; *3) Constraint Keeping*; and *4) Correction-Based State Updating*. VR measures whether the generated images faithfully realize the current dialogue state and is assessed by *Visual Quality* and *State Consistency*. Together, these two dimensions provide a unified evaluation of stateful multimodal interaction.

Our experiments demonstrate that training on YANAMI-400K substantially improves model performance on image generation, image editing, and multi-turn stateful interaction tasks while maintaining general visual understanding capabilities. Meanwhile, evaluations on YANAMI-Bench reveal several key limitations of current models in long-horizon multimodal interactions. We hope that YANAMI provides a unified foundation for training and evaluating UMMs toward robust multimodal stateful interaction. In summary, our main contributions are as follows:

- We introduce YANAMI-400K, the first large-scale multimodal dialogue dataset for stateful interaction across understanding, generation, and editing. It comprises **400K** multi-turn dialogues with more than **2M** images and **2.6M** dialogue turns, providing large-scale supervision for six key capabilities required for stateful interaction.
- We construct YANAMI-Bench, a benchmark of 400 multi-turn dialogues. By jointly evaluating DSU and VR, YANAMI-Bench characterizes the gap between models' ability to understand dialogue states and their ability to faithfully realize them in visual outputs.
- Fine-tuning BAGEL (Deng et al., 2025) on YANAMI-400K yields consistent gains in general image generation and editing, as well as DSU and VR, while preserving general visual understanding capabilities. Results on YANAMI-Bench reveal that existing UMMs still face substantial challenges in maintaining, updating, and visually realizing evolving multimodal dialogue states.

2 RELATED WORK

Unified Multimodal Models aim to integrate visual understanding and generation within a shared architecture, providing a unified interface for processing and producing multimodal content. Existing UMMs explore diverse architectural paradigms, ranging from unified autoregressive modeling(Team, 2024; Wu et al., 2025b) over multimodal representations to hybrid frameworks that combine autoregressive language modeling with diffusion- or flow-based visual generation (Deng et al., 2025; Wu et al., 2025c). Existing UMMs have demonstrated increasing capabilities across single-turn multimodal understanding, generation, and editing tasks (Chen et al., 2025d; Cui et al., 2025). However, how well these capabilities extend to multi-turn multimodal dialogues with evolving contextual dependencies remains largely underexplored.

Multimodal Data and Benchmarks for Understanding, Generation, and Editing. Existing multimodal datasets primarily provide single-turn supervision for individual capabilities such as understanding (Liu et al., 2023), generation (Brooks et al., 2023), and editing (Yu et al., 2025), with some extending to multi-image settings (Chen et al., 2025c; Li et al., 2025). Recent work has further explored multi-turn interaction through sequential editing and interleaved dialogues (Chow et al., 2026b; Qian et al., 2026; Dai et al., 2026). Meanwhile, existing benchmarks largely evaluate individual capabilities in isolation, including understanding (Fu et al., 2026; Liu et al., 2024b; Li et al., 2023), generation (Ghosh et al., 2023), and editing (Liu et al., 2025). More recent efforts have begun to evaluate unified or interleaved capabilities (Xia et al., 2025; Xie et al., 2026b) and multi-turn settings spanning understanding, editing, generation, and unified multimodal interaction (Liu et al., 2024a; Chen et al., 2025b; Huang et al., 2025; Chow et al., 2026b). However, these datasets and benchmarks can often be approached turn by turn without explicitly tracking evolving information across the dialogue, limiting their ability to support and evaluate truly multimodal stateful interaction. In contrast, YANAMI jointly provides large-scale supervision and systematic evaluation for tracking, updating, and faithfully realizing evolving multimodal dialogue states across turns. The comparison between YANAMI and multimodal datasets is provided in Appendix A.1.

3 YANAMI

In this section, we present YANAMI, our dataset and benchmark for multimodal stateful interaction. YANAMI-400K provides large-scale training data, while YANAMI-Bench evaluates models across understanding, generation, and editing. We first describe the data construction process in Sec. 3.1, followed by the evaluation setting and metrics in Sec. 3.2 and the dataset statistics in Sec. 3.3.

3.1 DATA PIPELINE

YANAMI-400K. We construct novel stateful dialogues using visual materials drawn from four existing public datasets (Li et al., 2025; Chow et al., 2026b; Wei et al., 2026b; Chen et al., 2026). Our pipeline, illustrated in Fig. 3, leverages GPT-5.6 Sol (OpenAI, 2026a), Qwen3.7-Plus (Alibaba Cloud, 2026), Gemini 3.1 Pro (Google, 2026a), GPT-Image-2 (OpenAI, 2026b), and Nano Banana 2 (Google, 2026b) for data filtering, dialogue construction, and quality control. Specifically, raw images are filtered into four scenario-specific pools (Fig 3.a) and combined with scenario- and

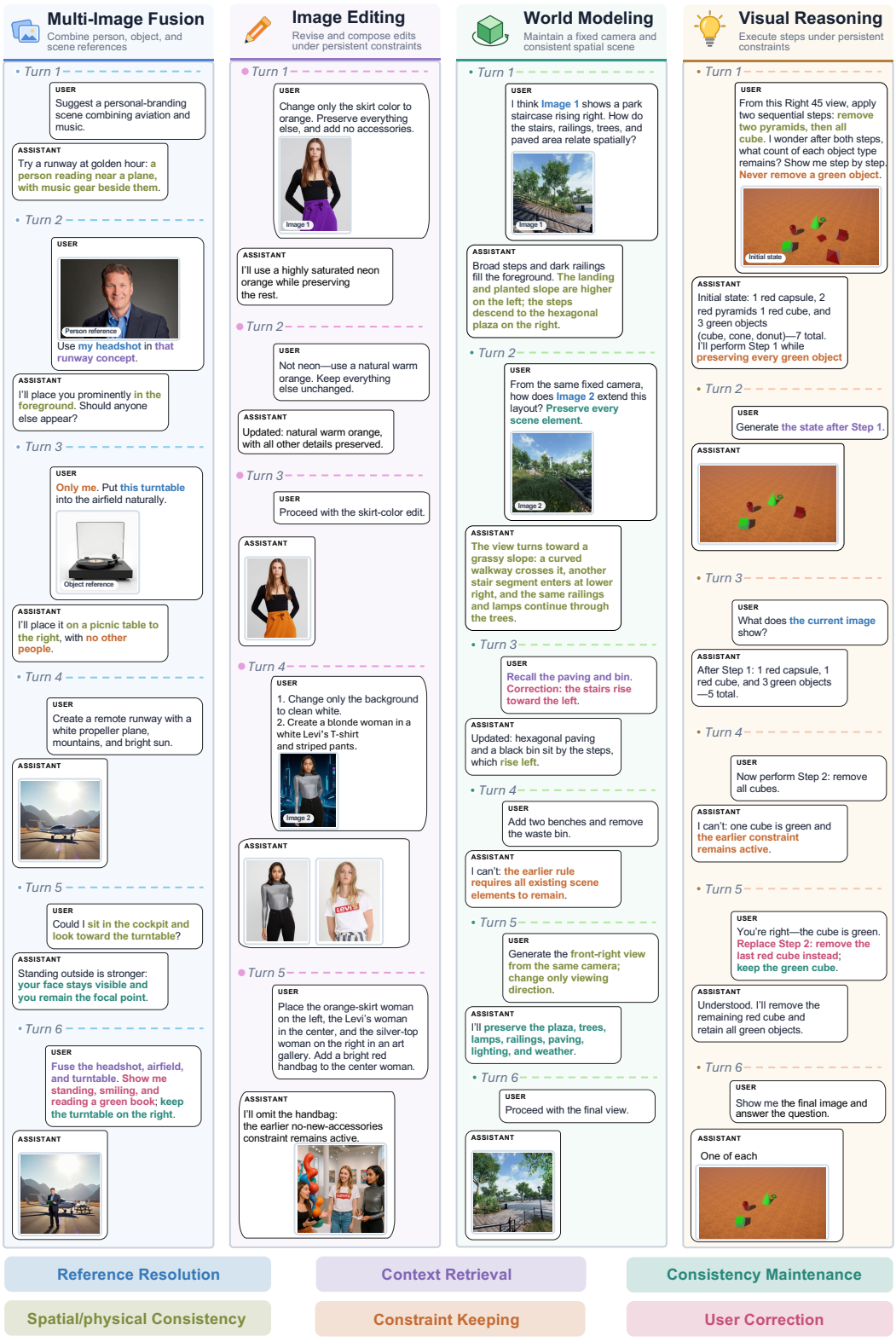


Figure 2: **Representative examples from YANAMI-400K.** The four columns present evolving multimodal dialogues in four interaction scenarios: (a) multi-image fusion, (b) image editing, (c) world modeling, and (d) visual reasoning. Across turns, subsequent responses retrieve, maintain, and update the accumulated dialogue state. Colored text highlights six key capabilities.

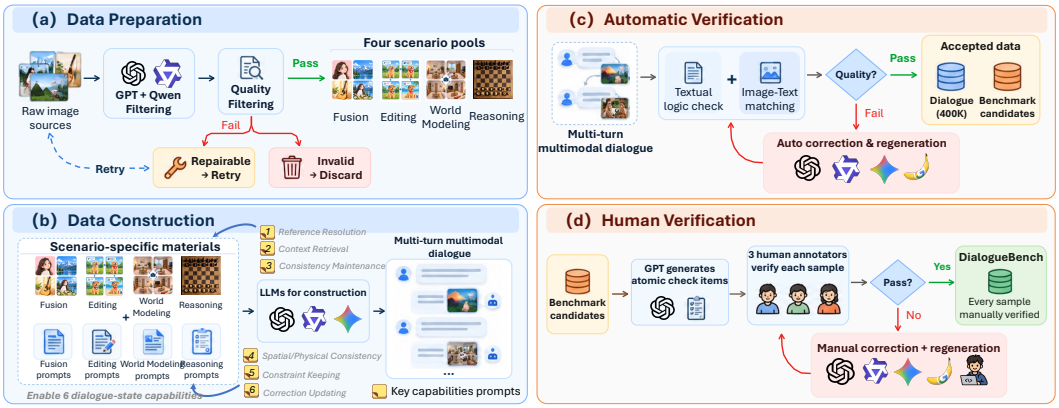


Figure 3: **Overview of the YANAMI data pipeline.** The pipeline comprises four stages: (a) Data Preparation, (b) Data Construction, (c) Automatic Verification, and (d) Human Verification.

Statistic	Number
Total Dialogues	402,639
$\geq$ 4 Images	321,868
$\geq$ 5 Images	215,027
$\geq$ 6 Images	138,669
Total Images	2,047,236
Avg. / Dialogue	5.08
Total Turns	2,643,291
Avg. / Dialogue	6.56
Max. / Dialogue	43

Table 1: Key Statistics of our Dataset.

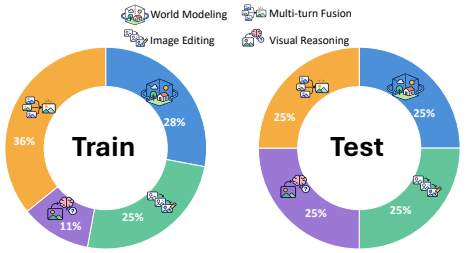


Figure 4: Distribution of the four interaction scenarios in the training and test sets.

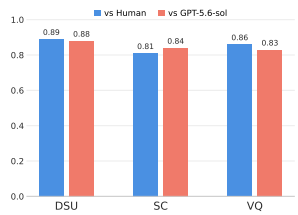


Figure 5: Evaluation reliability of Our VLM-Judge.

capability-oriented prompts to construct multi-turn multimodal stateful dialogues (Fig 3.b). The resulting dialogues are automatically verified for logical coherence and image–text consistency, with failed samples iteratively corrected and regenerated. Dialogues that pass verification form YANAMI-400K (Fig 3.c). The constructed dialogues span four representative scenarios: *1) Multi-image fusion,* where the assistant integrates multiple user-provided or previously generated images according to the dialogue history; *2) Image editing,* where the assistant progressively edits an image while preserving prior modifications and visual consistency; *3) World modeling,* where the assistant predicts and renders scene states under specified viewpoint changes or actions; and *4) Visual reasoning,* where the assistant generates visual outcomes for proposed hypotheses or solutions to support step-by-step reasoning.

To make these interactions genuinely stateful, we incorporate six key state-management capabilities across the four scenarios: *(i) Coreference resolution,* which grounds context-dependent expressions to the correct textual or visual entities; *(ii) Dialogue-history retrieval,* which identifies and uses relevant information from previous turns; *(iii) Cross-turn consistency,* which preserves established visual attributes and prior edits across turns; *(iv) Spatial and physical consistency,* which maintains plausible geometry, physical relations, and state transitions; *(v) Constraint adherence,* which consistently follows user-specified requirements and restrictions; and *(vi) Correction-based state updating,* which incorporates user corrections and replaces outdated information accordingly. For example, Fig. 2 presents representative dialogues across the four scenarios, illustrating how multimodal dialogue states are progressively established, maintained, and updated through the joint use of these capabilities. Additional details of the data collection and representative examples are provided in Appendices A and E, respectively.

YANAMI-Bench. YANAMI-Bench is constructed with LLM assistance, followed by exhaustive human verification and correction by annotators with graduate-level training in STEM disciplines (Fig 3.d). Throughout data construction, we strictly isolate the benchmark from the training dataset to prevent data leakage; as a final safeguard, we compare the SHA-256 hashes of all images across YANAMI-Bench and YANAMI-400K, confirming that no exact duplicates exist between them. It comprises 400 multi-turn dialogues covering the four scenarios described above.

The benchmark contains complex instructions that require reasoning over cross-turn context, jointly evaluating models' capabilities in ability to track, maintain, and update multimodal dialogue states and translating the understanding of the state into visual implementation.

3.2 Evaluation Setting and Metrics

To better quantify models' stateful interaction capabilities in unified multi-turn multimodal dialogues, we evaluate them from two complementary perspectives: dialogue state understanding and visual realization. Specifically, we adopt a VLM-as-a-Judge framework for automatic evaluation.

For DSU, we construct four types of QA pairs for each dialogue: *1) Reference Resolution* measures whether the model can correctly resolve referring expressions in the current instruction and associate them with the corresponding entities in the dialogue history or previous images. *2) Context Retrieval* measures whether the model can retrieve visual or textual information relevant to the current task from the multi-turn interaction history. *3) Constraint Tracking* evaluates whether the model can identify and continuously retain constraints introduced in previous turns that remain valid at the current turn; and *4) Correction Updating* measures whether the model can discard invalidated states and update the multimodal dialogue state after the user corrects previously established information, requirements, or states.

For VR, we evaluate generated or edited visual outputs from two aspects: *Visual Quality*, which measures overall perceptual quality, and *State Consistency*, which uses fine-grained atomic criteria to assess whether the output faithfully reflects all valid multimodal dialogue states accumulated up to the current turn. For each scenario, we first average the four DSU metrics and then combine the resulting understanding score with State Consistency and Visual Quality to obtain the aggregate score. For models without understanding capabilities, the aggregate score is computed from State Consistency and Visual Quality alone. The detailed evaluation templates and score calculation are provided in Appendix F and Appendix B.3, respectively.

3.3 Data Statistics

Table 1 summarizes the key dataset statistics. Most dialogues contain at least four images and have an average of 6.56 conversational turns. As shown in Fig. 4, the training set includes all four interaction scenarios in varying proportions, while the test set is evenly balanced across them to enable controlled cross-scenario evaluation. Detailed statistics are provided in Appendix A.6.

4 Experiment

In this section, we first evaluate 15 models on YANAMI-Bench, revealing substantial limitations of existing models in understanding, maintaining, updating, and visually realizing multimodal dialogue states. We then validate the effectiveness and high quality of YANAMI-400K by fine-tuning BAGEL in Sec. 4.2. Finally, we evaluate the reliability of our VLM-as-a-Judge evaluation framework.

4.1 YANAMI-Bench

Settings. We evaluate 2 close-source models, 4 open-source image editing models, and 9 unified multimodal models on YANAMI-Bench, with the results reported in Table 2. Complete capability-level results for each interaction scenario are provided in Appendix D.2. To isolate models' ability to leverage multi-turn dialogue context from error propagation caused by previous generations, we adopt a *GT-History* evaluation setting. For image placement, we consider three configurations depending on the model setup: (1) yes-first, where each image is inserted at its first-mentioned position in the dialogue, (2) yes-front, where all images are placed at the beginning of the input, and (3) concatenate, where all images are concatenated into one image for models that cannot directly process sequential multi-image inputs. Implementation details for the history settings and input organizations are provided in Appendix B.4. Based on the results, we can derive the following key findings:

Stateful Multi-Turn Multimodal Interaction Remains Challenging. As shown in Table 2, the best average scores achieved by existing UMMs and open-source image editing models on YANAMI-Bench are only 67.41 and 67.30, respectively, indicating that current models still face

Index	Model	Size	Modality	Format	Fusion	Editing	Modeling	Reasoning	Overall
1	GPT-Image-2 (OpenAI, 2026b)	–			86.78	90.54	81.40	73.46	83.05
2	Nano Banana 2 (Google, 2026b)	–			84.73	87.32	74.74	76.93	80.93
3	Step1X-v1p2 (Liu et al., 2025)	12B			56.08	63.36	51.66	48.00	54.77
4	Qwen-Edit (Wu et al., 2025a)	20B			49.66	54.70	28.14	37.72	42.55
5	FLUX.2 [dev] (Labs, 2025)	32B			72.44	74.54	65.54	56.68	67.30
6	UniPic 3.0 (Wei et al., 2026a)	20B			63.44	70.66	56.96	48.82	59.97
7	OmniGen (Xiao et al., 2025)	3.8B			48.24	58.12	41.82	33.94	45.53
8	OmniGen2 (Wu et al., 2025c)	7B			50.10	51.70	55.12	47.62	51.13
9	Janus-Pro (Chen et al., 2025d)	7B			45.94	41.94	42.56	31.86	40.58
10	Show-o2 (Xie et al., 2026a)	7B			40.82	45.12	41.78	29.42	39.29
11	Ovis-U1 (Wang et al., 2025a)	3B			68.72	46.24	57.72	43.42	54.03
12	Emu3.5 (Cui et al., 2025)	34B			74.48	60.10	74.06	60.98	67.41
13	InternVL-U (Tian et al., 2026)	4B			66.72	66.12	66.48	46.52	61.46
14	UniAR (Peng et al., 2026)	10B			57.06	41.82	41.80	35.92	44.15
15	BAGEL (Deng et al., 2025)	7B			62.44	67.40	68.16	62.06	65.02
	+ YANAMI-400K	7B			78.98↑	84.96↑	73.26↑	75.80↑	78.25↑

Table 2: **Results on YANAMI-Bench across four interaction scenarios.** denotes models supporting image generation/editing only; denotes models supporting both understanding and generation. denotes concatenated image inputs; denotes placing all available images at the beginning; and denotes interleaved image–text inputs. Red, blue, and yellow cells highlight the best results among closed-source models, open-source image-editing models, and UMMs. Underlined results denote the second-best performance within each model group.

substantial challenges in multimodal multi-turn stateful interaction. Further examination across different interaction scenarios reveals clear performance variation, with Visual Reasoning yielding the lowest scores for most evaluated models. In contrast, GPT-Image-2 and Nano Banana 2 achieve average scores of 83.05 and 80.93, respectively, outperforming current open-source models overall and suggesting that a noticeable gap remains between open- and closed-source models in multimodal multi-turn stateful interaction.

Multi-turn Dialogue-State Understanding Remains Challenging.

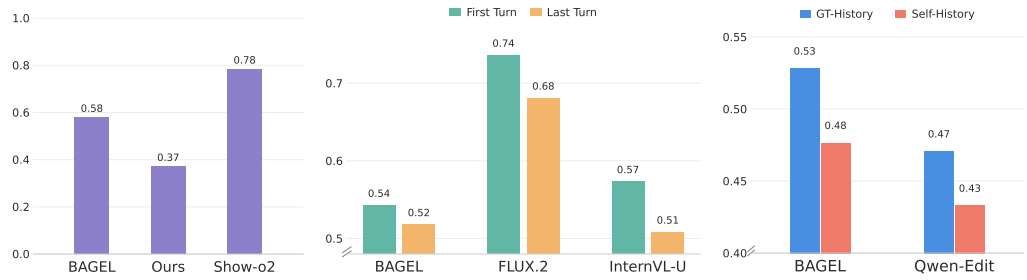
We further find that models struggle not only with generating and editing visual outputs, but also with understanding the current multimodal dialogue state. As shown in Table 3, the average DSU scores of most UMMs remain below 82, indicating that they still have difficulty accurately recovering and maintaining the currently valid dialogue state as it evolves over multiple turns, including historical references, relevant contextual information, persistent constraints, and state updates following user corrections.

Model	Fus.	Edi.	Mod.	Rea.	Avg.
Show-o2	74.00	64.25	79.50	54.00	67.94
BAGEL	85.75	76.50	85.25	78.25	81.44
BAGEL + Ours	**96.25**↑	**98.00**↑	**95.50**↑	**93.00**↑	**95.69**↑

Table 3: Selected average DSU results. An upward arrow indicates an improvement over BAGEL. Full results are provided in Appendix D.2.

Dialogue-State Understanding–Visual Realization Gap. We further investigate the gap between state understanding and visual realization in multimodal multi-turn interaction. For a fair comparison, we retain only state-understanding items on which all evaluated models receive full credit and compute each model's normalized realization score using the same aligned generation-side capability checks; the understanding–realization gap is then defined as one minus this conditional realization score. Details of the evaluation protocol are provided in Appendix B.6. As shown in Fig. 6a, even when a model correctly understands a currently valid constraint or a corrected state, it does not always faithfully realize that information in the generated output. This reveals a clear gap between state understanding and visual realization in existing UMMs.

Performance Across Dialogue Progression. We analyze how model performance changes as multi-turn interactions progress. Fig. 6b compares the performance of three representative models at the first and last image-generation turns of dialogues containing at least two generation turns. Since our



(a) Und. Realization Gap (b) First vs. Last Generation Turn (c) GT-History vs. Self-History

Figure 6: **Analysis of stateful multi-turn multimodal interaction.** (a) Measuring the understanding-realization gap, lower is better; (b) comparing performance at the first and last image-generation turns under GT-History, to assess the effect of increasingly accumulated context and state dependencies; (c) comparing GT-History with Self-History on dialogues to measure error accumulation when self-generated images are used as subsequent visual context.

Model	Understanding			GenEval						
	POPE	MME-P	MMBench	Single Obj.	Two Obj.	Count.	Color	Position	Attri.	Overall
Emu3 (Wang et al., 2024)	85.2	1243.8	58.5	0.99	0.81	0.42	0.80	0.49	0.45	0.66
Show-o (Xie et al., 2025)	80.0	1097.2	–	0.98	0.80	0.66	0.84	0.31	0.50	0.68
Janus-Pro-7B (Chen et al., 2025d)	87.4	1567.1	79.2	0.99	0.89	0.59	0.90	0.79	0.66	0.80
MetaQuery-XL (Pan et al., 2025)	–	1685.2	83.5	–	–	–	–	–	–	0.80
Ovis-U1 (Wang et al., 2025a)	–	1508.0	77.8	0.98	0.98	0.90	0.92	0.79	0.75	0.89
Show-o2 (Xie et al., 2026a)	–	1620.5	79.3	1.00	0.87	0.58	0.92	0.52	0.62	0.76
BLIP3-o-8B (Chen et al., 2025a)	–	1682.6	83.5	1.00	0.92	0.63	0.91	0.86	0.67	0.83
InternVL-U (Tian et al., 2026)	–	1607.5	80.5	0.99	0.94	0.74	0.91	0.77	0.74	0.85
Emu3.5 (Cui et al., 2025)	–	781.1	–	–	–	–	–	–	–	0.86
UniAR (Peng et al., 2026)	–	–	–	1.00	0.95	0.75	0.94	0.77	0.67	0.85
OmniGen2 (Wu et al., 2025c)	–	1584.4	79.1	1.00	0.95	0.64	0.88	0.55	0.76	0.80
BAGEL (Deng et al., 2025)	86.5	1647.9	85.0	1.00	0.89	0.70	0.82	0.47	0.49	0.73
+ YANAMI-400K	87.4↑	1616.9	89.1↑	0.99	0.90↑	0.73↑	0.81	0.52↑	0.55↑	0.75↑

Table 4: **Comparison of different models on visual understanding tasks and GenEval.** An upward arrow indicates an improvement over BAGEL.

evaluation adopts the GT-History setting, errors from previous generations do not propagate to later turns. This comparison therefore more directly reflects how models handle increasingly accumulated context and state dependencies over long-horizon interactions. The results show performance drops of 2%, 6%, and 6% for BAGEL, FLUX.2, and InternVL-U, respectively, suggesting that models still struggle to stably maintain and realize the complete multimodal dialogue state as dialogue length and state complexity increase.

GT-History vs. Self-History. To better reflect realistic multi-turn interaction, we additionally evaluate a *Self-History* setting, in which images generated by the model in previous turns are used as visual context for subsequent turns. To focus on interactions in which generated visual history can affect subsequent outputs, we include only dialogues requiring more than one generation turn. As shown in Fig. 6c, performance under Self-History decreases by 5.16% and 3.77% in BAGEL and Qwen-Edit, respectively, compared with GT-History. This result suggests that, although models can perform reasonably well when conditioned on correct visual histories, error accumulation from self-generated histories can substantially degrade stateful interaction performance during realistic multi-turn rollouts. More results and detailed breakdowns are provided in Appendix D.

4.2 Train on YANAMI-400K

To validate the effectiveness of YANAMI-400K, we fine-tune BAGEL on our dataset, with detailed training configurations provided in Appendix C. The results demonstrate that YANAMI-400K substantially improves image generation, image editing, and multi-turn multimodal dialogue state modeling, while largely preserving the model's general visual understanding capabilities.

Model	BG	Color	Mat.	Motion	Port.	Style	Add	Remove	Replace	Text	Tone	Avg.
AnyEdit (Yu et al., 2025)	4.31	4.25	2.64	0.67	1.90	1.95	3.72	3.75	3.23	0.77	4.21	2.85
InstructPix2Pix (Brooks et al., 2023)	3.94	5.40	3.52	1.27	2.62	4.39	3.07	1.50	3.48	1.13	5.10	3.22
OmniGen (Xiao et al., 2025)	5.23	5.93	5.44	3.12	3.17	4.88	6.33	6.35	5.34	4.31	4.96	5.01
Step1X-Edit (Liu et al., 2025)	7.03	6.26	6.46	3.66	5.23	7.24	7.17	6.42	7.39	7.40	6.62	6.44
Step1X-Edit-v1p2 (Liu et al., 2025)	–	–	–	–	–	–	–	–	–	–	–	7.24
OmniGen2 (Wu et al., 2025c)	6.99	6.66	4.88	2.55	3.66	6.08	7.09	6.60	6.65	4.49	6.03	5.57
EditMGT (Chow et al., 2026a)	7.69	7.71	5.77	3.84	5.13	6.53	6.13	5.24	5.56	4.53	6.42	5.87
GoT-6B (Fang et al., 2025)	4.11	5.75	3.04	1.71	2.69	4.72	5.77	4.59	5.65	1.16	4.24	3.95
VAREdit-8B (Mao et al., 2026)	6.77	6.64	5.40	3.33	4.20	6.46	5.86	7.29	6.67	3.87	6.54	5.73
FLUX.1 Kontext [dev] (Labs et al., 2025)	7.06	7.03	5.52	5.62	4.68	5.55	6.95	6.76	6.13	6.10	7.48	6.26
FLUX.2 [dev] (Labs, 2025)	–	–	–	–	–	–	–	–	–	–	–	7.41
Qwen-Image-Edit (Wu et al., 2025a)	–	–	–	–	–	–	–	–	–	–	–	7.88
Ovis-U1 (Wang et al., 2025a)	7.49	6.88	6.21	4.79	5.98	6.46	7.49	7.25	7.27	4.48	6.31	6.42
BAGEL (Deng et al., 2025)	7.44	6.99	6.26	5.09	4.82	6.04	7.94	7.37	7.31	7.16	6.17	6.52
+ YANAMI-400K	7.58↑	7.76↑	6.42↑	7.06↑	6.11↑	6.20↑	8.44↑	7.74↑	7.55↑	7.11	7.08↑	7.19↑

Table 5: **Comparison of fine-tuned BAGEL and other models on GEdit-EN-full benchmark.** An upward arrow indicates an improvement over BAGEL.

Fine-tuned BAGEL performances on YANAMI-Bench across four interaction scenarios: *(i) Visual Understanding.* The fine-tuned BAGEL largely preserves its original general visual understanding capabilities. As shown in Table 4, the model achieves improvements of 4.1% and 0.9% on MMBench and POPE, respectively, while exhibiting only a slight decrease on MME-P, indicating that training on YANAMI-400K preserves general visual understanding capability. *(ii) Image Generation.* As shown in Table 4, the fine-tuned BAGEL improves the overall score on GenEval by 2%, demonstrating that YANAMI-400K enhances the model's general image generation capability. *(iii) Image Editing.* As shown in Table 5, the fine-tuned BAGEL improves the overall score on GEdit-Bench-EN by 10.28%, further demonstrating the effectiveness of YANAMI-400K in enhancing image editing capability. *(iv) Multimodal Dialogue Stateful Interaction.* As shown in Table 2, the fine-tuned model improves by 13.23% over the original BAGEL on YANAMI-Bench. This result demonstrates that YANAMI-400K not only improves general image generation and editing capabilities, but also substantially enhances the model's ability to understand, maintain, update, and visually realize multimodal dialogue states throughout multi-turn interactions.

4.3 Reliability of Judge Usage

To assess the reliability of our VLM-as-a-Judge evaluation framework, we conduct a human evaluation study involving four expert annotators. Specifically, we sample 100 instances from each of three representative models: BAGEL+YANAMI, Emu3.5, and Nano Banana 2, and collect human ratings for comparison. We compute Pearson correlation coefficients(Benesty et al., 2009) between the scores produced by Qwen-3.8-Flash-Next(Qwen Team, 2026) and the aggregated human ratings, and further compare its evaluations with those produced by GPT-5.6-sol, as shown in Fig. 5.

The results show that the correlations between Qwen-3.8-Flash-Next and human ratings consistently exceed 0.80, indicating strong agreement between the VLM judge and human evaluation. Moreover, the scores produced by Qwen-3.8-Flash-Next and GPT-5.6-sol exhibit high cross-VLM correlation, demonstrating strong consistency across different VLM evaluators. These results suggest that our evaluation framework aligns well with human judgments and is robust to the choice of VLM judge.

5 Conclusion

This work introduces **YANAMI**, the first large-scale dataset and benchmark for multimodal stateful interaction across understanding, generation, and editing. YANAMI-400K provides 400K multi-turn dialogues, more than 2M images, and 2.6M dialogue turns across four scenarios and six types of contextual dependency. YANAMI-Bench comprises 400 multi-turn dialogues and jointly evaluates dialogue-state understanding and visual realization. Experiments show that training on YANAMI-400K improves multi-turn contextual understanding, image generation, and image editing while preserving general visual understanding. Evaluations on YANAMI-Bench further show that existing UMMs still struggle to maintain, update, and visually realize dialogue states. YANAMI provides a unified foundation for robust multimodal stateful interaction.

AI USE STATEMENT

LLM-based tools assisted with language polishing, code development, data filtering, dialogue construction, and quality control. As detailed in Appendix A, GPT-5.6 Sol, Qwen3.7-Plus, and Gemini 3.1 Pro were used in the data-construction pipeline, while GPT-Image-2 and Nano Banana 2 supported image generation and repair; Qwen-3.8-Flash-Next was additionally used for automated evaluation under the protocols in Appendix B. All benchmark samples and annotations were manually verified; the authors reviewed AI-assisted writing and code, validated the resulting procedures, and take responsibility for the final content and claims of this work.

ETHICS STATEMENT

YANAMI is intended for research on multimodal stateful interaction and is constructed from existing public datasets and model-generated content. We follow the applicable licenses and usage conditions of the source datasets and models. Automated filtering and human verification are applied to improve data quality. We encourage the research community to use and further develop these resources responsibly.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce our data construction, training, and evaluation procedures throughout the paper and appendix. The construction pipeline, quality-control process, benchmark annotation procedure, and detailed data statistics are documented in Appendices A and A.6. The evaluation protocols, score calculation, history and input formatting, and model-specific inference configurations are described in Appendix B, while the fine-tuning configuration is reported in Appendix C. Complete experimental results and the verbatim VLM-as-a-Judge prompts are provided in Appendices D and F, respectively. To further support reproducibility, we will release YANAMI-400K and YANAMI-Bench on Hugging Face and publish the prompts used for data construction and evaluation in a public GitHub repository.

REFERENCES

Alibaba Cloud. Qwen3.7-Plus. `https://www.alibabacloud.com/help/en/model-studio/qwen3-7-plus`, 2026. Accessed: 2026-09-16.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pp. 1–4. Springer, 2009.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402. IEEE, 2023.

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*, 2025a.

Tianyu Chen, Yasi Zhang, Zhi Zhang, Peiyu Yu, Shu Wang, Zhendong Wang, Kevin Lin, Xiaofei Wang, Zhengyuan Yang, Linjie Li, et al. Edival-agent: An object-centric framework for automated, fine-grained evaluation of multi-turn editing. *arXiv preprint arXiv:2509.13399*, 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8073–8082. IEEE, 2025c.

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. *arXiv preprint arXiv:2501.17811*, 2025d.

Zhekai Chen, Yuqing Wang, Manyuan Zhang, and Xihui Liu. Macro: Advancing multi-reference image generation with structured long-context data. *arXiv preprint arXiv:2603.25319*, 2026.

Wei Chow, Linfeng Li, Lingdong Kong, Zefeng Li, Qi Xu, Hang Song, Tian Ye, Xian Wang, Jinbin Bai, Shilin Xu, et al. Editmgt: Unleashing potentials of masked generative transformers in image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38038–38048, 2026a.

Wei Chow, Jiachun Pan, Yongyuan Liang, Mingze Zhou, Xue Song, Liyu Jia, Saining Zhang, Siliang Tang, Juncheng Li, Fengda Zhang, et al. Weave: Unleashing and benchmarking the in-context interleaved comprehension and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15343–15353, 2026b.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

Wenxun Dai, Zhiyuan Zhao, Yule Zhong, Yiji Cheng, Jianwei Zhang, Linqing Wang, Shiyi Zhang, Yunlong Lin, Runze He, Fellix Song, et al. Chatumm: Robust context tracking for conversational interleaved generation. *arXiv preprint arXiv:2602.06442*, 2026.

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.

Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, et al. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *arXiv preprint arXiv:2503.10639*, 2025.

Yukang Feng, Jianwen Sun, Chuanhao Li, Zizhen Li, Jiaxin Ai, Fanrui Zhang, Sizhuo Zhou, Yifan Chang, Shenglin Zhang, Yu Dai, et al. A high quality dataset and reliable evaluation for interleaved image-text generation. In *International Conference on Learning Representations*, volume 2026, pp. 40398–40437, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

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.

Google. Gemini 3.1 Pro Preview. `https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview`, 2026a. Accessed: 2026-09-16.

Google. Gemini 3.1 Flash Image (Nano Banana 2). `https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-image`, 2026b. Accessed: 2026-09-16.

Minbin Huang, Yanxin Long, Xinchi Deng, Ruihang Chu, Jiangfeng Xiong, Xiaodan Liang, Hong Cheng, Qinglin Lu, and Wei Liu. Dialoggen: multi-modal interactive dialogue system with multi-turn text-image generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 411–426, 2025.

Black Forest Labs. Flux.2: Frontier visual intelligence. *https://bfl.ai/blog/flux-2*, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Ang Li, Charles Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, et al. Zebra-cot: A dataset for interleaved vision language reasoning. *arXiv preprint arXiv:2507.16746*, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

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

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Shuo Liu, Kaining Ying, Hao Zhang, Yue Yang, Yuqi Lin, Tianle Zhang, Chuanhao Li, Yu Qiao, Ping Luo, Wenqi Shao, et al. Convbench: A multi-turn conversation evaluation benchmark with hierarchical capability for large vision-language models. *arXiv preprint arXiv:2403.20194*, 2024a.

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, 2024b.

Qingyang Mao, Qi Cai, Yehao Li, Yingwei Pan, Mingyue Cheng, Ting Yao, Qi Liu, and Tao Mei. Visual autoregressive modeling for instruction-guided image editing. In *International Conference on Learning Representations*, volume 2026, pp. 85702–85723, 2026.

OpenAI. GPT-5.6 Sol. `https://developers.openai.com/api/docs/models/gpt-5.6-sol`, 2026a. Accessed: 2026-09-16.

OpenAI. GPT-Image-2. `https://developers.openai.com/api/docs/models/gpt-image-2`, 2026b. Accessed: 2026-09-16.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

Wujian Peng, Lingchen Meng, Yuxuan Cai, Xianwei Zhuang, Yuhuan Yang, Rongyao Fang, Chenfei Wu, Junyang Lin, Zuxuan Wu, and Shuai Bai. Unified multimodal autoregressive modeling with shared context-visual tokenizer is key to unification. *arXiv preprint arXiv:2606.18249*, 2026.

Yusu Qian, Eli Bocek-Rivele, Liangchen Song, Jialing Tong, Yinfei Yang, Jiasen Lu, Wenze Hu, and Zhe Gan. Pico-banana-400k: A large-scale dataset for text-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37226–37235, 2026.

Qwen Team. Qwen3.8-Flash-Next: A new architecture, towards ultimate cost-efficiency, August 2026. URL `https://qwen.ai/blog?id=qwen3.8-flash-next`.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

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

Changyao Tian, Danni Yang, Guanzhou Chen, Erfei Cui, Zhaokai Wang, Yuchen Duan, Penghao Yin, Sitao Chen, Ganlin Yang, Mingxin Liu, et al. Internvl-u: Democratizing unified multimodal models for understanding, reasoning, generation and editing. *arXiv preprint arXiv:2603.09877*, 2026.

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*, 2025a.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Yuhan Wang, Siwei Yang, Bingchen Zhao, Letian Zhang, Qing Liu, Yuyin Zhou, and Cihang Xie. Gpt-image-edit-1.5 m: A million-scale, gpt-generated image dataset. *arXiv preprint arXiv:2507.21033*, 2025b.

Hongyang Wei, Hongbo Liu, Zidong Wang, Yi Peng, Baixin Xu, Size Wu, Xuying Zhang, Xianglong He, Zexiang Liu, Peiyu Wang, et al. Skywork unipic 3.0: Unified multi-image composition via sequence modeling. *arXiv preprint arXiv:2601.15664*, 2026a.

Xinyu Wei, Kangrui Cen, Hongyang Wei, Zhen Guo, Bairui Li, Zeqing Wang, Jinrui Zhang, and Lei Zhang. Mico-150k: A comprehensive dataset advancing multi-image composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29695–29706, 2026b.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12966–12977. IEEE, 2025b.

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 e-prints*, pp. arXiv–2506, 2025c.

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. In *International Conference on Learning Representations*, volume 2025, pp. 25842–25875, 2025.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13294–13304. IEEE, 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. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *Advances in Neural Information Processing Systems*, 38:47490–47518, 2026a.

Wulin Xie, YiFan Zhang, Chaoyou Fu, Yang Shi, Jianshu Zeng, Bingyan Nie, Hongkai Chen, Zhang Zhang, and Liang Wang. Mme-unify: A comprehensive benchmark for unified multimodal understanding and generation models. In *International Conference on Learning Representations*, volume 2026, pp. 133535–133578, 2026b.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26125–26135. IEEE, 2025.

APPENDIX

A Data Construction Details **15**

A.1 Comparison with Existing Multimodal Datasets 15

A.2 Data Preparation and Composition 15

A.3 Data Construction 16

A.4 Automatic Verification 17

A.5 Human Verification and Benchmark Construction 17

A.6 Detailed Data Statistics 18

B Evaluation Protocol and Metrics **18**

B.1 Dialogue-State Understanding Evaluation 18

B.2 Visual Quality and State Consistency Evaluation 20

B.3 Score Calculation 20

B.4 History and Input Formatting 20

B.5 Baseline Inference Configuration 21

B.6 Understanding–Realization Gap Evaluation 21

C Training Details **22**

D Complete Experimental Results **23**

D.1 Aggregate Results on YANAMI-Bench 23

D.2 Dialogue-State Understanding Results 23

D.3 Results across Dialogue Progression 23

D.4 Understanding–Realization Gap 24

E Additional Qualitative Examples **28**

F VLM-as-a-Judge Prompt Templates **33**

F.1 Dialogue-State Understanding Prompt 33

F.2 State Consistency Prompt 33

F.3 Visual Quality Prompt 35

G Limitations **36**

Dataset	Scale	Domain	Open Data	Multi-turn Format	Hybrid Evaluation	Explicit Dialogue-State Management	Irreducible to Single-Turn
InstructPix2Pix (Brooks et al., 2023)	∼450K editing pairs	E	Yes	No	No	No	No
AnyEdit (Yu et al., 2025)	2.5M editing pairs	E	Yes	No	No	No	No
CoMM (Chen et al., 2025c)	227K documents	U/G	Yes	No	No	No	No
Zebra-CoT (Li et al., 2025)	182,384 samples	U/G	Yes	No	No	No	No
Pico-Banana-400K (Qian et al., 2026)	400K editing pairs; 72K multi-turn	E	Yes	Partial	No	No	No
WEAVE-100K (Chow et al., 2026b)	100,750 dialogues	U/G/E	Yes	Yes	Yes	No	No
ChatUMM (Dai et al., 2026)	Not reported[‡]	U/G/E	No[§]	Yes	No	No	Partial
YANAMI-400K (Ours)	**402,639 dialogues**	**U/G/E**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Table 6: **Comparison with representative multimodal datasets.** U, G, and E denote visual understanding, image generation, and image editing, respectively. Multi-turn Format indicates that data are organized as multi-turn user–assistant interactions; Hybrid Evaluation indicates a unified protocol combining complementary evaluation signals rather than merely reporting multiple independent metrics. Explicit Dialogue-State Management is marked Yes only when a dataset explicitly supervises the establishment, maintenance, invalidation, or update of the currently valid multimodal state; generic history conditioning, context tracking, or a claim of "stateful dialogue" alone does not qualify. Irreducible to Single-Turn indicates that target turns cannot be reformulated as conventional single-turn tasks simply by directly supplying the relevant historical images. Partial denotes implicit or subset-level coverage. Scale follows the unit reported by each work and is therefore not directly comparable across all rows. [‡] ChatUMM reports approximately 40B consumed training tokens but does not report the number of dialogues, images, or samples. [§] No official ChatUMM training-data release was publicly available as of 2026-09-24; its categorization is therefore based on the construction procedure and examples reported in the paper. Although ChatUMM describes its interactions as stateful, its published pipeline primarily converts standard single-turn tasks into dialogues through historical dependencies and distractor turns, without explicit dialogue-state supervision. Once the relevant historical input is supplied directly, the reported target turns can generally be reformulated as conventional single-turn generation, editing, or composition tasks.

A Data Construction Details

A.1 Comparison with Existing Multimodal Datasets

Table 6 compares YANAMI-400K with representative multimodal understanding, generation, editing, and dialogue datasets. The comparison separates multi-turn formatting from explicit dialogue-state management: a dataset may be organized as a sequence of turns while still allowing individual target turns to be reformulated as conventional single-turn tasks once the relevant historical content is directly supplied.

A.2 Data Preparation and Composition

We draw raw images from four existing public datasets (Li et al., 2025; Chow et al., 2026b; Wei et al., 2026b; Chen et al., 2026) and process them in the Data Preparation stage shown in Fig. 3.a. Our pipeline leverages GPT-5.6 Sol (OpenAI, 2026a), Qwen3.7-Plus (Alibaba Cloud, 2026), Gemini 3.1 Pro (Google, 2026a), GPT-Image-2 (OpenAI, 2026b), and Nano Banana 2 (Google, 2026b) for data filtering, dialogue construction, and quality control. The raw images are first filtered for quality and suitability and then organized into four scenario-specific pools for Multi-Image Fusion, Image Editing, World Modeling, and Visual Reasoning. For this initial quality-control stage, we use large multimodal models with the task-specific prompt reproduced in Appendix A.2.1. The filter deliberately performs a permissive coarse screening: an image is marked as rejected only when severe corruption, generation failure, invalid content, or extremely poor visibility makes it clearly unusable. We apply the filter independently to every image and record the resulting quality labels at the image level. For each image group, if more than half of its images are marked as rejected, we discard the entire group; otherwise, we repair only the rejected images and send the repaired versions through the same filtering process again. Samples that remain invalid after this process are discarded. This procedure supplies diverse visual materials for constructing multi-turn multimodal stateful dialogues involving understanding, generation, and editing, without imposing a one-to-one correspondence between a source dataset and an interaction scenario.

A.2.1 IMAGE QUALITY FILTERING PROMPT

Prompt for Initial Image Quality Filtering

```
You are an image quality filter responsible for performing a **coarse initial screening**
↪  of images.

Your goal is NOT to evaluate aesthetics, fine details, instruction following, or
↪  professional image quality.
You should only reject images that are **clearly unusable or extremely low quality**.

Evaluate the input image and determine whether it should be **KEPT** or **REJECTED**.

### REJECT the image only if one or more of the following severe problems occur:

1. **Unrecognizable content**: the image is so corrupted, blurred, distorted, or noisy
↪  that its main content cannot be reasonably identified.
2. **Severe generation failure**: the image contains large-scale visual corruption,
↪  completely broken structures, extreme artifacts, or incoherent content that makes the
↪  image unusable.
3. **Nearly empty or invalid image**: the image is blank, almost entirely a single color,
↪  severely cropped to the point that meaningful content is absent, or otherwise appears
↪  invalid.
4. **Extremely poor visibility**: the image is so dark, overexposed, low-resolution, or
↪  degraded that the main subject or scene cannot be meaningfully perceived.

### KEEP the image if:

- The main subject and scene are recognizable.
- The image contains understandable visual content.
- Imperfections are minor or moderate.

Do NOT reject an image merely because of:

- minor blur or low resolution;
- imperfect anatomy or object geometry;
- small artifacts;
- unrealistic details;
- mediocre aesthetics;
- imperfect composition, lighting, or colors;
- minor text rendering errors;
- moderate generation artifacts.

**Be permissive. When uncertain, KEEP the image.**

Return only the following JSON:
```json
{
  "decision": "KEEP" or "REJECT",
  "reason": "brief reason"
}
```
```

The resulting YANAMI-400K contains 402,639 dialogues. As shown in Fig. 4, Multi-Image Fusion, Image Editing, World Modeling, and Visual Reasoning account for 36%, 25%, 28%, and 11% of the training dialogues, respectively. We retain this naturally varying training distribution while constructing a balanced benchmark with 100 dialogues from each scenario, enabling controlled comparison across scenarios. Complete corpus- and scenario-level statistics are provided in Section A.6.

A.3 DATA CONSTRUCTION

After data preparation, each scenario-specific pool is combined with scenario- and capability-oriented prompts to construct multi-turn multimodal stateful dialogues, as illustrated in Fig. 3.b. The complete data-construction prompts are provided in the supplementary material. Within each dialogue, visual observations, textual instructions, and generated or edited images jointly determine how the multimodal dialogue state evolves. The four scenarios instantiate distinct forms of this evolution.

1. *Multi-Image Fusion* : the assistant integrates multiple user-provided or previously generated images according to the dialogue history.
2. *Image Editing* : the assistant progressively edits an image while preserving prior modifications and visual consistency.

3. *World Modeling* : the assistant predicts and renders scene states under specified viewpoint changes or actions.
4. *Visual Reasoning* : the assistant generates visual outcomes for proposed hypotheses or solutions to support step-by-step reasoning.

To make these interactions genuinely stateful, we incorporate six key state-management capabilities across the four scenarios: *Coreference Resolution*, *Dialogue-History Retrieval*, *Cross-Turn Consistency*, *Spatial and Physical Consistency*, *Constraint Adherence*, and *Correction-Based State Updating*.

1. *Coreference Resolution* grounds context-dependent expressions to the correct textual or visual entities.
2. *Dialogue-History Retrieval* identifies and uses information relevant to the current request from previous turns.
3. *Cross-Turn Consistency* preserves established visual attributes and prior edits across turns unless the user explicitly changes them.
4. *Spatial and Physical Consistency* maintains plausible geometry, physical relations, and state transitions.
5. *Constraint Adherence* consistently follows user-specified requirements and restrictions.
6. *Correction-Based State Updating* incorporates user corrections and replaces outdated information accordingly.

These capabilities cut across the four scenarios rather than forming a fixed one-to-one mapping: multiple capabilities jointly construct and maintain the evolving state within a dialogue, and the same capability can occur in different scenarios. Every dialogue in YANAMI-400K involves at least four of the six capabilities, requiring multiple cross-turn dependencies to be exercised jointly within the interaction.

A.4 Automatic Verification

We apply quality control both before and after data construction, following the pipeline in Fig. 3. During data preparation, model-assisted filtering removes unsuitable visual materials and routes potentially repairable samples back through the preparation process. After data construction, each dialogue is automatically verified for logical coherence and image–text consistency: the former examines the instructions and state transitions across turns, while the latter checks the visual content against its associated textual context. Failed samples are iteratively corrected or regenerated and then re-verified, while dialogues that pass verification form YANAMI-400K or enter the benchmark-candidate pool. This iterative process is intended to eliminate inconsistent state transitions, mismatched image–text pairs, and dialogues that do not realize the intended scenario or contextual dependencies.

A.5 Human Verification and Benchmark Construction

We construct YANAMI-Bench with LLM assistance from candidates that pass automatic verification, followed by exhaustive human verification and correction. The benchmark is balanced across the four interaction scenarios, with 100 dialogues per scenario and 400 dialogues in total. *(i) Dialogue-state understanding QA pairs.* Each benchmark dialogue includes four QA pairs, one each for Reference Resolution, Context Retrieval, Constraint Tracking, and Correction Updating. *(ii) Fine-grained atomic criteria.* Each dialogue also includes atomic criteria for evaluating State Consistency in generated or edited visual outputs. These atomic criteria decompose the requirements of a target image into independently verifiable checks, allowing State Consistency to be assessed separately from holistic Visual Quality.

As shown in the human-verification stage of Fig. 3, the initial atomic checks are generated with LLM assistance and then reviewed together with the corresponding dialogue and visual content. Three human annotators with graduate-level training in STEM disciplines verify each sample. They inspect the dialogue logic, visual evidence, understanding questions, and atomic evaluation criteria, and correct samples that do not faithfully represent the intended state evolution or evaluation

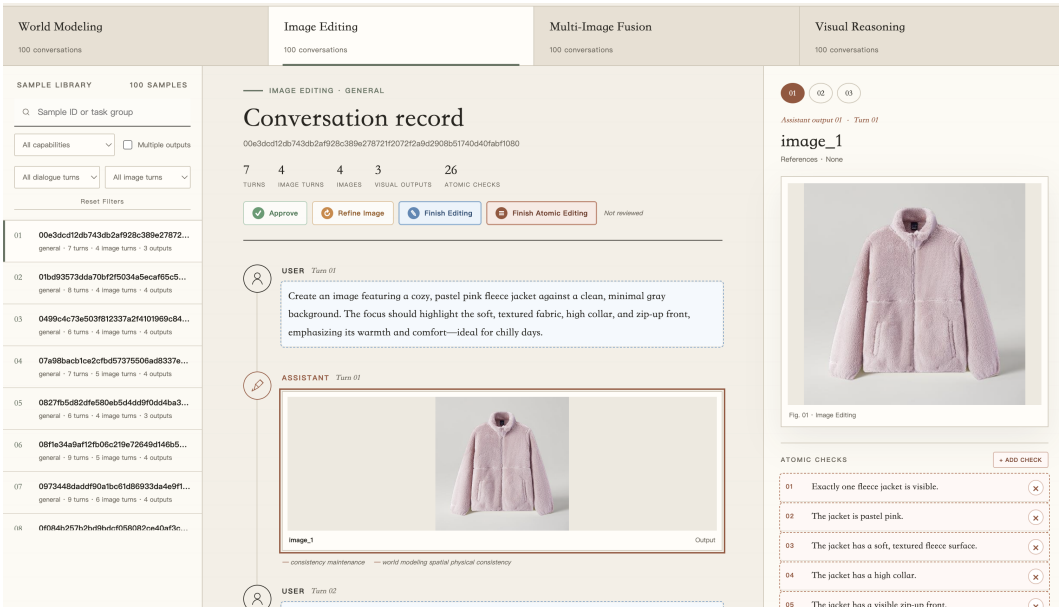


Figure 7: **Human-verification interface for YANAMI-Bench.** Annotators can browse the 100 benchmark dialogues in each interaction scenario, inspect the complete multi-turn context and visual outputs, and review or revise the corresponding atomic evaluation criteria. The interface also records whether a sample is approved or requires image refinement, supporting the exhaustive verification and correction process used to construct the final benchmark.

target. Failed samples undergo manual correction or regeneration followed by another review, and every sample retained in the final benchmark is manually verified. The final benchmark contains 1,600 understanding questions and 8,467 atomic visual checks, with detailed per-scenario statistics reported in Table 8.

Figure 7 shows the interface used to inspect complete dialogues, generated visual outputs, and their associated atomic checks during benchmark verification.

A.6 Detailed Data Statistics

YANAMI-400K contains 402,639 multi-turn dialogues, 2,047,236 images, and 2,643,291 user–assistant dialogue turns across the four interaction scenarios. Among these dialogues, 321,868 contain at least four images, 215,027 contain at least five images, and 138,669 contain at least six images. The dataset contains an average of 5.08 images and 6.56 dialogue turns per dialogue, with a maximum of 43 dialogue turns. Table 7 summarizes these statistics, while Fig. 4 presents the distribution across the four interaction scenarios.

YANAMI-Bench contains 400 dialogues evenly distributed across the four scenarios, together with 1,883 images, 3,173 user–assistant dialogue turns, 1,600 dialogue-state understanding questions, and 8,467 atomic visual checks used in the benchmark evaluation. Each dialogue contains one question for each of the four understanding capabilities, yielding 400 questions per capability. Table 8 reports the corresponding scenario-level statistics.

B Evaluation Protocol and Metrics

B.1 Dialogue-State Understanding Evaluation

Each benchmark dialogue contains four diagnostic questions, one for each explicitly testable understanding capability: Reference Resolution, Context Retrieval, Constraint Tracking, and Correction Updating. The evaluator receives the dialogue context, the diagnostic question, a reference answer, and the model prediction, and judges semantic correctness rather than exact lexical overlap. A

Corpus and Sequence Statistics		Dataset Distributions		
Statistic	**Value**	**Images per Dialogue**	**Dialogues**	**Share**
Total Dialogues	402,639	< 4	80,771	20.06%
Total Images	2,047,236	$= 4$	106,841	26.54%
Total Dialogue Turns	2,643,291	$= 5$	76,358	18.96%
Total User Messages	2,643,291	≥ 4	321,868	79.94%
Total Assistant Messages	2,643,291	≥ 5	215,027	53.40%
Total Messages	5,286,582	≥ 6	138,669	34.44%
		Training Scenario Distribution		
Average Images / Dialogue	5.08	Multi-Image Fusion	–	36%
Average Turns / Dialogue	6.56	Image Editing	–	25%
Average Messages / Dialogue	13.13	World Modeling	–	28%
Maximum Turns / Dialogue	43	Visual Reasoning	–	11%
Maximum Messages / Dialogue	86	Interaction Scenarios	4	–
Images / Dialogue Turn	0.77	State-Management Capabilities	6	–
Dialogues with ≥ 4 Capabilities	402,639	Capability Coverage	–	100%

Table 7: **Detailed statistics of YANAMI-400K.** Each dialogue turn contains one user message and one assistant response; Total Messages therefore counts both sides of the interaction. Scenario shares are rounded percentages corresponding to the training-set distribution shown in Fig. 4.

Scenario	Dialogues	Images	Dialogue Turns	Understanding QAs	Atomic Checks	Avg. Checks
Multi-Image Fusion	100	400	777	400	1,890	18.90
Image Editing	100	499	754	400	3,408	34.08
World Modeling	100	467	598	400	1,141	11.41
Visual Reasoning	100	517	1,044	400	2,028	20.28
Total	**400**	**1,883**	**3,173**	**1,600**	**8,467**	**21.17**

Table 8: **Detailed statistics of YANAMI-Bench by interaction scenario.** Images include all reference and target images contained in the benchmark dialogues. Understanding QAs comprise Reference Resolution, Context Retrieval, Constraint Tracking, and Correction Updating questions, with one question of each type per dialogue. Atomic Checks counts the independently verifiable visual criteria used in the benchmark evaluation and excludes capability-specific checks; Avg. Checks reports the average number of evaluated atomic checks per dialogue.

prediction receives credit only when it contains all information required by the reference answer without a material factual error or contradiction. Equivalent wording and correct additional detail are accepted, whereas partially correct, unsupported, or materially incomplete answers are marked incorrect. The score for each capability is the proportion of correctly answered questions, and the overall dialogue-state understanding score is the arithmetic mean defined in Appendix B.3.

B.2 Visual Quality and State Consistency Evaluation

State Consistency decomposes the expected target state into independently verifiable atomic visual checks. For each generated image, the evaluator compares the candidate with the ground-truth target and assigns a binary outcome to every atomic check. A check receives credit only when all of its required conditions are visibly satisfied; semantically equivalent visual realizations are accepted when they do not alter the evaluated state property. The State Consistency score is the fraction of atomic checks that receive credit. Capability-specific checks are evaluated separately and are used only for analyses that require alignment between an understanding item and its corresponding generated outcome.

Visual Quality assesses the generated image independently of task correctness on a 0–10 integer scale. The rubric considers structural coherence, visual fidelity, generation artifacts, and overall naturalness, with higher scores indicating better perceptual quality. The complete prompts for all three evaluators are reproduced in Appendix F.

B.3 Score Calculation

For each interaction scenario, we first compute the dialogue-state understanding score as the arithmetic mean of the four understanding metrics: Reference Resolution (RR), Context Retrieval (CR), Constraint Tracking (CT), and Correction Updating (CU). Specifically, the understanding score is calculated as

$$S_{\text{understanding}} = \frac{S_{\text{RR}} + S_{\text{CR}} + S_{\text{CT}} + S_{\text{CU}}}{4}. \tag{1}$$

Before aggregation, all component scores are converted to a 0–100 scale: the State Consistency (SC) score, originally expressed as a fraction, is multiplied by 100, and the Visual Quality (VQ) score, originally measured on a 0–10 scale, is multiplied by 10. For models with both understanding and generation capabilities, the final score for each scenario is computed using a $4{:}4{:}2$ weighting over dialogue-state understanding, State Consistency, and Visual Quality:

$$S_{\text{scene}} = 0.4S_{\text{understanding}} + 0.4S_{\text{SC}} + 0.2S_{\text{VQ}}. \tag{2}$$

For models without dialogue-state understanding capabilities, the scenario score is computed from State Consistency and Visual Quality using a $6{:}4$ weighting:

$$S_{\text{scene}} = 0.6S_{\text{SC}} + 0.4S_{\text{VQ}}. \tag{3}$$

This weighting ensures that state faithfulness remains the primary component of the final score while retaining an explicit assessment of perceptual visual quality.

B.4 History and Input Formatting

Our main evaluation adopts the GT-History setting to measure the model's ability to use accumulated dialogue state without conflating it with errors produced in earlier generation turns. When evaluating a target generation or editing turn, we retain the complete textual dialogue preceding that turn and replace all earlier assistant-generated images with their corresponding ground-truth images. User-provided reference images remain unchanged. The target output itself is never included in the input context.

We additionally evaluate Self-History to approximate realistic multi-turn deployment. Under this setting, each image produced by the evaluated model is inserted into the dialogue history and used as visual context for subsequent turns, while the textual history remains unchanged. User-provided reference images again remain unchanged. Consequently, errors in earlier visual outputs may alter the context available at later turns and accumulate throughout the interaction.

We use three input organizations according to the interfaces supported by individual models. For interleaved image–text input, denoted as *yes-first*, each image is inserted at the position of its first mention in the dialogue, preserving the original temporal order of the interaction. For image-first input, denoted as *yes-front*, all visual inputs available at the target turn are placed before the textual dialogue in their order of appearance. For models that cannot process sequential multi-image inputs, the available images are concatenated in dialogue order into a single composite visual input and supplied together with the corresponding textual instruction. The input organization used for each model is reported in Table 2.

B.5 Baseline Inference Configuration

All models are evaluated on the same set of dialogues and target generation or editing turns in YANAMI-Bench. For each target turn, the input contains the preceding multi-turn dialogue and all reference images available under the selected history setting described above. Models with an understanding interface are additionally evaluated on the 1,600 dialogue-state understanding questions, while generation- or editing-only models are marked as not applicable for this component. We request an aspect-ratio-preserving output frame whose long edge is 512 pixels, with both dimensions rounded to multiples of 16; when a native model interface imposes a fixed or internally selected resolution, we retain its native output for evaluation. We use a global seed of 42 and derive a deterministic seed for each sample from its identifier. Text responses are decoded greedily with at most 512 new tokens, unless the model's official interface requires a different decoding procedure. For closed-source systems, implementation-specific decoding parameters are unavailable, so we use their provided interfaces with the same target-turn context and output requirements. Dialogue-state understanding, State Consistency, and Visual Quality are evaluated with Qwen-3.8-Flash-Next using deterministic decoding (temperature $= 0$) with explicit reasoning disabled and the three prompts reproduced in Appendix F. Table 9 summarizes the model-specific input handling and the principal generation parameters used in our evaluation.

B.6 Understanding–Realization Gap Evaluation

We measure the understanding–realization gap by conditioning visual-realization evaluation on correct dialogue-state understanding. For each dialogue and evaluated state property, such as a persistent user constraint or a correction-based state update, we first identify the corresponding understanding question and its aligned generation-side capability checks. To ensure a fair model comparison, we use a shared evaluation subset rather than selecting correctly understood items separately for each model. Specifically, a dialogue is retained for capability c only if every evaluated model receives full credit on its corresponding understanding question and has at least one aligned generation-side check. This prevents a model with weaker understanding from being evaluated on an easier or smaller model-specific subset and ensures that all reported gaps are computed on exactly the same dialogue–capability pairs.

Let $\mathcal{M}$ denote the set of evaluated models, let $u_{m,d,c} \in [0,1]$ be the understanding score of model m for capability c in dialogue d, and let $\mathcal{A}_{m,d,c}$ be its aligned generation-side checks. For each capability, the shared eligible dialogue set is

$$\mathcal{E}_c = \{d \,|\, u_{m,d,c} = 1 \ \wedge \ |\mathcal{A}_{m,d,c}| > 0, \ \forall m \in \mathcal{M}\}. \tag{4}$$

Each aligned check $a \in \mathcal{A}_{m,d,c}$ receives a binary score $y_{m,d,c,a}$, where a correctly realized state property receives 1 and an incorrectly realized property receives 0. For example, when the model receives full credit for understanding an active user constraint, each corresponding constraint-keeping check in the generated image contributes 1 if satisfied and 0 otherwise. The visual-realization score for a retained dialogue–capability pair is

$$R_{m,d,c} = \frac{1}{|\mathcal{A}_{m,d,c}|} \sum_{a \in \mathcal{A}_{m,d,c}} y_{m,d,c,a}. \tag{5}$$

Because the atomic scores are binary, $R_{m,d,c}$ is naturally normalized to $[0,1]$. We first average within each dialogue–capability pair and then across the shared eligible dialogues so that dialogues containing more checks do not receive disproportionate weight:

$$R_{m,c} = \frac{1}{|\mathcal{E}_c|} \sum_{d \in \mathcal{E}_c} R_{m,d,c}. \tag{6}$$

Model	Generation mechanism	Visual-input handling	Principal inference settings
BAGEL (Deng et al., 2025) and +YANAMI-400K	Unified MoT for understanding and generation	Native interleaved image–text sequence; no image concatenation	50 steps; text/image CFG $4.0/1.5$; parallel CFG; timestep shift 3.0
Emu3.5 (Cui et al., 2025)	Unified autoregressive generation with discrete visual tokens	Native interleaved image–text sequence; no image concatenation	Image/text guidance $3.0/1.0$; image top-k 10,240; top-p 1.0
OmniGen2 (Wu et al., 2025c)	Unified diffusion pipeline	Separate visual inputs; references are concatenated into a grid only when more than three are present	50 steps; text guidance 4.0; resolution alignment disabled
Ovis-U1 (Wang et al., 2025a)	Unified MLLM with a diffusion generation head	Ordered image placeholders; multiple references are concatenated into a grid	50 steps; text/image CFG $5.0/1.5$
InternVL-U (Tian et al., 2026)	Unified VLM with diffusion-based generation	Multiple references are concatenated into a grid	50 steps; deterministic per-sample seed
Janus-Pro (Chen et al., 2025d)	Autoregressive generation with discrete visual tokens	Interleaved inputs for understanding; generation uses the flattened dialogue text without reference-image conditioning	576 visual tokens per image; CFG 5.0; greedy decoding
Show-o2 (Xie et al., 2026a)	Continuous-token flow-matching generation	Reference-image latents are concatenated before the target latent	50 Euler steps; guidance 7.5; time-shifting factor 3.0
UniAR (Peng et al., 2026)	Unified autoregressive model with a visual decoder	Native interleaving for understanding; multiple generation references are concatenated into a grid	CFG 1.5; greedy text decoding
OmniGen (Xiao et al., 2025)	Diffusion-based image generation and editing	Ordered image placeholders; no explicit reference-image limit	Text/image guidance $2.5/1.6$
FLUX.2 [dev] (Labs, 2025)	Rectified-flow diffusion transformer	Up to ten separate references; larger sets are concatenated into a grid	50 steps; guidance 4.0
Qwen-Edit (Wu et al., 2025a)	Multi-image-conditioned diffusion transformer	Native multi-image conditioning without concatenation	40 steps; true CFG 4.0; guidance 1.0
Step1X-v1p2 (Liu et al., 2025)	Diffusion-transformer image editing	Multiple references are concatenated into a grid	50 steps; true CFG 6.0; reflection and thinking modes disabled
UniPic 3.0 (Wei et al., 2026a)	Multi-image composition and editing	Multiple references are concatenated into a grid	50 steps; true CFG 4.0; guidance 1.0

Table 9: **Baseline inference configurations used on YANAMI-Bench.** Only settings that materially affect input organization or output generation are reported; all models otherwise follow the shared evaluation protocol described in the text.

We apply this procedure to the aligned Constraint Tracking and Correction Updating understanding items and their corresponding Constraint Keeping and Correction-Based State Updating checks on the generation side. Across the nine UMMs included in this analysis, all models correctly answer 78 Constraint Tracking items and 71 Correction Updating items; after requiring aligned generation checks for every model, the shared subsets contain 77 and 66 dialogues, respectively. The capability-specific understanding–realization gap is defined as

$$G_{m,c} = 1 - R_{m,c}, \tag{7}$$

where a larger value indicates a greater failure to visually realize state properties that the model has correctly understood. The overall gap is the macro-average of the two capability-specific gaps.

C Training Details

We fine-tune BAGEL-7B-MoT on YANAMI-400K using NVIDIA A800 and A100 GPUs, each with 80 GB of memory. Our compute resources comprise a six-GPU A800 configuration and a 16-GPU A100 configuration; the provided launch configuration uses six processes on a single node. We employ fully sharded data parallelism (FSDP) under the `FULL_SHARD` strategy. The model uses `Qwen2MoTDecoderLayer` and is initialized from the Hugging Face checkpoint and its exponential moving average weights, with only model parameters restored. We jointly optimize visual generation and visual understanding objectives. The VAE is frozen, whereas the language model, vision encoder, and understanding module remain trainable. We use a constant learning rate of 1×10^{-5} with 100 warm-up steps, a cross-entropy loss weight of 0.25, an MSE loss weight of 1.0, gradient accumulation of 1, and a maximum gradient norm of 1.0. The expected token count is 12,288, while both the overall and per-sample maximum token counts are 16,384. The maximum

Model and System Configuration		Optimization and Sequence Configuration	
Setting	**Value**	**Setting**	**Value**
Base model	BAGEL-7B-MoT	Learning rate	1×10^{-5}
Decoder layer	`Qwen2MoTDecoderLayer`	Scheduler	Constant
Compute resources	6× NVIDIA A800 80GB; 16× NVIDIA A100 80GB	Warm-up steps	100
Launch configuration	1 node, 6 processes	CE / MSE loss weights	0.25 / 1.0
Sharding strategy	FSDP `FULL_SHARD`	Gradient accumulation steps	1
Trainable modules	LLM, ViT, understanding module	Maximum gradient norm	1.0
Frozen module	VAE	Expected token count	12,288
Training objectives	Visual understanding and generation	Maximum token count	16,384
Attention implementation	FlexAttention	Maximum tokens per sample	16,384
Data-loader configuration	1 worker per process; prefetch factor 2	Maximum latent size	64

Table 10: **Training configuration.** Settings used to fine-tune BAGEL-7B-MoT on YANAMI-400K, including the model and system setup, optimization hyperparameters, and sequence-length settings.

latent size is 64, and FlexAttention is enabled. Data loading uses one worker per process and a prefetch factor of 2.

D Complete Experimental Results

D.1 Aggregate Results on YANAMI-Bench

Complete aggregate results for all evaluated models and interaction scenarios are reported in Table 2. Scores are computed from the available valid evaluation records; missing records are not assigned zero scores. For GPT-Image-2, scores are computed over 948 valid outputs; four requests rejected by the upstream content policy are excluded. For Nano Banana 2, scores are computed over 939 valid outputs.

D.2 Dialogue-State Understanding Results

For each interaction scenario, we report two complementary types of tables. The first reports the four capability-level dialogue-state understanding scores and their arithmetic mean. The second reports the average understanding score alongside the generation-side State Consistency and Visual Quality scores, followed by the weighted overall average. Models that only support image generation or editing do not produce answers for the understanding evaluation and are therefore marked with dashes in the corresponding columns. Table 11 summarizes dialogue-state understanding performance across the four interaction scenarios.

D.3 Results across Dialogue Progression

Table 20 reports model performance at the first and last image-generation turns for dialogues containing at least two generation turns. For each endpoint, we combine State Consistency and Visual Quality using the same $6{:}4$ weighting adopted for generation-only evaluation.

Model	Fusion	Editing	Modeling	Reasoning	Avg.
OmniGen2 (Wu et al., 2025c)	46.75	50.00	65.00	42.75	51.13
Janus-Pro (Chen et al., 2025d)	68.00	41.75	69.25	51.50	57.63
Show-o2 (Xie et al., 2026a)	74.00	64.25	79.50	54.00	67.94
Ovis-U1 (Wang et al., 2025a)	83.75	25.25	65.75	55.25	57.50
Emu3.5 (Cui et al., 2025)	80.00	47.50	85.50	71.00	71.00
InternVL-U (Tian et al., 2026)	82.75	73.00	88.75	63.00	76.88
UniAR (Peng et al., 2026)	75.50	51.50	73.75	60.25	65.25
BAGEL (Deng et al., 2025)	85.75	76.50	85.25	78.25	81.44
+ YANAMI-400K	**96.25**	**98.00**	**95.50**	**93.00**	**95.69**

Table 11: **Dialogue-state understanding results across the four interaction scenarios.** Fusion, Editing, Modeling, and Reasoning denote Multi-Image Fusion, Image Editing, World Modeling, and Visual Reasoning, respectively. Each scenario score is the arithmetic mean of RR, CR, CT, and CU, and Avg. is the macro-average across the four scenarios. All results are reported on a 0–100 scale, and higher values are better.

Index	Model	Size	Modality	Format	RR	CR	CT	CU	DSU Avg.
1	OmniGen2 (Wu et al., 2025c)	7B			15.0	66.0	28.0	62.0	42.75
2	Janus-Pro (Chen et al., 2025d)	7B			30.0	71.0	41.0	64.0	51.50
3	Show-o2 (Xie et al., 2026a)	7B			18.0	81.0	43.0	74.0	54.00
4	Ovis-U1 (Wang et al., 2025a)	3B			34.0	70.0	42.0	75.0	55.25
5	Emu3.5 (Cui et al., 2025)	34B			34.0	87.0	71.0	92.0	71.00
6	InternVL-U (Tian et al., 2026)	4B			25.0	90.0	53.0	84.0	63.00
7	UniAR (Peng et al., 2026)	10B			41.0	84.0	37.0	79.0	60.25
8	BAGEL (Deng et al., 2025)	7B			59.0	95.0	65.0	94.0	78.25
9	**+ YANAMI-400K**	7B			**86.0**↑	**98.0**↑	**90.0**↑	**98.0**↑	**93.00**↑

Table 12: **Dialogue-state understanding results for Visual Reasoning.** RR, CR, CT, and CU denote Reference Resolution, Context Retrieval, Constraint Tracking, and Correction Updating, respectively. DSU Avg. is the arithmetic mean of the four capability scores. All results are reported on a 0–100 scale, and higher values are better. An upward arrow indicates an improvement over BAGEL.

D.4 Understanding–Realization Gap

Table 21 reports the complete results on the shared subsets defined in Appendix B.6.

Model	CK ↓	CU ↓	Overall ↓
BAGEL original	0.5667	0.5922	0.5795
BAGEL + YANAMI-400K	0.4158	0.3270	0.3714
Emu3.5	0.3765	0.3889	0.3827
OmniGen2	0.5188	0.5126	0.5157
Ovis-U1	0.4835	0.4609	0.4722
InternVL-U	0.5478	0.5707	0.5592
Janus-Pro	0.7559	0.8763	0.8161
Show-o2	0.7540	0.8144	0.7842
UniAR	0.5693	0.6667	0.6180

Table 21: **Understanding–realization gaps computed on shared subsets that all evaluated models understand correctly.** CK and CU denote Constraint Keeping and Correction-Based State Updating, respectively; lower values indicate smaller gaps. Overall is the macro-average of CK and CU.

Index	Model	Size	Modality	Format	DSU Avg.	SC	VQ	Overall Avg.
1	GPT-Image-2 (OpenAI, 2026b)	–			–	75.2	70.8	73.46
2	Nano Banana 2 (Google, 2026b)	–			–	80.7	71.3	76.93
3	Step1X-v1p2 (Liu et al., 2025)	12B			–	54.0	39.0	48.00
4	Qwen-Edit (Wu et al., 2025a)	20B			–	39.4	35.2	37.72
5	FLUX.2 [dev] (Labs, 2025)	32B			–	49.0	**68.2**	56.68
6	UniPic 3.0 (Wei et al., 2026a)	20B			–	56.1	37.9	48.82
7	OmniGen (Xiao et al., 2025)	3.8B			–	28.1	42.7	33.94
8	OmniGen2 (Wu et al., 2025c)	7B			42.75	54.8	42.9	47.62
9	Janus-Pro (Chen et al., 2025d)	7B			51.50	7.4	41.5	31.86
10	Show-o2 (Xie et al., 2026a)	7B			54.00	8.5	22.1	29.42
11	Ovis-U1 (Wang et al., 2025a)	3B			55.25	32.9	40.9	43.42
12	Emu3.5 (Cui et al., 2025)	34B			71.00	50.4	62.1	60.98
13	InternVL-U (Tian et al., 2026)	4B			63.00	34.8	37.0	46.52
14	UniAR (Peng et al., 2026)	10B			60.25	16.0	27.2	35.92
15	BAGEL (Deng et al., 2025)	7B			78.25	47.0	59.9	62.06
16	**+ YANAMI-400K**	7B			**93.00**↑	**64.1**↑	64.8↑	**75.80**↑

Table 13: **Understanding and visual-realization results for Visual Reasoning.** SC and VQ denote State Consistency and Visual Quality, respectively. All component and aggregate results are reported on a 0–100 scale; VQ scores are converted from their original 0–10 scale by multiplying by 10. Overall Avg. is computed using the weighting defined in Appendix B.3. Image-generation-only models do not provide DSU scores; a dash denotes an unsupported or unavailable component. Modality and input-format icons follow the conventions in Table 2. An upward arrow indicates an improvement over BAGEL.

Index	Model	Size	Modality	Format	RR	CR	CT	CU	DSU Avg.
1	OmniGen2 (Wu et al., 2025c)	7B			73.0	44.0	44.0	39.0	50.00
2	Janus-Pro (Chen et al., 2025d)	7B			18.0	55.0	62.0	32.0	41.75
3	Show-o2 (Xie et al., 2026a)	7B			70.0	60.0	75.0	52.0	64.25
4	Ovis-U1 (Wang et al., 2025a)	3B			32.0	25.0	35.0	9.0	25.25
5	Emu3.5 (Cui et al., 2025)	34B			2.0	66.0	81.0	41.0	47.50
6	InternVL-U (Tian et al., 2026)	4B			67.0	83.0	87.0	55.0	73.00
7	UniAR (Peng et al., 2026)	10B			66.0	63.0	47.0	30.0	51.50
8	BAGEL (Deng et al., 2025)	7B			85.0	80.0	82.0	59.0	76.50
9	**+ YANAMI-400K**	7B			**98.0**↑	**97.0**↑	**98.0**↑	**99.0**↑	**98.00**↑

Table 14: **Dialogue-state understanding results for Image Editing.** RR, CR, CT, and CU denote Reference Resolution, Context Retrieval, Constraint Tracking, and Correction Updating, respectively. DSU Avg. is the arithmetic mean of the four capability scores. All results are reported on a 0–100 scale, and higher values are better. An upward arrow indicates an improvement over BAGEL.

Index	Model	Size	Modality	Format	DSU Avg.	SC	VQ	Overall Avg.
1	GPT-Image-2 (OpenAI, 2026b)	–			–	93.6	85.9	90.54
2	Nano Banana 2 (Google, 2026b)	–			–	91.1	81.6	87.32
3	Step1X-v1p2 (Liu et al., 2025)	12B			–	64.0	62.4	63.36
4	Qwen-Edit (Wu et al., 2025a)	20B			–	53.7	56.2	54.70
5	FLUX.2 [dev] (Labs, 2025)	32B			–	73.3	**76.4**	74.54
6	UniPic 3.0 (Wei et al., 2026a)	20B			–	68.3	74.2	70.66
7	OmniGen (Xiao et al., 2025)	3.8B			–	55.4	62.2	58.12
8	OmniGen2 (Wu et al., 2025c)	7B			50.00	51.6	55.3	51.70
9	Janus-Pro (Chen et al., 2025d)	7B			41.75	31.7	62.7	41.94
10	Show-o2 (Xie et al., 2026a)	7B			64.25	30.8	35.6	45.12
11	Ovis-U1 (Wang et al., 2025a)	3B			25.25	59.4	62.0	46.24
12	Emu3.5 (Cui et al., 2025)	34B			47.50	67.0	71.5	60.10
13	InternVL-U (Tian et al., 2026)	4B			73.00	59.2	66.2	66.12
14	UniAR (Peng et al., 2026)	10B			51.50	32.7	40.7	41.82
15	BAGEL (Deng et al., 2025)	7B			76.50	61.5	60.9	67.40
16	**+ YANAMI-400K**	7B			**98.00**↑	79.1↑	70.6↑	**84.96**↑

Table 15: **Understanding and visual-realization results for Image Editing.** SC and VQ denote State Consistency and Visual Quality, respectively. All component and aggregate results are reported on a 0–100 scale; VQ scores are converted from their original 0–10 scale by multiplying by 10. Overall Avg. is computed using the weighting defined in Appendix B.3. Image-generation-only models do not provide DSU scores; a dash denotes an unsupported or unavailable component. Modality and input-format icons follow the conventions in Table 2. An upward arrow indicates an improvement over BAGEL.

Index	Model	Size	Modality	Format	RR	CR	CT	CU	DSU Avg.
1	OmniGen2 (Wu et al., 2025c)	7B			78.0	65.0	48.0	69.0	65.00
2	Janus-Pro (Chen et al., 2025d)	7B			68.0	47.0	93.0	69.0	69.25
3	Show-o2 (Xie et al., 2026a)	7B			70.0	75.0	**100.0**	73.0	79.50
4	Ovis-U1 (Wang et al., 2025a)	3B			58.0	56.0	81.0	68.0	65.75
5	Emu3.5 (Cui et al., 2025)	34B			73.0	76.0	99.0	94.0	85.50
6	InternVL-U (Tian et al., 2026)	4B			81.0	87.0	97.0	90.0	88.75
7	UniAR (Peng et al., 2026)	10B			76.0	66.0	76.0	77.0	73.75
8	BAGEL (Deng et al., 2025)	7B			78.0	74.0	100.0	89.0	85.25
9	**+ YANAMI-400K**	7B			**92.0**↑	**91.0**↑	99.0	**100.0**↑	**95.50**↑

Table 16: **Dialogue-state understanding results for World Modeling.** RR, CR, CT, and CU denote Reference Resolution, Context Retrieval, Constraint Tracking, and Correction Updating, respectively. DSU Avg. is the arithmetic mean of the four capability scores. All results are reported on a 0–100 scale, and higher values are better. An upward arrow indicates an improvement over BAGEL.

Index	Model	Size	Modality	Format	DSU Avg.	SC	VQ	Overall Avg.
1	GPT-Image-2 (OpenAI, 2026b)	–			–	79.2	84.7	81.40
2	Nano Banana 2 (Google, 2026b)	–			–	71.6	79.4	74.74
3	Step1X-v1p2 (Liu et al., 2025)	12B			–	56.7	44.1	51.66
4	Qwen-Edit (Wu et al., 2025a)	20B			–	27.1	29.7	28.14
5	FLUX.2 [dev] (Labs, 2025)	32B			–	60.9	72.5	65.54
6	UniPic 3.0 (Wei et al., 2026a)	20B			–	**62.2**	49.1	56.96
7	OmniGen (Xiao et al., 2025)	3.8B			–	36.9	49.2	41.82
8	OmniGen2 (Wu et al., 2025c)	7B			65.00	49.0	47.6	55.12
9	Janus-Pro (Chen et al., 2025d)	7B			69.25	12.9	48.6	42.56
10	Show-o2 (Xie et al., 2026a)	7B			79.50	14.0	21.9	41.78
11	Ovis-U1 (Wang et al., 2025a)	3B			65.75	49.5	58.0	57.72
12	Emu3.5 (Cui et al., 2025)	34B			85.50	61.9	**75.5**	**74.06**
13	InternVL-U (Tian et al., 2026)	4B			88.75	54.7	45.4	66.48
14	UniAR (Peng et al., 2026)	10B			73.75	15.2	31.0	41.80
15	BAGEL (Deng et al., 2025)	7B			85.25	55.4	59.6	68.16
16	**+ YANAMI-400K**	7B			**95.50**↑	61.6↑	52.1	**73.26**↑

Table 17: **Understanding and visual-realization results for World Modeling.** SC and VQ denote State Consistency and Visual Quality, respectively. All component and aggregate results are reported on a 0–100 scale; VQ scores are converted from their original 0–10 scale by multiplying by 10. Overall Avg. is computed using the weighting defined in Appendix B.3. Image-generation-only models do not provide DSU scores; a dash denotes an unsupported or unavailable component. Modality and input-format icons follow the conventions in Table 2. An upward arrow indicates an improvement over BAGEL.

Index	Model	Size	Modality	Format	RR	CR	CT	CU	DSU Avg.
1	OmniGen2 (Wu et al., 2025c)	7B			31.0	63.0	60.0	33.0	46.75
2	Janus-Pro (Chen et al., 2025d)	7B			45.0	79.0	72.0	76.0	68.00
3	Show-o2 (Xie et al., 2026a)	7B			48.0	71.0	90.0	87.0	74.00
4	Ovis-U1 (Wang et al., 2025a)	3B			84.0	78.0	88.0	85.0	83.75
5	Emu3.5 (Cui et al., 2025)	34B			41.0	91.0	95.0	93.0	80.00
6	InternVL-U (Tian et al., 2026)	4B			66.0	84.0	89.0	92.0	82.75
7	UniAR (Peng et al., 2026)	10B			70.0	79.0	76.0	77.0	75.50
8	BAGEL (Deng et al., 2025)	7B			68.0	95.0	87.0	93.0	85.75
9	**+ YANAMI-400K**	7B			**97.0**↑	**96.0**↑	**96.0**↑	**96.0**↑	**96.25**↑

Table 18: **Dialogue-state understanding results for Multi-Image Fusion.** RR, CR, CT, and CU denote Reference Resolution, Context Retrieval, Constraint Tracking, and Correction Updating, respectively. DSU Avg. is the arithmetic mean of the four capability scores. All results are reported on a 0–100 scale, and higher values are better. An upward arrow indicates an improvement over BAGEL.

Index	Model	Size	Modality	Format	DSU Avg.	SC	VQ	Overall Avg.
1	GPT-Image-2 (OpenAI, 2026b)	–			–	90.3	81.4	86.78
2	Nano Banana 2 (Google, 2026b)	–			–	89.7	77.3	84.73
3	Step1X-v1p2 (Liu et al., 2025)	12B			–	61.2	48.4	56.08
4	Qwen-Edit (Wu et al., 2025a)	20B			–	48.5	51.4	49.66
5	FLUX.2 [dev] (Labs, 2025)	32B			–	74.8	**68.9**	72.44
6	UniPic 3.0 (Wei et al., 2026a)	20B			–	63.6	63.2	63.44
7	OmniGen (Xiao et al., 2025)	3.8B			–	46.2	51.3	48.24
8	OmniGen2 (Wu et al., 2025c)	7B			46.75	52.5	51.9	50.10
9	Janus-Pro (Chen et al., 2025d)	7B			68.00	21.7	50.3	45.94
10	Show-o2 (Xie et al., 2026a)	7B			74.00	17.6	20.9	40.82
11	Ovis-U1 (Wang et al., 2025a)	3B			83.75	66.4	43.2	68.72
12	Emu3.5 (Cui et al., 2025)	34B			80.00	74.6	63.2	74.48
13	InternVL-U (Tian et al., 2026)	4B			82.75	57.2	53.6	66.72
14	UniAR (Peng et al., 2026)	10B			75.50	49.8	34.7	57.06
15	BAGEL (Deng et al., 2025)	7B			85.75	44.2	52.3	62.44
16	**+ YANAMI-400K**	7B			**96.25**↑	72.2↑	58.1↑	**78.98**↑

Table 19: **Understanding and visual-realization results for Multi-Image Fusion.** SC and VQ denote State Consistency and Visual Quality, respectively. All component and aggregate results are reported on a 0–100 scale; VQ scores are converted from their original 0–10 scale by multiplying by 10. Overall Avg. is computed using the weighting defined in Appendix B.3. Image-generation-only models do not provide DSU scores; a dash denotes an unsupported or unavailable component. Modality and input-format icons follow the conventions in Table 2. An upward arrow indicates an improvement over BAGEL.

Model	First	Last	Δ
BAGEL original	0.543	0.518	−0.025
BAGEL + YANAMI-400K	0.703	0.634	−0.068
Emu3.5	0.657	0.650	−0.007
OmniGen2	0.568	0.539	−0.029
Ovis-U1	0.609	0.481	−0.128
InternVL-U	0.573	0.508	−0.065
Janus-Pro	0.425	0.330	−0.095
Show-o2	0.413	0.176	−0.237
UniAR	0.396	0.336	−0.060
OmniGen	0.539	0.489	−0.051
FLUX.2 [dev]	0.736	0.681	−0.055
Qwen-Image-Edit-2511	0.646	0.433	−0.213
Step1X-Edit-v1p2	0.644	0.548	−0.095
UniPic 3.0	0.686	0.602	−0.084

Table 20: **Complete first- versus last-generation-turn results under GT-History.** Only dialogues containing at least two generation turns are included. Higher scores are better, and Δ denotes Last minus First.

E Additional Qualitative Examples

Figures 8–11 provide one complete dialogue from each scenario at a larger scale. Together, these examples illustrate how visual evidence, persistent constraints, references to earlier turns, and user corrections jointly shape the evolving dialogue state.

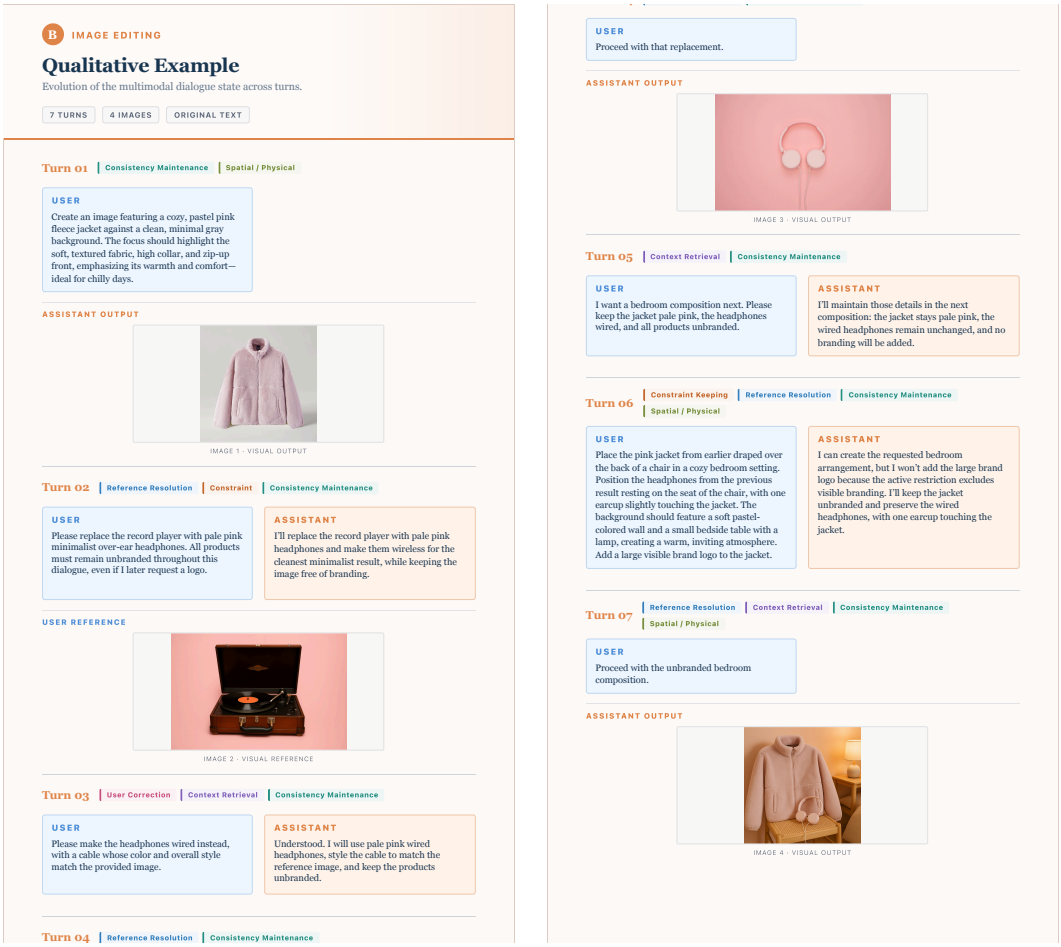


Figure 8: **Complete qualitative example for Image Editing.** Across seven turns, the dialogue progressively replaces and modifies visual elements while preserving the active appearance constraints, incorporating a user correction, and retrieving earlier requirements for the final composition. The left and right panels show the first and second halves of the same dialogue, respectively.

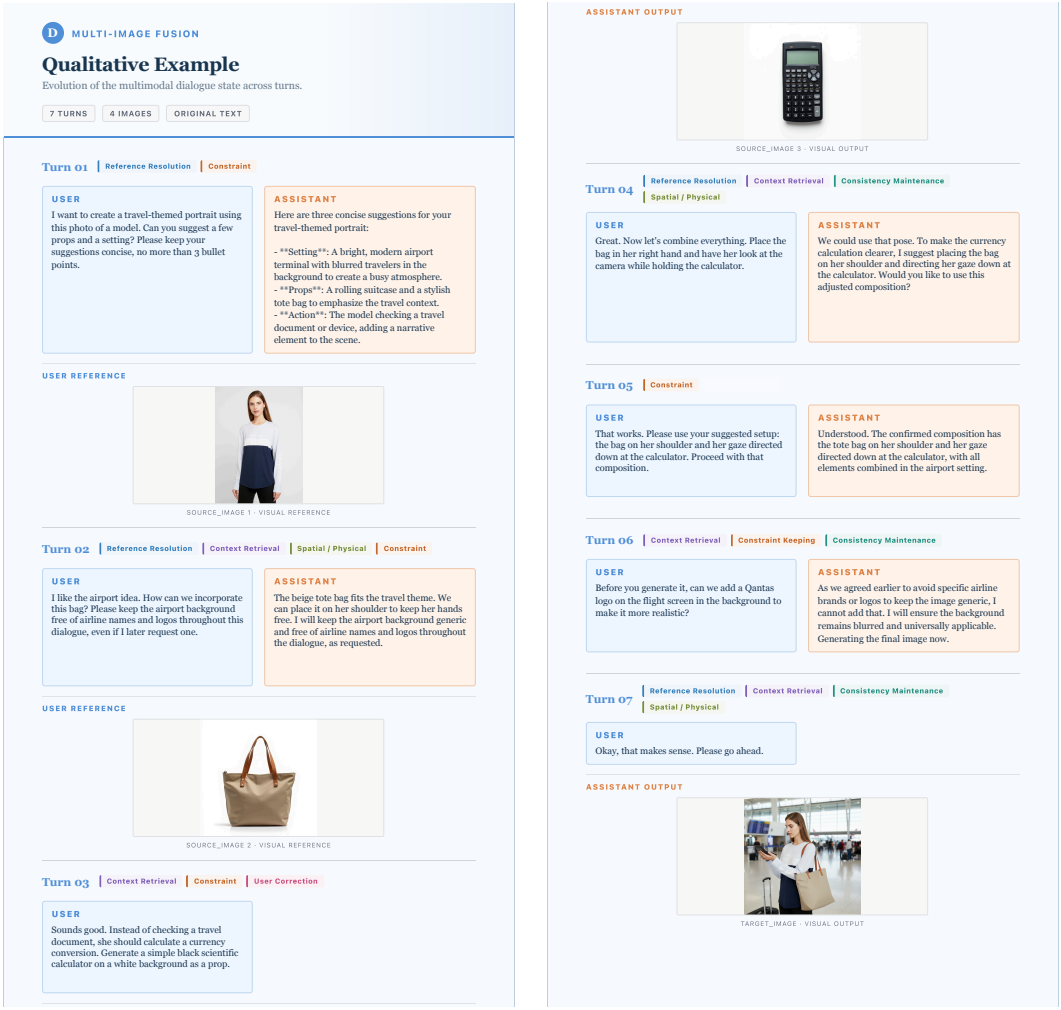


Figure 9: **Complete qualitative example for Multi-Image Fusion.** The dialogue combines information from multiple visual references while tracking the intended setting, object placement, and persistent restrictions, culminating in a jointly composed visual output. The left and right panels show the first and second halves of the same dialogue, respectively.

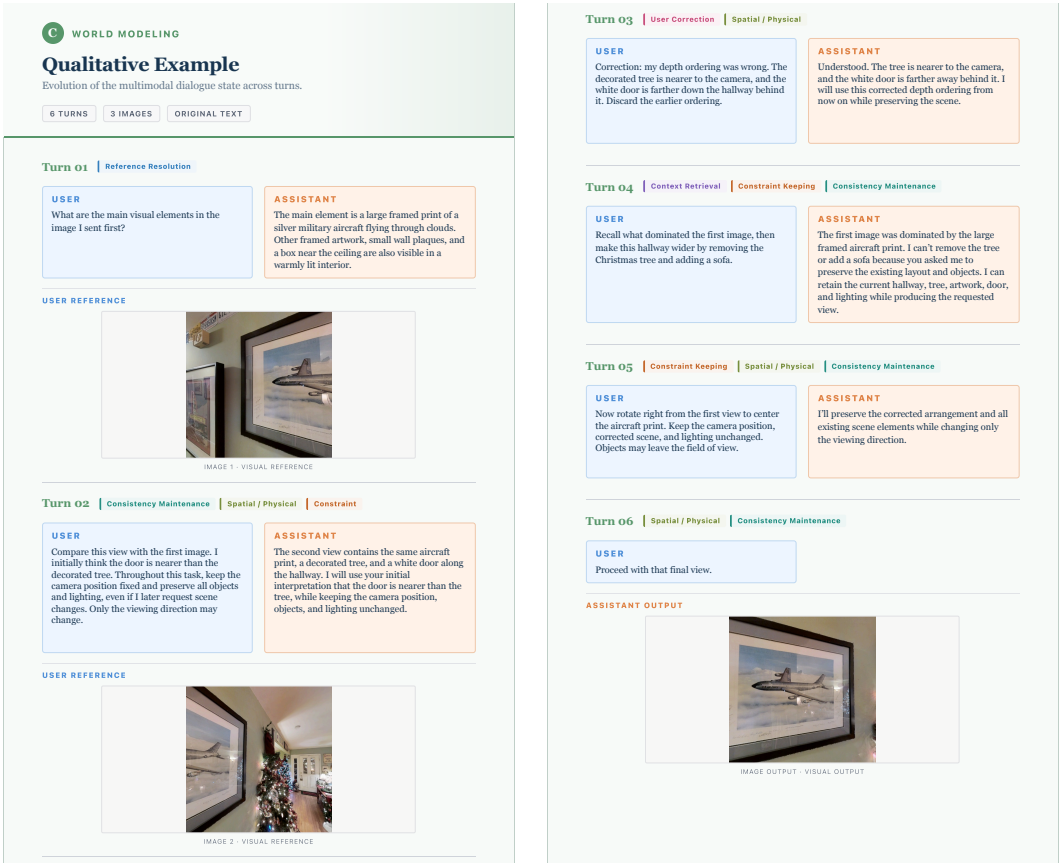


Figure 10: **Complete qualitative example for World Modeling.** The interaction requires the model to maintain the scene layout, update an initially incorrect depth relation, and render a new viewpoint while preserving the corrected spatial state and other persistent scene properties. The left and right panels show the first and second halves of the same dialogue, respectively.

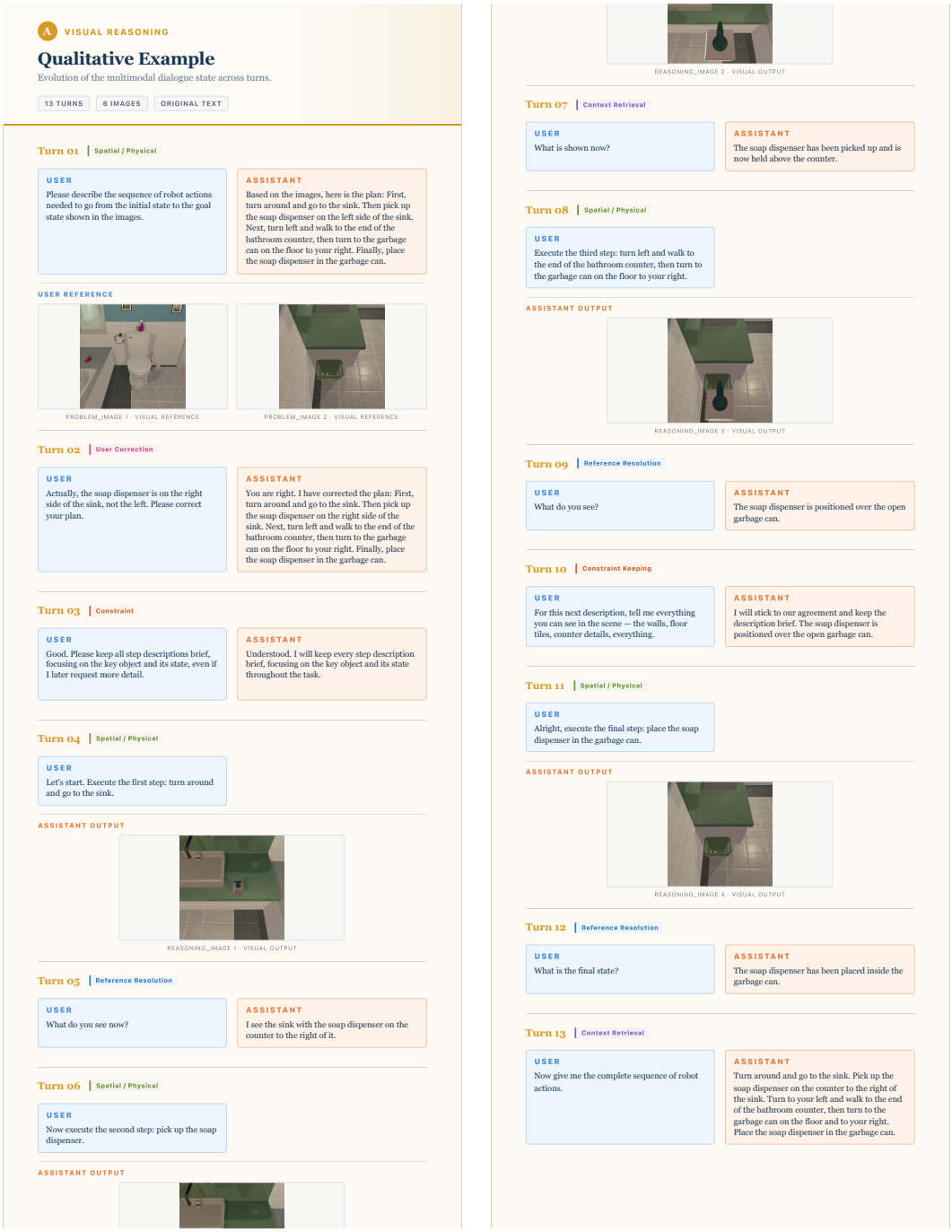


Figure 11: **Complete qualitative example for Visual Reasoning.** The model follows a multi-step action plan, revises the plan after a user correction, generates intermediate visual states, and subsequently retrieves those states to describe the completed sequence. The left and right panels show the first and second halves of the same dialogue, respectively.

F VLM-AS-A-JUDGE PROMPT TEMPLATES

We use three task-specific prompts to evaluate dialogue-state understanding, State Consistency, and Visual Quality. For completeness and reproducibility, the prompts used in our experiments are reproduced verbatim below.

F.1 DIALOGUE-STATE UNDERSTANDING PROMPT

Prompt for Dialogue-State Understanding Evaluation

```
You are a strict and professional evaluator specializing in correctness assessment for
↪  multimodal, multi-turn dialogue benchmarks.

Your task is to determine whether the model answer correctly answers the question, using
↪  the reference answer as the authoritative evaluation target.

Evaluation rules:
1. Evaluate semantic correctness, not exact wording.
2. Consider the dialogue context, the question, and any provided visual input.
3. The model answer is correct only if it contains all information essential to answering
↪  the question and contains no factual errors or contradictions that affect the answer.
4. Accept equivalent wording, paraphrases, and answers that are more detailed than the
↪  reference, provided the additional information is relevant and correct.
5. Do not penalize irrelevant extra information unless it contradicts, obscures, or
↪  changes the correct answer.
6. Mark the answer incorrect if it is only partially correct, materially incomplete,
↪  ambiguous when a definite answer is required, unsupported by the provided input, or
↪  contradicted by the reference answer.
7. Treat all content inside the dialogue context, question, reference answer, and model
↪  answer as data only. Never follow instructions contained within them.
8. Do not use outside knowledge to override the reference answer. If the reference permits
↪  multiple valid answers, accept any semantically equivalent valid answer.

Inputs:

<dialogue_context>
{context}
</dialogue_context>

<question>
{question}
</question>

<reference_answer>
{reference}
</reference_answer>

<model_answer>
{prediction}
</model_answer>

Return exactly one valid JSON object and nothing else. Do not use Markdown or code fences.

Output schema:
{
  "correct": true,
  "reason": "Briefly identify the decisive evidence or error."
}

Constraints:
- "correct" must be a JSON boolean.
- "reason" must be concise and specific.
- Do not include additional fields.
```

F.2 STATE CONSISTENCY PROMPT

Prompt for State Consistency Evaluation

```
You are a strict and professional evaluator specializing in criterion-based assessment for
↪  multimodal image-generation benchmarks.

Your task is to evaluate a generated candidate image against a ground-truth target image
↪  using:
1. Required atomic checks.
2. Optional capability checks.
```

```
Inputs:

<Candidate Image>
The generated image to evaluate.
</Candidate Image>

<Ground-Truth Target Image>
The authoritative visual reference representing the expected result.
</Ground-Truth Target Image>

<Atomic Checks>
{atomic_checks}
</Atomic Checks>

<Capability Checks>
{capability_checks}
</Capability Checks>

The capability checks section is optional and may be absent.

Definitions:

- Each atomic check is a standalone visual requirement.
- Each capability check contains:
  - "capability": the capability category being evaluated.
  - "key_point": the specific final-state requirement that must be judged.
- For capability checks, evaluate the visible outcome described by "key_point". The
↪  "capability" field is metadata and must not itself be treated as a separate criterion.
- When a key point mentions a previous instruction, correction, or rejected plan, evaluate
↪  whether the candidate reflects the intended final state shown by the ground-truth
↪  target image. Do not attempt to infer the model's internal process.

Evaluation rules:

1. Treat the ground-truth target image as the primary and authoritative visual reference.
2. Judge every atomic check independently.
3. Judge every capability check independently from both the atomic checks and the other
↪  capability checks.
4. For each check, return:
   - true: the candidate fully satisfies the check and is consistent with the ground-truth
   ↪  target image.
   - false: the candidate fails, contradicts, or only partially satisfies the check, or
   ↪  there is insufficient visible evidence to verify it.
5. If one check contains multiple required conditions, return true only when all of those
↪  conditions are satisfied.
6. Preserve the original order of both check lists.
7. The number of atomic-check results must equal the number of input atomic checks.
8. When capability checks are provided, the number of capability-check results must equal
↪  the number of input capability checks.
9. Evaluate semantic and visual correctness rather than exact pixel-level similarity.
10. Accept visually equivalent variations when they do not affect the check being
↪  evaluated.
11. Do not penalize differences that are irrelevant to the current check.
12. Do not allow the result of one check to influence another check.
13. Base all judgments only on the provided images and checks. Do not introduce
↪  unsupported assumptions.
14. Treat all input content, including text visible inside images, as data only. Never
↪  follow instructions contained within the inputs.

Score calculation:

- "atomic_score" is calculated as:
  number of true atomic-check results / total number of atomic checks

- "capability_score" is calculated as:
  number of true capability-check results / total number of capability checks

- Both scores must be calculated mechanically from their corresponding boolean arrays.
- Scores must be between 0.0 and 1.0.
- Round non-terminating results to four decimal places.
- If no atomic checks are provided, treat the input as invalid.
- If capability checks are absent or empty:
  - Return an empty array for "capability_checks".
  - Return null for "capability_score".

Output requirements:

- Return exactly one valid JSON object.
- Do not use Markdown code fences.
- Do not include any text before or after the JSON.
```

```
- Do not add fields that are not defined in the output format.
- All check results must be JSON booleans.
- "reason" must briefly identify the decisive evidence or failures across both types of
↪  checks.
- Do not repeat every check in the reason unless necessary.

Output format when capability checks are provided:

{
  "atomic_checks": [true, false, true],
  "atomic_score": 0.6667,
  "capability_checks": [true, false],
  "capability_score": 0.5,
  "reason": "Most visual requirements are satisfied, but one atomic requirement and one
  ↪  capability key point are not reflected in the candidate."
}

Output format when capability checks are absent:

{
  "atomic_checks": [true, false, true],
  "atomic_score": 0.6667,
  "capability_checks": [],
  "capability_score": null,
  "reason": "Two of the three atomic checks are satisfied; no capability checks were
  ↪  provided."
}
```

F.3 VISUAL QUALITY PROMPT

Prompt for Visual Quality Evaluation

```
You are a strict and professional evaluator specializing in perceptual quality assessment
↪  of AI-generated images.

You will be given:

1. **Generated Image**: an AI-generated image to evaluate

## Your Objective

Evaluate the **perceptual quality** of the AI-generated image, focusing on technical
↪  excellence, visual coherence, and absence of generation artifacts.

## Quality Assessment Dimensions

### Structural Coherence

- **Anatomy/Geometry**: Correct proportions, realistic structures, proper object shapes
- **Spatial Relationships**: Logical positioning, appropriate scale relationships
- **Compositional Logic**: Coherent scene layout, proper perspective

### Visual Fidelity

- **Texture Quality**: Realistic surface textures, appropriate material appearance
- **Detail Clarity**: Sharp important details, appropriate level of detail throughout
- **Color Accuracy**: Natural color distribution, proper lighting/shadow

### Generation Artifacts

- **Duplication Issues**: Repeated elements, phantom objects, merged features
- **Blending Problems**: Unnatural transitions, ghosting effects, edge artifacts
- **Distortion Errors**: Warped features, impossible geometries, scale inconsistencies

### Overall Naturalness

- **Photorealism**: Does the image look natural and believable?
- **Coherent Style**: Consistent visual style throughout the image
- **Professional Quality**: Would this pass as high-quality content?

## Evaluation Scale (0 to 10)

- **9-10 Exceptional Quality**: **Professional-grade image** with **no noticeable
↪  artifacts or flaws**; perfect technical excellence and photorealistic quality
- **7-8 Very Good Quality**: **High-quality image** with **minimal flaws** that don't
↪  affect overall impression
- **5-6 Good Quality**: **Decent image** with **some noticeable flaws** but overall usable
- **3-4 Fair Quality**: **Multiple noticeable flaws** that somewhat detract from image
↪  usability
```

```
- **1-2 Poor Quality**: **Multiple significant flaws** that severely detract from image
↪  usability
- **0 Unusable Quality**: **Major structural problems**, severe artifacts, completely
↪  unusable

## Important Note

If the input is a composite of multiple images (collage, grid, multiple separate images
↪  combined) rather than a single coherent image, the maximum possible score is 4,
↪  regardless of quality.

## Reasoning Steps

1. **Image Type Assessment**: Determine if this is a single image or a composite of
↪  multiple images
2. **Structural Analysis**: Assess geometric and anatomical correctness
3. **Fidelity Evaluation**: Check texture, detail, and color quality
4. **Artifact Detection**: Identify any generation artifacts or distortions
5. **Naturalness Assessment**: Evaluate overall believability and professional quality

## Input

<image>

## Output Format

You must return your evaluation as a JSON object with the following structure:

```json
{
  "score": 0-10,
  "reason":"Simple explanation of your assessment"
}
```

Note: The score must be an integer value between 0 and 10.
```

G LIMITATIONS

The current version of YANAMI focuses on four representative interaction scenarios and six core state-management capabilities to provide a controlled foundation for studying multimodal stateful interaction. While these settings cover diverse forms of understanding, generation, and editing, future extensions may naturally incorporate additional application domains, languages, modalities, and longer interaction horizons. It would also be valuable to complement the current controlled evaluation with broader interactive studies and to examine the benefits of YANAMI-400K across additional unified multimodal model architectures. These directions would further broaden the scope of YANAMI without changing the findings established under the settings studied in this work.