MACRO: ADVANCING MULTI-REFERENCE IMAGE GENERATION WITH STRUCTURED LONG-CONTEXT DATA

Anonymous authors
Paper under double-blind review



Figure 1: **Overview of MacroData.** MacroData contains 400K high-quality samples with up to 10 input images across four long-context multi-reference image generation tasks: (a) **Customization**: generating compositions conditioned on multiple reference images, (b) **Illustration**: producing illustrative images based on multimodal context, (c) **Spatial**: predicting novel view images given multiple views, specifically including outside-in objects and inside-out scenes, and (d) **Temporal**: forecasting future frames based on historical sequence. Each task is composed of 100k samples, split into different numbers of reference images, including 1-3, 4-5, 6-7, and 8-10 input images.

ABSTRACT

Generating images conditioned on multiple visual references is critical for real-world applications such as multi-subject composition, narrative illustration, and novel view synthesis, yet current models suffer from severe performance degradation as the number of input references grows. We identify the root cause as

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.

a fundamental data bottleneck: existing datasets are dominated by single- or few-reference pairs and lack the structured, long-context supervision needed to learn dense inter-reference dependencies. To address this, we introduce MacroData, a large-scale dataset of 400K samples, each containing up to 10 reference images, systematically organized across four complementary dimensions—Customization, Illustration, Spatial reasoning, and Temporal dynamics—to provide comprehensive coverage of the multi-reference generation space. Recognizing the concurrent absence of standardized evaluation protocols, we further propose MacroBench, a benchmark of 4,000 samples that assesses generative coherence across graded task dimensions and input scales. Extensive experiments show that fine-tuning on MacroData yields substantial improvements in multi-reference generation, and ablation studies further reveal synergistic benefits of cross-task co-training and effective strategies for handling long-context complexity. Our website and data are available for anonymous review.

1 INTRODUCTION

In-context generation (Wu et al., 2025b; Labs et al., 2025; Sheynin et al., 2024; Deng et al., 2025; Xie et al., 2023; Chen et al., 2025) has become a prominent paradigm in image synthesis (Rombach et al., 2022; Podell et al., 2023; Chen et al., 2023; Labs, 2024; Xie et al., 2024a; Lipman et al., 2022; Esser et al., 2024), enabling models to generate images directly conditioned on interleaved text and visual references. While recent advances have achieved impressive results on single- and few-reference tasks such as identity-preserving generation and style transfer, many real-world scenarios—multi-subject composition, narrative illustration, novel view synthesis—naturally demand reasoning over a larger set of reference images. This *multi-reference* setting is substantially more challenging and remains largely unsolved.

Even the most capable open-source models struggle in this regime. For instance, OmniGen2 (Wu et al., 2025b) is constrained to a maximum of five input images, restricting its utility for more complex compositions, and Bagel (Deng et al., 2025), despite being theoretically designed to accept unlimited inputs, exhibits severe performance degradation beyond three references. We attribute this deficiency primarily to a critical scarcity of high-quality, structured multi-reference training data. Unlike single-reference generation, which mainly involves extracting and reproducing features from one source, multi-reference generation requires models to comprehend dense inter-reference dependencies—temporal dynamics, spatial consistency, and identity preservation across inputs—and seamlessly integrate them into a coherent output. Existing datasets (Ye et al., 2025a; Liu et al., 2025; Wang et al., 2025; Li et al., 2025a; Ye et al., 2025b), however, are predominantly composed of single- or few-reference pairs and thus fail to provide the necessary training signals for learning these complex relationships.

To bridge this gap, we introduce **MacroData** (**M**ulti-image d**A**taset for **C**ontext-**R**eferencing generati**O**n), a large-scale dataset comprising **400K** samples with up to 10 reference images per sample. Rather than targeting a single capability, MacroData is structured around four complementary dimensions essential for multi-reference generation: 1) **Customization**, composing multiple referents into coherent original scenes; 2) **Illustration**, generating context-aware images that complement textual narratives; 3) **Spatial**, synthesizing novel viewpoints from multiple input views; and 4) **Temporal**, predicting future keyframes from video image sequences. To ensure logical consistency and visual fidelity, we design a robust pipeline that distills knowledge from advanced closed-source models and meticulously filters real-world corpora, rather than relying on noisy web-scraped data (Figs. 3 to 6).

Beyond the training data, we observe that the lack of a standardized evaluation protocol has also hindered progress in multi-reference generation. Existing benchmarks such as OmniContext (Wu et al., 2025b) are limited to customization tasks with at most three input images, leaving the broader setting unevaluated. We therefore propose **MacroBench**, a comprehensive benchmark comprising 4,000 samples that evaluates generative coherence across all four task dimensions and graded varying input scales from 1 to 10 images. Following recent works (Ye et al., 2025b;a; Wei et al., 2025), we employ an LLM-as-Judge mechanism with task-specific criteria to rigorously assess context sensitivity and narrative adherence.

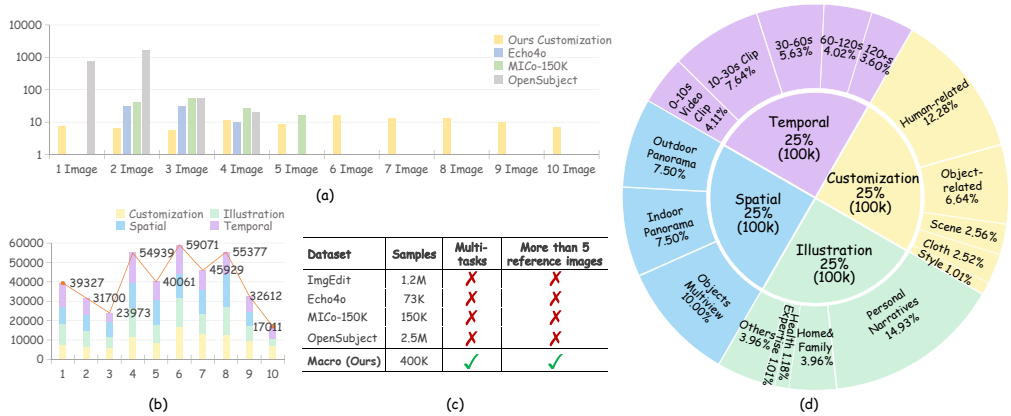

Figure 2: **Statistics of MacroData.** MacroData comprises four tasks, each containing 100k samples. (a) The number of input images in the customization subtask averaged 5.84 images per sample, with a maximum of 10. (b) The number of input images across all tasks averaged 5.44 images per sample. (c) Comparison among different datasets. (d) The distribution of data composition for each task.

Equipped with MacroData and MacroBench, we conduct extensive experiments by fine-tuning state-of-the-art open-source models (e.g., Bagel (Deng et al., 2025)) on our dataset. Results show that MacroData unlocks substantial improvements in multi-reference generation, significantly narrowing the gap with closed-source models. Furthermore, we provide in-depth ablation studies on data scaling and task synergies to guide future research and explore potential strategies for further enhancing multi-reference generation performance.

Our main contributions are summarized as follows: 1) We introduce **MacroData**, a large-scale multi-reference generation dataset comprising 400K samples with up to 10 reference images, structured across four complementary dimensions—Customization, Illustration, Spatial, and Temporal—to facilitate inter-reference dependency learning. 2) We propose **MacroBench**, a standardized benchmark with 4,000 samples that evaluates multi-reference generative coherence across graded task dimensions and input scales, filling a critical evaluation void in this domain. 3) We conduct extensive experiments demonstrating that MacroData significantly enhances long-context multi-reference generation, and identify effective strategies including cross-task co-training and token selection for handling complex multi-reference contexts.

2 Related Work

In-Context Image Generation Model. In-context image generation requires models to jointly comprehend preceding visual and textual inputs and synthesize coherent images conditioned on them. To this end, diverse architectural paradigms have been explored (Ma et al., 2025; Xie et al., 2024b; Team, 2024; Zhou et al., 2024; Shi et al., 2024), with recent state-of-the-art models (Deng et al., 2025; Wu et al., 2025b;a; Xie et al., 2025; Cui et al., 2025) achieving strong results through autoregressive, hybrid, or diffusion-based architectures with specialized vision representations. Among them, Bagel (Deng et al., 2025) introduces a Mixture-of-Transformer design that separately processes understanding and generation tokens, while OmniGen2 (Wu et al., 2025b) co-trains diffusion models with LVLM hidden states for tighter vision-language alignment. Despite these advances, open-source models remain limited to processing at most 3–5 input images (Deng et al., 2025; Wu et al., 2025b;a), with performance degrading sharply as reference counts increase. We attribute this limitation largely to the absence of structured data for multi-reference scenarios, which motivates our data-centric approach.

In-Context Image Generation Dataset. Constructing high-quality training data for in-context generation remains a significant challenge. Existing datasets are built primarily through two strategies: distillation from strong generative models (Tan et al., 2025; Wu et al., 2025c; Ye et al., 2025a; Wei et al., 2025; Ye et al., 2025b) and retrieval from real-world corpora (Xu et al., 2025; Liu et al., 2025). For instance, Echo4o (Ye et al., 2025a) and MICo (Wei et al., 2025) prompt closed-source models to synthesize identity-consistent pairs, while OpenSubject (Liu et al., 2025) extracts and matches relevant images from web pages and videos. However, these datasets are limited in two critical aspects: they focus narrowly on customization and editing tasks, and rarely include more than three to five

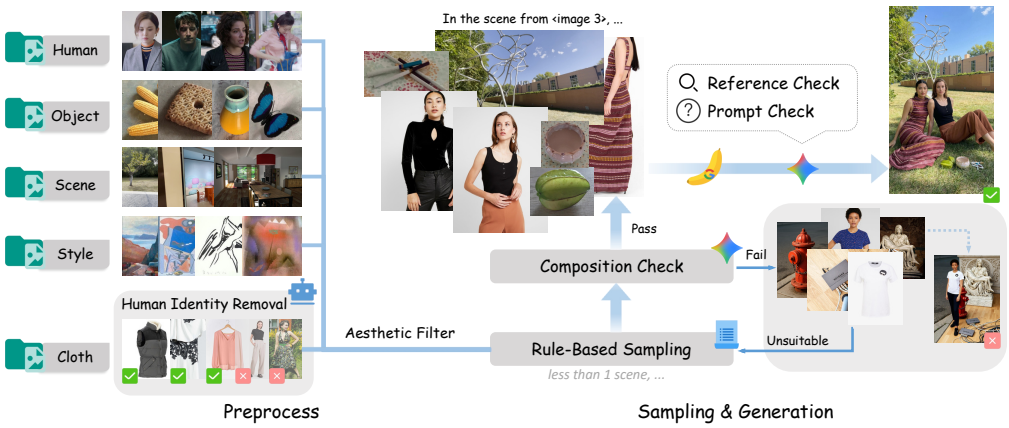

Figure 3: **Customization Subset Pipeline** composites preprocessed metadata via rule-based and VLM-reasoned sampling and finally applies a bidirectional assessment to ensure reference fidelity and prompt consistency.

reference images per sample. This dual limitation—in both task diversity and input scale—hinders the development of models for general, long-context multi-reference generation.

In-Context Image Generation Benchmark. Evaluating in-context generation poses unique challenges, as outputs must be assessed for consistency with multiple heterogeneous inputs spanning different modalities and semantic roles. Recent benchmarks (Wei et al., 2025; Ye et al., 2025b; Wu et al., 2025b; Liu et al., 2025) adopt an LLM-as-Judge paradigm following text-to-image evaluation practices (Ghosh et al., 2023; Hu et al., 2024). For example, OmniContext (Wu et al., 2025b) employs GPT-4.1 (OpenAI, 2024) to score prompt adherence and subject consistency. However, existing benchmarks are restricted to customization and editing scenarios with at most a few input images, lacking evaluation coverage for spatial reasoning, temporal coherence, and the systematic scaling of input references.

3 MacroData

3.1 Overview

We introduce MacroData, a comprehensive dataset of **400K** samples for multi-reference image generation. Addressing the scarcity of many-to-one data, it supports up to 10 inputs per sample with an average of **5.44**. As shown in Fig. 2(a) and (c), it significantly surpasses datasets like OpenSubject (Liu et al., 2025) and Echo4o (Ye et al., 2025a), which lack samples beyond five references. In contrast, MacroData offers substantial long-context coverage, concentrated among more than 4 references and peaking at ∼59K samples for 6 references as depicted in Fig. 2(b). This abundance of long-context samples is crucial for models to leverage extensive visual contexts.

Task Definition. To ensure data balance, MacroData is equally divided into four distinct tasks with 100K samples each, as detailed in Fig. 2(d): 1) *Customization:* Covers 5 subject categories (human, object, scene, cloth, and style). 2) *Illustration:* Features 100 diverse topics (e.g., narratives and health) derived from interleaved contexts. 3) *Spatial:* Focuses on 3D consistency for multiview objects and panoramas. 4) *Temporal:* Captures dynamics across video clips of varying durations (0 to 120+ seconds). Detailed construction pipeline is provided in Section A.

3.2 Customization Subset

Source Collection and Preprocessing. As depicted in Fig. 3, we aggregate source metadata from OpenSubject (Liu et al., 2025) for human, MVImgNet (Yu et al., 2023) for object, DL3DV (Ling et al., 2024) for scene, Vibrant Clothes Rental Dataset (Borgersen et al., 2024) for cloth and WikiArt (WikiArt, n.d.) for style. Each undergoes tailored preprocessing: we uniformly sample video keyframes for scenes, filter out clothing images containing faces via VLMs to prevent identity leakage, and categorize styles using tags to guarantee diversity. Finally, aesthetic scoring is applied to select high-quality images.

Sampling and Generation. We sample metadata across these categories to build diverse composition sets, using LLMs to evaluate and iteratively resample any logically or spatially unreasonable

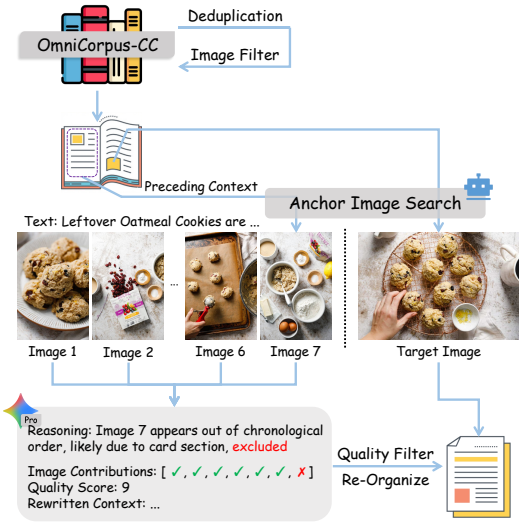

Figure 4: **Illustration Subset Pipeline** identifies highly relevant anchor images from interleaved data as targets and utilizes VLMs to rewrite and filter the context for narrative coherence.

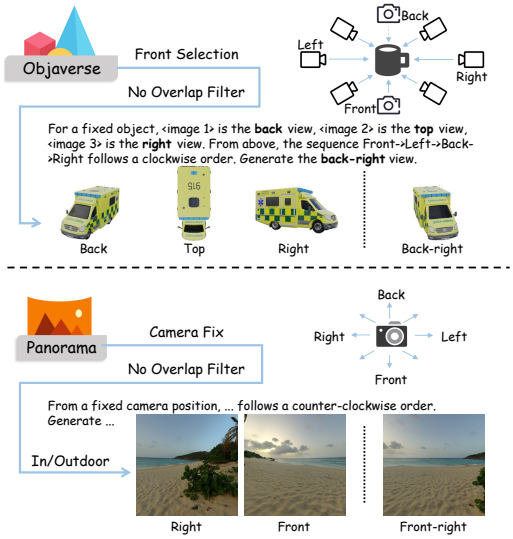

Figure 5: **Spatial Subset Pipeline** samples input and target views from canonical directions for outside-in objects and inside-out panoramas, applying spatial overlap filters for plausibility.

combinations. The valid sets are then synthesized into target images. To ensure fidelity, we perform a VLM-based bidirectional assessment, filtering out samples where input images are not faithfully reflected or where text prompts are semantically inconsistent with the generated images.

3.3 Illustration Subset

Anchor Image Search. As illustrated in Fig. 4, we leverage large-scale interleaved image-text sequences from web crawls of OmniCorpus-CC-210M (Li et al., 2024) as raw source material. Since such data is inherently noisy, we employ VLMs to identify "anchor images" that exhibit high semantic relevance to both the accompanying text and preceding images. These anchors are designated as generation targets, with the preceding context serving as input conditions.

Sample Reorganization. Since the preceding contexts are often noisy and textually verbose, we employ strong VLMs to regenerate each sequence. The VLM re-evaluates semantic relevance, discards images inconsistent with the narrative flow, and synthesizes a concise, coherent textual context. Each reorganized sample is then assigned a quality score, and low-scoring samples are filtered out to finalize the Illustration dataset.

3.4 Spatial Subset

Outside-in Objects. As shown in the upper section of Fig. 5, we construct this subtask from multi-view 3D object renderings data from G-buffer Objaverse dataset (Qiu et al., 2024). The outside-in setup focuses on capturing a central object from surrounding external viewpoints. To bridge the domain gap with real imagery, we filter out low-quality samples (e.g., transparent or white-textured objects) based on color saturation and brightness in HSV space. From a set of canonical views, we designate one as the target and randomly sample diverse input views from the remainder, strictly ensuring visual overlap for physical plausibility.

Inside-out Scenes. For inside-out scenes (Fig. 5, bottom), which capture the surrounding environment by looking outward from a central internal point, we start from panoramic images of DIT360 (Feng et al., 2025), Pano360 (Kocabas et al., 2021), and Polyhaven (Polyhaven, n.d.), filtering out non-standard formats (e.g., fisheye) and categorizing the rest into indoor and outdoor scenes using VLMs. We define canonical viewing directions from the panorama's center, designate a target view, and sample input views from the remaining directions, ensuring spatial overlap with the target.

3.5 Temporal Subset

Video Clip Extraction. To mitigate redundancy in raw video content from OmniCorpus-YT (Li et al., 2024), we apply shot boundary detection (Soucek & Lokoc, 2024) to segment videos into

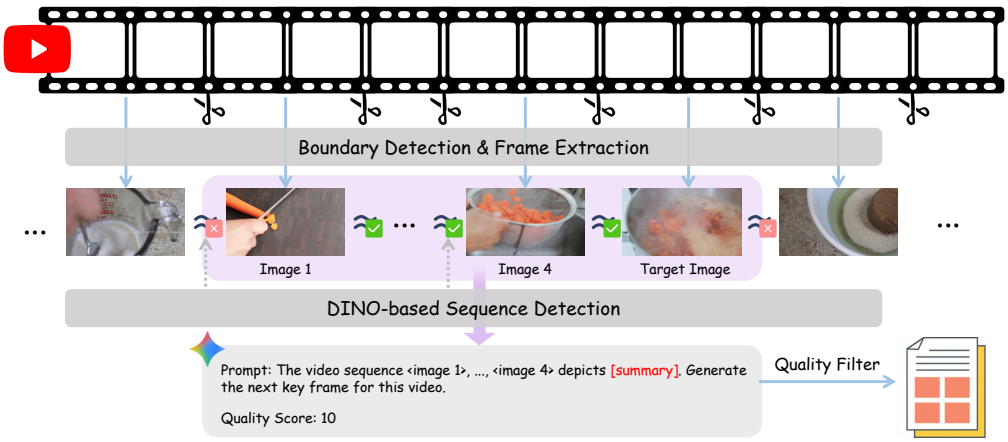

Figure 6: **Temporal Subset Pipeline** extracts video keyframes, groups them into coherent sequences using DINOv2, and utilizes a VLM to generate a summary prompt for predicting the final frame.

semantically distinct clips and extract the central frame of each clip as a representative keyframe, effectively compressing temporal information while preserving key visual transitions.

Sample Construction. We group keyframes into coherent sequences within scene boundaries identified via DINOv2 (Oquab et al., 2023) visual feature similarity thresholds, ensuring smooth transitions. A VLM is then applied to generate a descriptive summary and a quality score for each sequence, and low-scoring samples are discarded. The final frame of each valid sequence is designated as the generation target, completing the Temporal dataset.

4 MacroBench

We propose MacroBench, a comprehensive benchmark evaluating in-context generation across Customization, Illustration, Spatial, and Temporal dimensions, supporting up to 10 input images.

Tasks. MacroBench is structured along two dimensions: 1) **Task-level**: covering Customization, Illustration, Spatial, and Temporal tasks; 2) **Input-level**: categorized by input image counts (1–3, 4–5, 6–7, and 8–10). This two-dimensional structure enables fine-grained analysis of how model performance varies across both task complexity and input scale.

Data Curation. Following the pipeline in Section 3, we curate diverse evaluation sources: metadata (humans, objects, scenes, cloths, styles) for Customization; documents for Illustration; objects and panoramas for Spatial; and videos for Temporal. All evaluation data is strictly held out from the training set. We randomly sample 250 input pairs per task and image count category, yielding a total of 4,000 pairs, 1,000 for each task and for each image count.

Judge Model Setting. Following recent benchmarks (Ghosh et al., 2023; Ye et al., 2025b; Wu et al., 2025b; Wei et al., 2025; Liu et al., 2025), we adopt an LLM-as-Judge evaluation paradigm. We find that commonly-used judge models GPT-4.1 (OpenAI, 2024) exhibit degraded evaluation quality when processing multiple reference images, particularly for Spatial tasks requiring spatial reasoning. Through comparison in Section B.3, we select Gemini-3-Flash (DeepMind, 2025a) as the judge model for reliable evaluations across all dimensions and input scales at reasonable cost.

We design task-specific metrics (scored 1–10) to evaluate distinct capabilities: *Customization* assesses input consistency and instruction adherence via Image Consistency Score (ICS) and Prompt Following Score (PFS); *Illustration* measures alignment with context text and input images via Text Consistency Score (TCS) and ICS; *Spatial* evaluates view transformation and content preservation via View Transformation Score (VTS) and Content Consistency Score (CCS); and *Temporal* examines content and temporal coherence via CCS and Image Sequence Consistency Score (ISCS). For Customization, ICS is computed per input image, and the overall ICS is the harmonic mean across inputs to penalize inconsistency with any individual reference. To ensure balanced performance across dimensions, the final score for each task is the geometric mean of its two metrics (e.g., $\sqrt{\text{ICS} \times \text{PFS}}$ for Customization), with the overall MacroBench score averaged across tasks. Prompts and evaluation details are provided in Section B.

Table 1: **Quantitative Comparison Results on MacroBench.** "+ X" denotes the baseline model specifically fine-tuned on the corresponding dataset. Results of "1–3" and "4–5" are combined into "1–5", and "6-7" and "8-10" are combined into "6-10".

Model	Customization		Illustration		Spatial		Temporal		Average↑
	1-5	6-10	1-5	6-10	1-5	6-10	1-5	6-10	
Closed-source models									
Nano Banana Pro (DeepMind, 2025c)	9.41	7.60	**9.12**	**8.88**	3.28	3.20	8.21	7.25	7.12
GPT-Image-1.5 (OpenAI, 2025)	**9.62**	**8.76**	8.90	8.84	**3.32**	**4.24**	**8.48**	**7.84**	**7.50**
Open-source models									
Bagel (Deng et al., 2025)	5.37	2.53	4.51	4.34	0.67	0.53	3.36	2.93	3.03
Bagel + Echo4o (Ye et al., 2025a)	6.95	3.98	4.52	4.14	0.76	0.80	3.19	2.41	3.34
Bagel + MICo (Wei et al., 2025)	6.32	2.84	4.79	4.62	0.81	0.80	3.85	3.53	3.44
OmniGen2 (Wu et al., 2025b)	5.21	2.27	4.72	3.80	0.67	1.01	3.00	2.41	2.89
OmniGen2 + MICo (Wei et al., 2025)	4.97	2.18	4.50	3.98	0.53	0.97	2.81	2.45	2.80
OmniGen2 + OpenSubject (Liu et al., 2025)	5.09	2.34	4.48	3.64	1.06	1.51	3.21	2.62	2.99
Qwen-Image-Edit-2511 (Wu et al., 2025a)	6.41	0.92	4.69	2.17	1.57	0.63	4.02	1.14	2.69
Bagel + Ours	**8.92**	**7.12**	**5.69**	**5.56**	**3.32**	**3.48**	**6.23**	**5.40**	**5.71**
OmniGen2 + Ours	7.30	4.45	4.68	4.25	1.82	1.40	4.97	4.07	4.11
Qwen + Ours	8.31	4.69	5.04	3.49	2.76	2.60	5.03	2.98	4.36

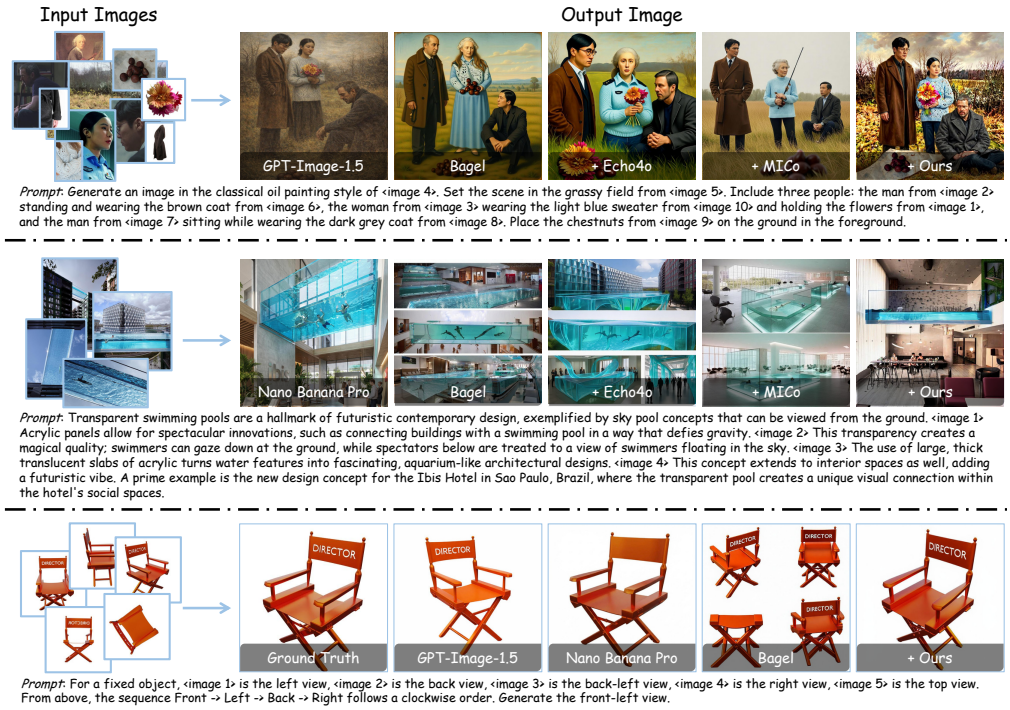

Figure 7: **Qualitative Results.** Results of different datasets are fine-tuned from the same model.

5 Experiments

In this section, we first present experiment settings in Section 5.1, then validate the effectiveness of MacroData on both MacroBench and OmniContext (Wu et al., 2025b) in Section 5.2, then conduct ablation studies in Section 5.3.

5.1 Experimental Settings

Baselines. We fine-tune three open-source in-context generative models on our MacroData: Bagel (Deng et al., 2025), OmniGen2 (Wu et al., 2025b), and Qwen-Image-Edit-2511 (Wu et al., 2025a). All fine-tuning runs also include a portion of T2I data for text-to-image generation to preserve general generation ability. We compare against closed-source models Nano Banana Pro (DeepMind, 2025c) and GPT-Image-1.5 (OpenAI, 2025), as well as models fine-tuned on alternative datasets: Echo4o (Ye et al., 2025a), MICo (Wei et al., 2025), and OpenSubject (Liu et al., 2025). Note that GPT-Image-1.5 imposes a 1,000-token text limit, so only a subset of Illustration tasks can be evaluated; we report the average over successfully evaluated samples.

Table 2: **Results on OmniContext.** "Char. + Obj." denotes Character + Object. Black and gray rows indicate Customization-only and full-dataset training, respectively. "†" marks reported results; others are reproduced under Section 5.1.

Model	SINGLE		MULTIPLE			SCENE			Average↑
	Character	Object	Character	Object	Char. + Obj.	Character	Object	Char. + Obj.	
Closed-source models									
Nano Banana Pro (DeepMind, 2025c)	9.07	9.43	8.91	9.23	8.79	8.94	8.61	8.70	8.96
GPT-Image-1.5 (OpenAI, 2025)	**9.33**	**9.44**	**9.40**	**9.43**	**9.11**	**9.22**	**9.14**	**8.90**	**9.25**
Open-source models									
Bagel (Deng et al., 2025)	7.45	7.40	5.54	6.56	7.17	4.59	5.49	5.95	6.27
Bagel + Echo4o† (Ye et al., 2025a)	-	-	8.07	7.50	**8.29**	**8.62**	8.00	8.08	8.09
Bagel + MICo (Wei et al., 2025)	8.07	8.29	7.24	7.97	7.75	6.08	7.19	6.79	7.42
OmniGen2 (Wu et al., 2025b)	8.22	8.24	7.45	7.30	7.78	7.31	6.47	7.24	7.50
OmniGen2 + MICo (Wei et al., 2025)	8.14	7.76	7.30	7.49	7.91	6.38	6.57	7.14	7.34
OmniGen2 + OpenSubject (Liu et al., 2025)	8.38	7.77	7.29	7.73	7.66	7.16	6.86	7.37	7.53
Qwen-Image-Edit-2511 (Wu et al., 2025a)	8.49	9.10	8.34	**8.71**	8.11	7.22	7.90	7.92	8.22
Bagel + Ours Customization	8.49	8.76	**8.69**	8.17	8.22	8.11	7.87	7.88	8.26
Bagel + Ours All	8.36	8.69	8.02	8.48	7.88	7.37	7.72	7.43	8.00
OmniGen2 + Ours Customization	**8.64**	8.16	8.17	8.11	8.28	8.00	7.62	8.00	8.12
OmniGen2 + Ours All	8.52	8.51	7.52	7.94	8.15	6.59	6.83	7.98	7.75
Qwen + Ours All	8.60	**9.21**	8.55	8.69	8.13	7.75	**8.45**	**8.19**	**8.45**

Table 3: **Cross-task Validation.** "+ Task" denotes training only on the corresponding task subset, while "+ All" uses the complete dataset.

Model	Customization		Illustration		Spatial		Temporal		Average↑
	1-5	6-10	1-5	6-10	1-5	6-10	1-5	6-10	
Bagel	5.37	2.53	4.51	4.34	0.67	0.53	3.36	2.93	3.03
Bagel + All	**8.92**	**7.12**	5.69	5.56	**3.32**	**3.48**	**6.23**	**5.40**	**5.71**
Bagel + Customization	8.61	6.43	5.19	4.92	0.81	0.64	4.43	3.17	4.27
Bagel + Illustration	5.46	2.16	**5.93**	**5.70**	1.21	1.22	5.14	4.59	3.92
Bagel + Spatial	5.08	2.24	4.89	4.78	2.72	2.98	4.99	4.13	3.97
Bagel + Temporal	5.25	1.87	4.96	4.84	1.05	1.07	6.17	5.19	3.80

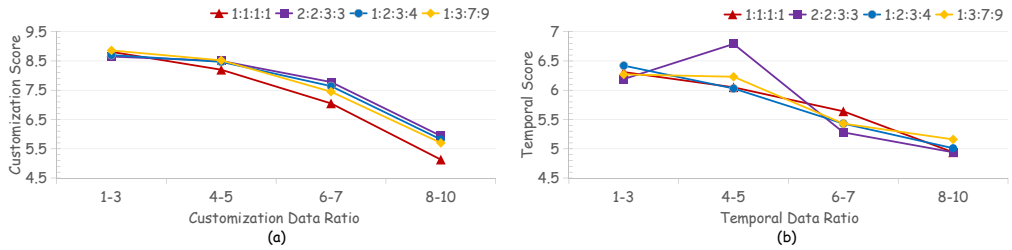

Figure 8: **Impact of Data Ratio** on (a) Customization and (b) Temporal. The categories are 1–3, 4–5, 6–7, and 8–10 references; each curve denotes a different training-data ratio.

Model Configuration. To control sequence length, inputs use 1024×1024, 768×768, and 512×512 resolution for 1–3, 4–5, and 6–10 images, respectively; outputs are capped at 768×768. Bagel (Deng et al., 2025) ViT (Dosovitskiy et al., 2020) inputs are further capped at 336×336.

Benchmarks. We use MacroBench and OmniContext (Wu et al., 2025b) for multi-reference generation and GenEval (Ghosh et al., 2023) for T2I. Open-source models generate at 768×768; closed-source models use their defaults.

5.2 Main Results

MacroBench. As shown in Table 1, MacroData consistently improves all open-source models. Bagel + Ours reaches 5.71, up from 3.03 and ranking third behind only Nano Banana Pro and GPT-Image-1.5. It approaches Nano Banana Pro on Customization, surpasses it on Spatial, and outperforms Echo4o (Ye et al., 2025a), MICo (Wei et al., 2025), and OpenSubject (Liu et al., 2025) under matched architectures.

Furthermore, while increasing input images from 1–5 to 6–10 degrades performance generally, models trained on MacroData exhibit improved robustness. For instance, our fine-tuned Qwen mitigates the base model's Customization drop from 5.49 to 3.62 and Illustration drop from 2.50 to 1.55. MacroData also provides stable gains in challenging Spatial tasks where base models typically score below 1.0. More details are provided in the Appendix.

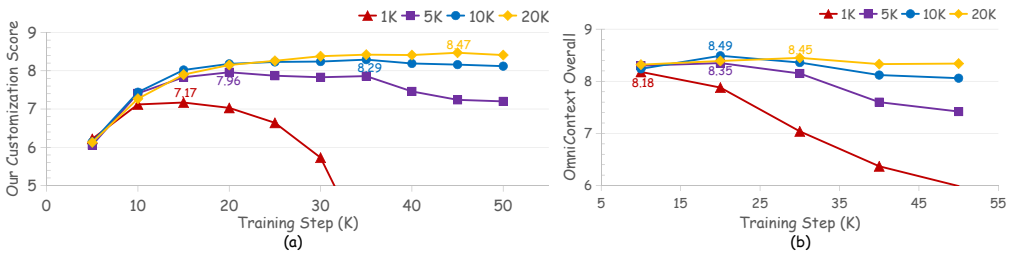

Figure 9: **Data Scaling.** Customization performance on (a) MacroBench and (b) OmniContext (Wu et al., 2025b) across training steps and dataset sizes.

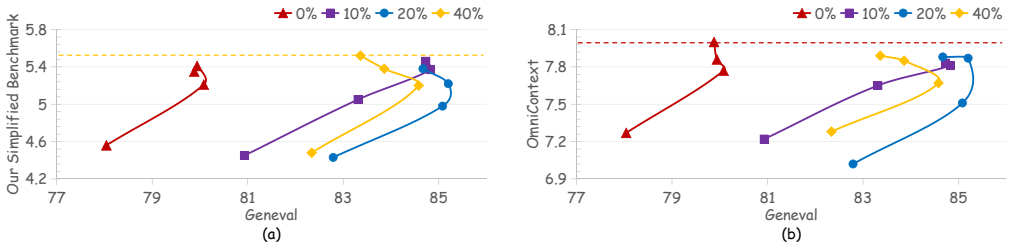

Figure 10: **T2I Data Ratio.** Multi-reference performance on (a) MacroBench and (b) OmniContext (Wu et al., 2025b) versus GenEval (Ghosh et al., 2023) for different T2I training ratios.

Qualitative Results. Figure 7 shows coherent composition of up to 10 references, long-context Illustration following, and accurate Spatial viewpoint synthesis.

OmniContext Benchmark. Following prior work (Ye et al., 2025a; Wei et al., 2025; Liu et al., 2025), we fine-tune Bagel (Deng et al., 2025) and OmniGen2 (Wu et al., 2025b) on full MacroData and its Customization subset, reporting the better reproduced or published baseline for fairness. As shown in Table 2, Bagel + Ours Customization surpasses purpose-built Echo4o (Ye et al., 2025a), scoring 8.26 versus 8.09, while Qwen + Ours All reaches 8.45.

5.3 Ablation Study

Cross-task Validation. We train Bagel on each 100K task subset for one-quarter of the full iterations. Table 3 shows that each subset best improves its matching task, while joint training achieves the best average and confirms complementary supervision. Illustration-only wins its task because it is more text-heavy and semantically distinct.

Progressive vs. Non-Progressive Data Ratios. Figure 8 compares sampling ratios of 1:1:1:1, 2:2:3:3, 1:2:3:4, and 1:3:7:9 over the 1–3, 4–5, 6–7, and 8–10 input bins. Upweighting large-input samples improves high-input Customization without harming low-input scores, but provides no benefit on Temporal. We therefore use 2:2:3:3 for Customization and 1:1:1:1 for the other tasks.

Data Scaling. As shown in Fig. 9, scaling Customization data from 1K through 5K and 10K to 20K improves MacroBench and OmniContext, with the largest gains by 10K and diminishing returns thereafter, while training becomes progressively more stable. We therefore use 100K samples per MacroData task for robust convergence.

Text-to-Image Tradeoff. We train with T2I data ratios of 0%, 10%, 20%, and 40%, then evaluate on GenEval and a representative MacroBench subset containing 50 samples per input count across tasks. As shown in Fig. 10, T2I co-training improves GenEval while retaining strong multi-reference performance, but gains largely saturate beyond 10%. We therefore adopt a 10% ratio to preserve general text-to-image generation without unnecessarily diluting multi-reference supervision.

6 Conclusion

We introduce MacroData, a 400K-sample dataset for multi-reference image generation with up to 10 references across Customization, Illustration, Spatial, and Temporal, together with MacroBench for standardized evaluation across tasks and input scales. Experiments show that training on MacroData consistently improves long-context multi-reference generation, while ablations highlight the benefits of task synergy and practical data-design choices. We hope these resources support future research on in-context generation with complex visual contexts.

AI USE STATEMENT

We used generative AI tools for dataset construction and automated evaluation, as described in Sections 3 and 4. For MacroData, Nano Banana Pro was used to generate target images for the Customization subset; Qwen-3VL-8B, Gemini-3-Pro, and Gemini-3-Flash were used for data filtering and categorization, context reorganization, text generation, and quality assessment across the relevant subsets. For MacroBench, Gemini-3-Flash was used as an automated judge to score generated images. We also used AI assistants for code completion, manuscript polishing, LaTeX formatting, and compilation troubleshooting. All experimental results and manuscript text were reviewed by the authors, and all conclusions were drawn through human analysis. All reported human evaluations were performed by people, not simulated by AI. These include the human ratings used to assess agreement with automated scores. The authors take responsibility for the final manuscript and all AI-assisted content and artifacts.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of our data construction pipelines, preprocessing and filtering procedures, training configurations, and evaluation protocols in the main paper and supplementary material. Our project website and data repository are available for anonymous review.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Karl Audun Kagnes Borgersen, Morten Goodwin, Morten Grundetjern, and Jivitesh Sharma. A Dataset for Adapting Recommender Systems to the Fashion Rental Economy. In *18th ACM Conference on Recommender Systems*, pp. 945–950, Bari Italy, October 2024. ACM. ISBN 9798400705052. doi: 10.1145/3640457.3688174. URL https://dl.acm.org/doi/10.1145/3640457.3688174.

Kaitong Cai, Jusheng Zhang, Jing Yang, Yijia Fan, Pengtao Xie, Jian Wang, and Keze Wang. Flashvlm: Text-guided visual token selection for large multimodal models. *arXiv preprint arXiv:2512.20561*, 2025a.

Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, et al. Mixture of contexts for long video generation. *arXiv preprint arXiv:2508.21058*, 2025b.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-$alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12501–12511, 2025.

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.

Google DeepMind. Gemini 3 flash. https://deepmind.google/models/gemini/flash/, 2025a.

Google DeepMind. Gemini 3 pro. https://deepmind.google/models/gemini/pro/, 2025b.

Google DeepMind. Gemini 3 pro image (nano banana pro). https://deepmind.google/models/gemini-image/pro/, 2025c.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Haoran Feng, Dizhe Zhang, Xiangtai Li, Bo Du, and Lu Qi. Dit360: High-fidelity panoramic image generation via hybrid training. *arXiv preprint arXiv:2510.11712*, 2025.

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.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

Muhammed Kocabas, Chun-Hao P Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J Black. Spec: Seeing people in the wild with an estimated camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11035–11045, 2021.

Black Forest Labs. Flux.1 – official inference repository for flux models. `https://github.com/black-forest-labs/flux`, 2024. Accessed: 2026-02-26.

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.

Mingsong Li, Lin Liu, Hongjun Wang, Haoxing Chen, Xijun Gu, Shizhan Liu, Dong Gong, Junbo Zhao, Zhenzhong Lan, and Jianguo Li. Multiedit: Advancing instruction-based image editing on diverse and challenging tasks. *arXiv preprint arXiv:2509.14638*, 2025a.

Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, et al. Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text. *arXiv preprint arXiv:2406.08418*, 2024.

Xingyang Li, Muyang Li, Tianle Cai, Haocheng Xi, Shuo Yang, Yujun Lin, Lvmin Zhang, Songlin Yang, Jinbo Hu, Kelly Peng, et al. Radial attention: $o(n \log n)$ sparse attention with energy decay for long video generation. *arXiv preprint arXiv:2506.19852*, 2025b.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

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

Yexin Liu, Manyuan Zhang, Yueze Wang, Hongyu Li, Dian Zheng, Weiming Zhang, Changsheng Lu, Xunliang Cai, Yan Feng, Peng Pei, et al. Opensubject: Leveraging video-derived identity and diversity priors for subject-driven image generation and manipulation. *arXiv preprint arXiv:2512.08294*, 2025.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7739–7751, 2025.

OpenAI. Gpt-4-1. `https://openai.com/index/gpt-4-1`, 2024.

OpenAI. Gpt-image-1-5. `https://openai.com/index/new-chatgpt-images-is-here/`, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Karl Pearson. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242, 1895.

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

Polyhaven. Poly haven. `https://polyhaven.com/`, n.d. Accessed September 20, 2026.

Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9914–9925, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8871–8879, 2024.

Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Lmfusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024.

Tomás Soucek and Jakub Lokoc. Transnet v2: An effective deep network architecture for fast shot transition detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11218–11221, 2024.

Charles Spearman. *The proof and measurement of association between two things.* Appleton-Century-Crofts, 1961.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14940–14950, 2025.

Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. Quest: Query-aware sparsity for efficient long-context llm inference. *arXiv preprint arXiv:2406.10774*, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 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*, 2025.

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. *arXiv preprint arXiv:2512.07348*, 2025.

WikiArt. Wikiart. `https://www.wikiart.org/`, n.d. Accessed September 20, 2026.

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.

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

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18682–18692, 2025c.

Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, et al. Sparse videogen: Accelerating video diffusion transformers with spatial-temporal sparsity. *arXiv preprint arXiv:2502.01776*, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024a.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024b.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. *arXiv preprint arXiv:2310.08580*, 2023.

Hengyuan Xu, Wei Cheng, Peng Xing, Yixiao Fang, Shuhan Wu, Rui Wang, Xianfang Zeng, Daxin Jiang, Gang Yu, Xingjun Ma, et al. Withanyone: Towards controllable and id consistent image generation. *arXiv preprint arXiv:2510.14975*, 2025.

Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, et al. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *arXiv preprint arXiv:2508.09987*, 2025a.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025b.

Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9150–9161, 2023.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Yuxing Wei, Lean Wang, Zhiping Xiao, et al. Native sparse attention: Hardware-aligned and natively trainable sparse attention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23078–23097, 2025.

Jintao Zhang, Chendong Xiang, Haofeng Huang, Jia Wei, Haocheng Xi, Jun Zhu, and Jianfei Chen. Spargeattention: Accurate and training-free sparse attention accelerating any model inference. *arXiv preprint arXiv:2502.18137*, 2025.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36:34661–34710, 2023.

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

APPENDIX

A DETAILED DATA CONSTRUCTION PIPELINE

A.1 CUSTOMIZATION SUBSET

For the source collection, we utilize datasets of immense scale to ensure diversity. Specifically, the metadata encompasses over 2 million identities from OpenSubject (Liu et al., 2025), 200,000 object videos classified into 238 categories from MVImgNet (Yu et al., 2023), 10,000 scene videos from DL3DV (Ling et al., 2024), 50,000 images from the Vibrant Clothes Rental Dataset (Borgersen et al., 2024), and 10,000 artworks from WikiArt (WikiArt, n.d.). Furthermore, the human, object, scene, and cloth categories are augmented with data from Echo4o (Ye et al., 2025a). During preprocessing, keyframes for DL3DV (Ling et al., 2024) are extracted by uniformly sampling five frames per video. To prevent identity leakage in clothing data, we explicitly employ Qwen-3VL-8B (Bai et al., 2025) to filter out images containing human faces. For style data from WikiArt, we categorize images using fine-grained "artist-genre-style" tags and restrict selection to a maximum of three images per category. This rigorous pipeline yields a finalized source dataset of 50,000 humans, 50,000 objects, 32,500 scenes, 30,000 cloths, and 29,292 style samples.

During the composition phase, metadata is mixed using a strict sampling ratio of 8:6:3:2:1 across the human, object, scene, cloth, and style categories. We employ Gemini-3-Flash (DeepMind, 2025a) to evaluate combinations, resampling if a set is deemed unreasonable. Valid sets are then processed by Nano Banana Pro (DeepMind, 2025c) for target image generation. Gemini-3-Flash (DeepMind, 2025a) is subsequently utilized to conduct a bidirectional consistency assessment between inputs, prompts, and generated outputs. The final curated 100,000 samples for this task are distributed as 20,000, 20,000, 30,000, and 30,000 across the 1–3, 4–5, 6–7, and 8–10 image number categories, respectively.

A.2 ILLUSTRATION SUBSET

The raw source material from OmniCorpus-CC-210M (Li et al., 2024) contains an overwhelming 210 million interleaved image-text sequences. To efficiently navigate this massive scale and identify optimal "anchor images," we utilize the Qwen-3VL-8B (Bai et al., 2025) model. Following random sampling of candidate targets and their preceding contexts, we deploy Gemini-3-Pro (DeepMind, 2025b) to execute the sample reorganization phase. This model is explicitly instructed to re-evaluate semantic relevance, synthesize a concise textual context, and assign a final quality score for filtering.

The resulting 100,000 high-quality illustration samples are perfectly balanced, with exactly 25,000 samples allocated to each of the four image number categories.

A.3 Spatial Subset

For the outside-in object subtask utilizing the 10-category G-buffer Objaverse dataset (Qiu et al., 2024), we leverage a highly structured multi-view rendering setup. The source provides 24 views within an elevation range of 5° to 30° (at 15° intervals), 12 views between -5° and 5° (at 30° intervals), alongside single top and bottom views. We define a precise canonical set of 10 views: top, bottom, left, right, front, back, and four diagonal perspectives. In this outside-in configuration, the horizontal sequence of front, left, back, and right is arranged in a clockwise order. Input views are meticulously drawn from either the 15° or corresponding 30° rotation sets to guarantee visual overlap. This yields 40,000 samples for the outside-in subtask, with 10,000 samples per image number category.

For the inside-out scene subtask, after filtering non-standard panoramic formats, we explicitly utilize Qwen-3VL-8B (Bai et al., 2025) to classify the valid panoramas into indoor and outdoor categories. Employing the same 10 canonical view definitions as the object subtask, the inside-out horizontal sequence is instead arranged in a counter-clockwise order. Camera initialization involves a random yaw angle coupled with a pitch strictly bounded between -10° and 10°. This pipeline generates 30,000 indoor and 30,000 outdoor samples, with each subset stratified into 7,500 samples per image number category.

A.4 Temporal Subset

The raw temporal data originates from OmniCorpus-YT (Li et al., 2024), which comprises 10 million YouTube videos. Due to the sheer volume, we sample a representative subset of 1 million video links for raw downloading before applying TransNetv2 (Soucek & Lokoc, 2024) for clip segmentation and central keyframe extraction. Following shot boundary identification via DINOv2 (Oquab et al., 2023), we prompt Gemini-3-Flash (DeepMind, 2025a) to act as the evaluator, generating a descriptive summary for each visually continuous sequence and assigning the requisite quality score for filtering. The final target designates the last image of each valid sequence, resulting in 100,000 temporal samples evenly stratified with 25,000 samples per image number category.

A.5 More Visualizations of MacroData

We present additional data samples in Figs. 19 to 22, which separately illustrate different tasks with varying numbers of input references in our MacroData. As shown, the Customization subset encompasses diverse types of inputs, while the Illustration subset covers a wide range of topics. The Spatial subset comprises three primary categories: objects, indoor panorama, and outdoor panorama, each composed of different target viewpoints for prediction. The Temporal subset demonstrates examples with varying sequence lengths.

B Benchmark

B.1 Benchmark Prompt

As shown in Figs. 23 to 26, we provide the complete prompts used for each task in MacroBench. For every query, the judge model receives: (1) the task-specific instruction, (2) all input reference images, and (3) the generated output image. For Spatial and Temporal tasks, the ground-truth target image is additionally provided as an explicit reference to facilitate fine-grained assessment.

B.2 Calculation Details

Single-score Aggregation. All raw metric scores lie in $[0, 10]$. For each sample, the two task-specific metric scores (M_1, M_2) are aggregated into a single scalar via the geometric mean:

$$S = \sqrt{M_1 \times M_2}. \tag{1}$$

Table 4: Cross-judge score correlations on non-GT validation samples. The w/o Macro uses five official models ($N = 400$); the with Macro adds three models fine-tuned on MacroData ($N = 640$).

Scope	Judge Pair	N	Pearson	Spearman	Kendall
w/o Macro	Gemini–GPT-5.4	400	**0.814**	0.811	**0.662**
	Gemini–Opus	400	0.800	0.810	0.645
	Gemini–GPT-4.1	400	0.693	0.721	0.564
	GPT-5.4–Opus	400	0.809	**0.816**	0.656
	GPT-5.4–GPT-4.1	400	0.779	0.803	0.645
	Opus–GPT-4.1	400	0.765	0.773	0.616
with Macro	Gemini–GPT-5.4	640	**0.800**	**0.807**	**0.651**
	Gemini–Opus	640	0.775	0.797	0.628
	Gemini–GPT-4.1	640	0.658	0.697	0.534
	GPT-5.4–Opus	640	0.793	0.800	0.639
	GPT-5.4–GPT-4.1	640	0.758	0.774	0.614
	Opus–GPT-4.1	640	0.743	0.753	0.593

Table 5: Model ranking consistency across AI judges and human scores. Human scores are available only for the five official models. Parentheses denote the rank under the corresponding judge.

Model	Gemini	GPT-5.4	Opus	GPT-4.1	Human
GPT-Image-1.5	$7.40_{(1)}$	$6.87_{(1)}$	$7.08_{(1)}$	$9.50_{(1)}$	$6.87_{(2)}$
Nano Banana Pro	$7.34_{(2)}$	$6.77_{(2)}$	$6.81_{(2)}$	$8.99_{(2)}$	$7.22_{(1)}$
Bagel + Ours	$6.42_{(3)}$	$6.20_{(3)}$	$6.39_{(3)}$	$7.94_{(3)}$	–
OmniGen2 + Ours	$4.40_{(4)}$	$4.65_{(5)}$	$5.30_{(4)}$	$6.70_{(5)}$	–
Qwen + Ours	$4.31_{(5)}$	$4.74_{(4)}$	$5.17_{(5)}$	$7.31_{(4)}$	–
Bagel	$3.04_{(6)}$	$3.50_{(6)}$	$4.61_{(6)}$	$6.00_{(6)}$	$4.51_{(3)}$
Qwen-Image-Edit-2511	$2.56_{(7)}$	$2.51_{(8)}$	$3.80_{(8)}$	$4.53_{(8)}$	$3.30_{(4)}$
OmniGen2	$2.40_{(8)}$	$2.97_{(7)}$	$4.22_{(7)}$	$5.35_{(7)}$	$3.08_{(5)}$

The geometric mean is chosen over the arithmetic mean because it penalizes severe failure on either dimension more aggressively: if one metric collapses to zero, the overall score also collapses, regardless of how high the other metric is.

The metric pairs for each task are:

- **Customization:** $S = \sqrt{\mathrm{ICS} \times \mathrm{PFS}}$, where ICS is the harmonic mean of per-reference Image Consistency Scores.
- **Illustration:** $S = \sqrt{\mathrm{TCS} \times \mathrm{ICS}}$.
- **Spatial:** $S = \sqrt{\mathrm{VTS} \times \mathrm{CCS}}$.
- **Temporal:** $S = \sqrt{\mathrm{CCS} \times \mathrm{ISCS}}$.

Harmonic Mean for Customization ICS. Because Customization requires faithfully reproducing every individual reference subject, ICS is first computed for each reference image and then aggregated via the harmonic mean:

$$\mathrm{ICS} = \frac{n}{\sum_{i=1}^{n} \frac{1}{\mathrm{ICS}_i}}, \tag{2}$$

where n is the number of input reference images. The harmonic mean ensures that low fidelity to *any single* reference significantly suppresses the overall ICS.

Overall MacroBench Score. The overall MacroBench score for a model is the arithmetic mean of its per-task scores across the four tasks.

Table 6: Correlation with human scores. Human-only uses 400 human-annotated non-GT samples. Human+GT appends 120 GT samples whose output equals the target image and score is 10.

Method	Human-only				Human+GT			
	N	Pearson	Spearman	Kendall	N	Pearson	Spearman	Kendall
Gemini	400	**0.833**	**0.844**	**0.689**	520	**0.874**	**0.867**	**0.730**
GPT-5.4	400	0.762	0.773	0.609	520	0.802	0.809	0.655
Opus	400	0.720	0.730	0.552	520	0.779	0.786	0.614
GPT-4.1	400	0.599	0.635	0.471	520	0.652	0.700	0.547
DINOv2	400	0.185	0.190	0.131	520	0.530	0.623	0.453
CLIP-I	400	0.021	0.021	0.016	520	0.426	0.540	0.382

B.3 Validation Consistency

To validate the reliability of the evaluation protocol, we conduct a larger consistency study from three complementary views: cross-judge score agreement, model-ranking agreement, and judge–human agreement. All judge scores are converted into the same single score using Eq. (1).

Validation Samples. We use two non-GT validation scopes. The *w/o Macro* scope contains 400 model-generated samples from five official models: Nano Banana Pro (DeepMind, 2025c), GPT-Image-1.5 (OpenAI, 2025), Bagel (Deng et al., 2025), OmniGen2 (Wu et al., 2025b), and Qwen-Image-Edit-2511 (Wu et al., 2025a). For each model, we sample five cases for every task and image-count category, yielding $5 \times 4 \times 4 \times 5 = 400$ samples. The *with Macro* scope further adds Bagel + Ours, OmniGen2 + Ours, and Qwen + Ours with the same case indices, yielding $8 \times 4 \times 4 \times 5 = 640$ samples. For judge–human analysis, the 400 official-model samples are annotated by humans. We also report a Human+GT setting that appends 120 ground-truth samples from Illustration, Spatial, and Temporal tasks, where the output is the target image and the human score is set to 10.

Cross-judge Score Agreement. We independently score the validation samples with Gemini-3-Flash (DeepMind, 2025a), GPT-5.4, Opus-4.6, and GPT-4.1 (OpenAI, 2024), and report Pearson (r) (Pearson, 1895), Spearman (ρ) (Spearman, 1961), and Kendall (τ) (Kendall, 1938) correlations in Table 4. The three stronger judges (Gemini, GPT-5.4, and Opus) show highly consistent overall scores, with Pearson correlations around 0.78–0.81 across both validation scopes. GPT-4.1 has a higher absolute-score bias and lower pairwise correlations, but still preserves medium-to-high agreement with the other judges.

Model-ranking Agreement. Beyond sample-level correlations, Table 5 compares the model ranking induced by each judge. The top three models are identical across all four AI judges: GPT-Image-1.5, Nano Banana Pro, and Bagel + Ours. The macro models consistently outperform their corresponding official backbones under every judge. Minor swaps only occur between close middle-ranked models (OmniGen2 + Ours and Qwen + Ours) and between the two weakest official models. The human ranking is available for the five official models only; it agrees with the AI judges on the top-two and bottom-three groups, with only the first two models swapped.

Judge–human Agreement. Table 6 shows that Gemini-3-Flash has the strongest overall agreement with human scores among the VLM judges, including traditional metrics like CLIP-I (Radford et al., 2021) and DINOv2 (Oquab et al., 2023) as introduced below. On the Human-only setting, Gemini reaches Pearson/Spearman correlations of 0.833/0.844, followed by GPT-5.4 and Opus. Adding GT anchors increases the correlations of all methods, given the clear high-score reference points provided by GT outputs.

B.4 Traditional Metric Analysis

We additionally report CLIP-I (Radford et al., 2021) and DINOv2 (Oquab et al., 2023) image-similarity metrics on MacroBench. CLIP-I uses CLIP ViT-L/14 image features and cosine similarity; DINOv2 uses ViT-L/14 features and cosine similarity. For Customization, the output is compared against each input reference and aggregated by the harmonic mean. For Illustration, Spatial, and

Table 7: Per-task Pearson correlation with human scores. The Human-only setting is more diagnostic for non-GT model outputs; Human+GT additionally includes exact target-image anchors for Illustration, Spatial, and Temporal.

Setting	Task	Gemini	GPT-5.4	Opus	GPT-4.1	CLIP-I	DINOv2
Human-only	Customization	**0.894**	0.847	0.884	0.826	0.344	0.384
	Illustration	0.755	**0.785**	0.748	0.694	0.467	0.548
	Spatial	**0.611**	0.507	0.368	0.432	0.265	0.312
	Temporal	0.791	0.807	**0.816**	0.786	0.407	0.222
Human+GT	Customization	**0.894**	0.847	0.884	0.826	0.344	0.384
	Illustration	**0.827**	0.748	0.791	0.717	0.706	0.745
	Spatial	**0.914**	0.874	0.839	0.598	0.771	0.770
	Temporal	**0.816**	0.813	0.802	0.765	0.674	0.558

Table 8: CLIP-I and DINOv2 results summarized by task. Avg is the macro-average over four tasks.

Model	CLIP-I					DINOv2				
	Cust.	Illus.	Spat.	Temp.	Avg	Cust.	Illus.	Spat.	Temp.	Avg
Nano Banana Pro	0.764	0.850	0.919	0.899	**0.858**	0.602	0.725	0.794	**0.820**	0.735
GPT-Image-1.5	**0.769**	**0.854**	0.913	0.885	0.855	**0.614**	**0.756**	0.788	0.813	0.743
Bagel	0.762	0.816	0.885	0.867	0.832	0.601	0.720	0.732	0.795	0.712
Bagel + Echo4o	0.767	0.802	0.879	0.864	0.828	**0.614**	0.708	0.750	0.793	0.716
Bagel + MICo	0.754	0.820	0.893	0.870	0.834	0.596	0.716	0.740	0.798	0.713
OmniGen2	0.748	0.822	0.898	0.862	0.833	0.594	0.716	0.750	0.792	0.713
OmniGen2 + MICo	0.734	0.827	0.905	0.862	0.832	0.581	0.710	0.759	0.786	0.709
OmniGen2 + OpenSubject	0.759	0.829	0.905	0.868	0.840	0.604	0.715	0.767	0.796	0.721
Qwen-Image-Edit-2511	0.766	0.785	0.882	0.843	0.819	0.592	0.638	0.725	0.754	0.677
Bagel + Ours	0.762	0.837	**0.935**	0.888	0.855	0.604	0.733	**0.837**	0.813	**0.747**
OmniGen2 + Ours	0.751	0.815	0.901	0.884	0.838	0.596	0.707	0.743	0.799	0.711
Qwen + Ours	0.748	0.807	0.932	0.867	0.839	0.595	0.692	0.815	0.781	0.720

Temporal, the output is compared against the target image. Table 8 reports task-wise summaries, while Tables 9 and 10 provide the full task-by-image-count grids for all evaluated models.

The key observation is that both embedding metrics have a narrow dynamic range. For example, the CLIP-I average varies only from 0.819 to 0.858 across the closed-source and open-source models, while DINOv2 varies from 0.677 to 0.747. These similarities often reward low-level visual or semantic overlap but do not reliably capture multi-reference instruction following, identity binding, viewpoint reasoning, or temporal consistency. This explains their much weaker correlation with human judgments in Tables 6 and 7.

C Experiments Details

We first detail the training and evaluation settings, then report detailed quantitative and qualitative results, and finally present additional ablations and exploratory studies.

C.1 Training Settings

We display detailed hyperparameters for model training.

Bagel. Bagel (Deng et al., 2025) is fine-tuned using Fully Sharded Data Parallel (FSDP) across 32 NVIDIA H800 GPUs (4 nodes) with a frozen ViT encoder. Training utilizes a token-based dynamic batching strategy with a maximum of 32,768 tokens and a learning rate of 2×10^{-5}.

OmniGen2. OmniGen2 (Wu et al., 2025b) is fine-tuned using DeepSpeed across 32 NVIDIA H800 GPUs (4 nodes). Training runs with a global batch size of 64 and a learning rate of 8×10^{-7}. The baseline of OmniGen2 (Wu et al., 2025b) adopts a fixed image embedding for max 5 images. To

Table 9: CLIP-I results on the full task-by-image-count grid. Avg is the average over all cells.

Model	Customization				Illustration				Spatial				Temporal				Avg
	1–3	4–5	6–7	8–10	1–3	4–5	6–7	8–10	1–3	4–5	6–7	8–10	1–3	4–5	6–7	8–10	
Nano Banana Pro	.828	.759	.740	.727	.855	.852	.851	.841	.924	.921	.916	.913	.905	.893	.895	.900	.858
GPT-Image-1.5	.831	.761	.746	.738	.860	.849	.855	.854	.908	.913	.915	.915	.889	.882	.884	.883	.855
Bagel	.813	.753	.742	.738	.802	.820	.823	.818	.886	.882	.885	.887	.866	.867	.869	.864	.832
Bagel + Echo4o	.818	.762	.748	.741	.793	.800	.815	.800	.871	.879	.885	.882	.866	.862	.866	.863	.828
Bagel + MICo	.812	.749	.733	.725	.809	.824	.830	.816	.899	.887	.892	.892	.869	.869	.874	.870	.834
OmniGen2	.801	.738	.729	.725	.819	.824	.828	.817	.908	.899	.895	.891	.866	.857	.861	.866	.833
OmniGen2 + MICo	.790	.730	.716	.702	.825	.830	.834	.821	.907	.904	.906	.903	.862	.856	.864	.867	.832
OmniGen2 + OpenSubject	.809	.752	.738	.737	.827	.831	.835	.822	.903	.901	.907	.907	.868	.862	.867	.873	.840
Qwen-Image-Edit	.818	.766	.742	.739	.818	.800	.771	.750	.906	.904	.870	.850	.871	.861	.839	.802	.819
Bagel + Ours	.818	.757	.741	.732	.832	.840	.841	.835	.932	.933	.935	.938	.884	.890	.891	.885	.855
OmniGen2 + Ours	.801	.743	.733	.728	.804	.821	.818	.819	.908	.919	.891	.885	.880	.884	.886	.886	.838
Qwen + Ours	.805	.745	.726	.714	.814	.821	.812	.782	.933	.934	.931	.931	.883	.874	.865	.846	.839

support more than 5 input references, we extend the image embedding and randomly initialize them according to a normal distribution.

Qwen-Image-Edit-2511. We fine-tune the DiT component of Qwen-Image-Edit-2511 (Wu et al., 2025a) using DeepSpeed across 128 NVIDIA H800 GPUs (16 nodes). Training uses a learning rate of 1×10^{-5}.

C.2 Evaluation Settings

Apart from using a dynamic resolution strategy to manage context length for varying numbers of input references, we generally follow the default inference settings provided by the baselines (Deng et al., 2025; Wu et al., 2025b;a). Specifically, for OmniGen2 (Wu et al., 2025b), we expand the image embeddings from 5 to 10 using normal initialization to accommodate a maximum of 10 inputs.

C.3 More Ablation Studies

Dynamic Resolution. We compare our dynamic input-resolution schedule—1024px for 1–3 references, 768px for 4–5, and 512px for 6–10—against uniform 1024px inputs, while fixing the output resolution, random seed, and Gemini-3-Flash judge. As shown in Table 11, the overall averages are nearly unchanged (2.76 vs. 2.79 for Bagel and 5.59 vs. 5.56 for Bagel + Ours), and both settings exhibit similar degradation at high input counts. Thus, this degradation primarily reflects the multi-reference reasoning challenge rather than reduced input resolution.

C.4 Exploration on Potential Techniques

Token Selection Strategies. Adding more input images linearly expands the token sequence, introducing long-context inefficiencies similar to those in LLMs (Tang et al., 2024; Zhang et al., 2023; Xiao et al., 2023) and video generation (Cai et al., 2025b; Xi et al., 2025; Li et al., 2025b). A natural

Table 10: DINOv2 results on the full task-by-image-count grid. Avg is the average over all cells.

Model	Customization				Illustration				Spatial				Temporal				Avg
	1–3	4–5	6–7	8–10	1–3	4–5	6–7	8–10	1–3	4–5	6–7	8–10	1–3	4–5	6–7	8–10	
Nano Banana Pro	.669	.600	.576	.561	.734	.738	.718	.708	.805	.799	.787	.783	.821	.811	.809	.838	.735
GPT-Image-1.5	.684	.611	.587	.573	.761	.763	.755	.744	.774	.791	.788	.798	.817	.805	.807	.823	.743
Bagel	.663	.596	.578	.566	.709	.727	.725	.718	.734	.734	.732	.729	.791	.785	.793	.809	.712
Bagel + Echo4o	.676	.612	.591	.578	.704	.716	.715	.699	.740	.752	.757	.751	.790	.784	.788	.810	.716
Bagel + MICo	.658	.592	.574	.561	.712	.726	.722	.703	.750	.735	.736	.738	.791	.790	.798	.814	.713
OmniGen2	.655	.589	.572	.561	.718	.725	.725	.696	.781	.768	.728	.722	.786	.778	.791	.814	.713
OmniGen2 + MICo	.636	.577	.562	.550	.710	.725	.716	.691	.783	.765	.751	.740	.778	.766	.787	.813	.709
OmniGen2 + OpenSubject	.666	.603	.580	.569	.717	.727	.724	.691	.775	.774	.761	.759	.788	.782	.796	.819	.721
Qwen-Image-Edit	.662	.606	.557	.545	.699	.650	.619	.582	.754	.759	.707	.680	.797	.777	.738	.704	.677
Bagel + Ours	.665	.601	.582	.566	.724	.743	.737	.727	.835	.833	.837	.845	.809	.803	.816	.825	.747
OmniGen2 + Ours	.654	.594	.575	.562	.706	.722	.707	.693	.768	.781	.715	.709	.787	.791	.801	.818	.711
Qwen + Ours	.657	.594	.572	.556	.718	.716	.698	.634	.812	.819	.813	.815	.794	.773	.787	.770	.720

Table 11: **Dynamic-resolution ablation.** Average MacroBench scores by task and input-count bin.

Model	Res.	Customization				Illustration				Spatial				Temporal				Avg
		1–3	4–5	6–7	8–10	1–3	4–5	6–7	8–10	1–3	4–5	6–7	8–10	1–3	4–5	6–7	8–10	
Bagel	Dynamic	6.65	3.07	1.38	0.47	4.66	4.35	4.54	4.14	0.87	0.47	0.53	0.53	3.90	2.82	3.13	2.72	2.76
	1024px	6.65	3.16	1.41	0.55	4.66	4.41	4.59	4.43	0.87	0.49	0.45	0.53	3.90	2.93	3.03	2.53	2.79
Bagel + Ours	Dynamic	8.97	8.76	7.51	4.85	5.78	5.59	5.50	5.62	3.40	3.24	3.21	3.74	6.34	6.12	5.64	5.15	5.59
	1024px	8.97	8.72	7.19	5.01	5.78	5.66	5.64	5.56	3.40	2.88	3.33	3.41	6.34	6.10	5.78	5.15	5.56

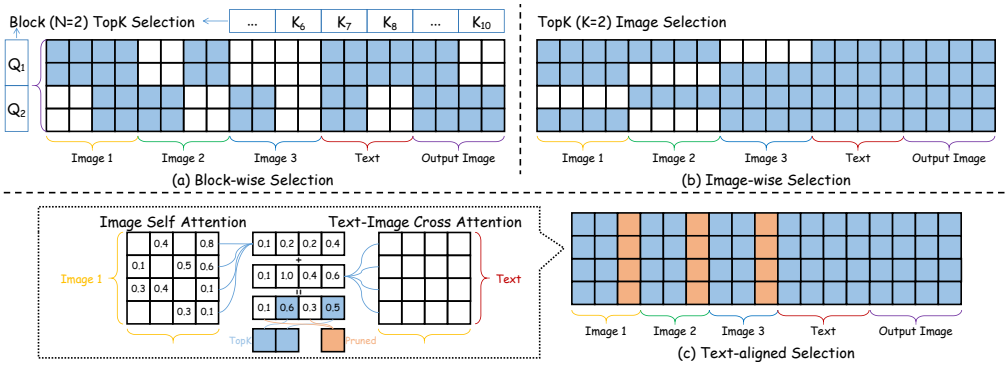

Figure 11: **Token Selection Strategies.** (a) Block-wise Selection, (b) Image-wise Selection, (c) Text-aligned Selection.

solution is to filter the context token sequence to retain only key tokens during attention calculation, referred to as token selection. Guiding the model to focus on the most informative tokens rather than the entire sequence, it effectively mitigates computational overhead and avoids being distracted by useless information. We evaluate three strategies as illustrated in Fig. 11:

Block-wise: Retains top-K token blocks based on mean query-key attention scores among blocks, also known as block sparse attention (Yuan et al., 2025; Zhang et al., 2025; Tang et al., 2024). *Image-wise:* Selects top-K images per token during the diffusion process via attention scores between each output query and the mean of image keys. *Text-aligned:* Selects top-K tokens per image during prefilling via text-image and image-image attention scores, discarding or pruning the rest (Cai et al., 2025a).

Table 12 presents Customization results on MacroBench (retention ratios for block-wise and text-aligned; image counts for image-wise) . Image-wise selection underperforms the baseline, revealing information loss and emphasizing the necessity of cross-reference interactions. Conversely, block-wise and text-aligned strategies outperform the baseline by effectively capturing crucial multi-reference information. Text-aligned selection excels even at low retention ratios, with pruning further boosting performance. Applying selection only on VAE or ViT tokens reveals a trade-off: retaining VAE tokens is crucial given few images, whereas retaining ViT tokens becomes increasingly beneficial as image counts grow.

To further illustrate the differences among the token selection strategies, we present quantitative results for the three selection strategies in Fig. 13. Specifically, the results employ a retention rate of 90% for block-wise, 50% for image-wise, and 30% for text-aligned selection.

As shown, block-wise selection preserves high-quality generation; however, due to the dynamic dropping of certain tokens, the generated images occasionally appear unnatural, as evidenced by the fourth case. Image-wise selection primarily suffers from the degradation of cross-reference blending. Constrained by the retention rate, the generation tends to focus on specific images, particularly when the number of input images is small, as demonstrated in the first, second, and third cases, ultimately leading to failures in coherent blending. For instance, the woman's face disappears in the first case, the identity of the second individual is lost in the second case, and the jacket is not generated appropriately in the third case.

In contrast, despite retaining only 30% of the tokens, text-aligned selection effectively preserves the most critical information, achieving competitive results across all cases. Nevertheless, due to the selective discarding of certain tokens, some fine-grained details are missing, including the flowers in the fifth case and the paintings in the sixth case.

Table 12: **Selection Comparison.**

Settings	1-3	4-5	6-7	8-10	Avg↑
Bagel + Ours	9.00	8.84	8.01	6.23	8.02
Block-wise					
50%	7.23	8.49	7.52	5.99	7.31
80%	8.92	8.88	8.05	6.60	8.11
90%	9.04	9.06	8.19	6.54	**8.21**
Image-wise					
30% Image	8.20	7.99	6.83	5.49	7.13
50% Image	8.32	7.95	7.40	5.99	7.42
w/ only VAE	8.63	8.39	7.62	6.04	**7.67**
w/ only ViT	8.98	8.43	7.25	5.40	7.52
Text-aligned					
80%	9.02	8.78	7.94	6.29	8.01
50%	9.13	9.04	8.00	6.36	8.13
30%	9.03	9.02	7.90	6.48	8.11
w/ only VAE	9.04	8.84	8.01	6.82	**8.18**
w/ only ViT	9.11	8.89	8.13	6.54	8.17
w/ 10% pruning	9.02	9.02	8.20	6.39	8.16

Table 13: **Think & Collage Results.**

Model	1-3	4-5	6-7	8-10	Avg↑
Bagel	6.78	3.96	2.87	2.19	3.95
Bagel + Ours	9.00	8.84	8.01	6.23	8.02
w/ Think					
Bagel	6.66	4.20	2.72	1.98	3.89
Bagel + Ours	6.04	4.50	3.10	2.40	4.01
w/ Collage					
Bagel	5.84	2.79	2.76	2.08	3.37
Bagel + Ours	7.73	6.11	4.96	3.59	5.60



Figure 12: **Think & Collage Method.**

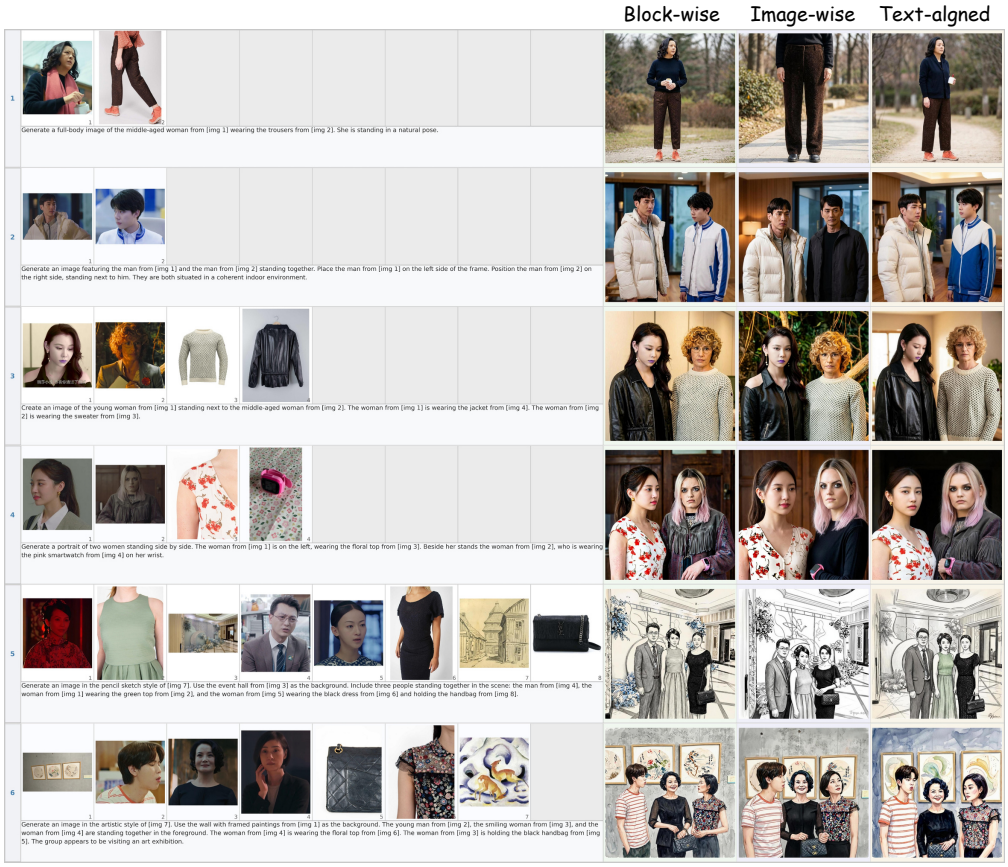

Figure 13: **Qualitative Results** of different token selection strategies.

Think Before Generation. We also examine the "think-before-generation" strategy (Guo et al., 2025), where models generate reasoning text before the final image, as depicted in Fig. 12 (a). However, this approach underperforms no-thinking baselines on multi-reference tasks as shown in Table 13. We hypothesize that without explicit training for multi-reference reasoning, models inherently struggle to synthesize information across multiple images, leading to suboptimal results.

Collage as the Proxy. To bypass input capacity limits, a common workaround is using a collage: stitching multiple reference images into a single spatial grid as a proxy for multi-reference inputs, as displayed in Fig. 12 (b). We evaluate this on Bagel (Deng et al., 2025) by packing inputs into one collage and appending positional descriptions "From left to right... are ¡image 1¿ to ¡image n¿." to the text prompt. Similar to the thinking strategy, collaging underperforms baselines (Table 13). We attribute this performance drop to a loss of detail, where compressing multiple images into a single collage limits the resolution of each component image.

C.5 Detailed Quantitative and Qualitative Results on MacroBench

We present detailed results on our MacroBench in Tables 14 to 17, reporting the scores across different image count categories for each task. As demonstrated, open-source models exhibit a pronounced performance degradation when conditioned on more than three reference images. Notably, Qwen-Image-Edit-2511 (Wu et al., 2025a) suffers a catastrophic drop from 8.27 (1–3 references) to 4.55 (4–5 references) in the Customization task.

Furthermore, these detailed tables reveal that current open-source models and existing datasets (Ye et al., 2025a; Wei et al., 2025; Liu et al., 2025) struggle to support long-context multi-reference generation, even on their specifically targeted Customization tasks (Table 14). For instance, although MICo (Wei et al., 2025) advocates for multi-reference generation, its training data—constructed via simple decomposition and recomposition—fails to provide the robust inter-reference reasoning capabilities necessary for models to process a large number of input images effectively.

Finally, the detailed metrics highlight a distinct contrast between progressive and non-progressive tasks. As shown, metrics for the progressive Customization task exhibit a clear descending trend as the number of input images increases. Conversely, performance on non-progressive tasks (Illustration, Spatial, and Temporal) remains comparatively stable across all image count categories for most baseline models. This pronounced degradation at higher image counts further underscores the necessity to increase the volume of 6–10 image samples, as implemented during our construction.

We display the generated results for each task in MacroBench, produced by Bagel (Deng et al., 2025) fine-tuned on our MacroData, as illustrated in Figs. 14 to 17. In the *Customization* task, the model demonstrates outstanding performance in composing multiple reference images into a single coherent and natural scene, as shown in Fig. 14, even when presented with more than 8 input images. In the *Illustration* task, the model further exhibits the capability to generate complementary images that enrich and enhance interleaved textual descriptions. For the *Spatial* task, the model demonstrates a strong understanding of 3D spatial relationships, successfully synthesizing novel views from specified viewpoints. Finally, in the *Temporal* task, the model effectively captures the transformation patterns across video frames and generates plausible future scenes.

D Failure Cases

We present failure cases of Bagel (Deng et al., 2025) fine-tuned on our MacroData in Fig. 18, including blue input images as references, instructions, green ground truths (no GT is provided for customization), and red generated failure results.

Customization. primarily suffers from the reference number problem. As shown in Fig. 18(a), when too many input images are provided, the model tends to "forget" certain references. For instance, the fourth woman in this case disappears, and an output with only three humans instead of the required four is generated. This consequently leads to the disappearance of clothing associated with the omitted identity, and introduces identity mixing to some extent, like the hair bundle of the fourth woman is incorrectly blended onto the second woman.

Illustration. exhibits problems related to long-context information gathering and text rendering. As displayed in Fig. 18(b), the original context requires generating accessories that appeared in the first

Table 14: **Detailed Results on Customization Task.** Columns 1–3, 4–5, 6–7, 8–10 denote difficulty bins; Avg is the macro-average over all bins.

Model	1–3	4–5	6–7	8–10	Avg↑
Closed-source models					
Nano Banana Pro (DeepMind, 2025c)	9.58	9.24	8.18	7.01	8.50
GPT-Image-1.5 (OpenAI, 2025)	9.67	9.57	9.30	8.23	9.19
Open-source models					
Bagel (Deng et al., 2025)	6.78	3.96	2.87	2.19	3.95
Bagel + Echo4o (Ye et al., 2025a)	7.89	6.01	4.71	3.24	5.46
Bagel + MICo (Wei et al., 2025)	7.67	4.97	3.29	2.39	4.58
OmniGen2 (Wu et al., 2025b)	6.66	3.76	2.55	1.99	3.74
OmniGen2 + MICo (Wei et al., 2025)	6.19	3.74	2.40	1.95	3.57
OmniGen2 + OpenSubject (Liu et al., 2025)	6.47	3.70	2.54	2.13	3.71
Qwen-Image-Edit-2511 (Wu et al., 2025a)	8.27	4.55	1.15	0.69	3.67
Bagel + Ours	**9.00**	**8.84**	**8.01**	**6.23**	**8.02**
OmniGen2 + Ours	7.71	6.88	5.17	3.73	5.87
Qwen + Ours	8.77	7.85	5.89	3.49	6.50

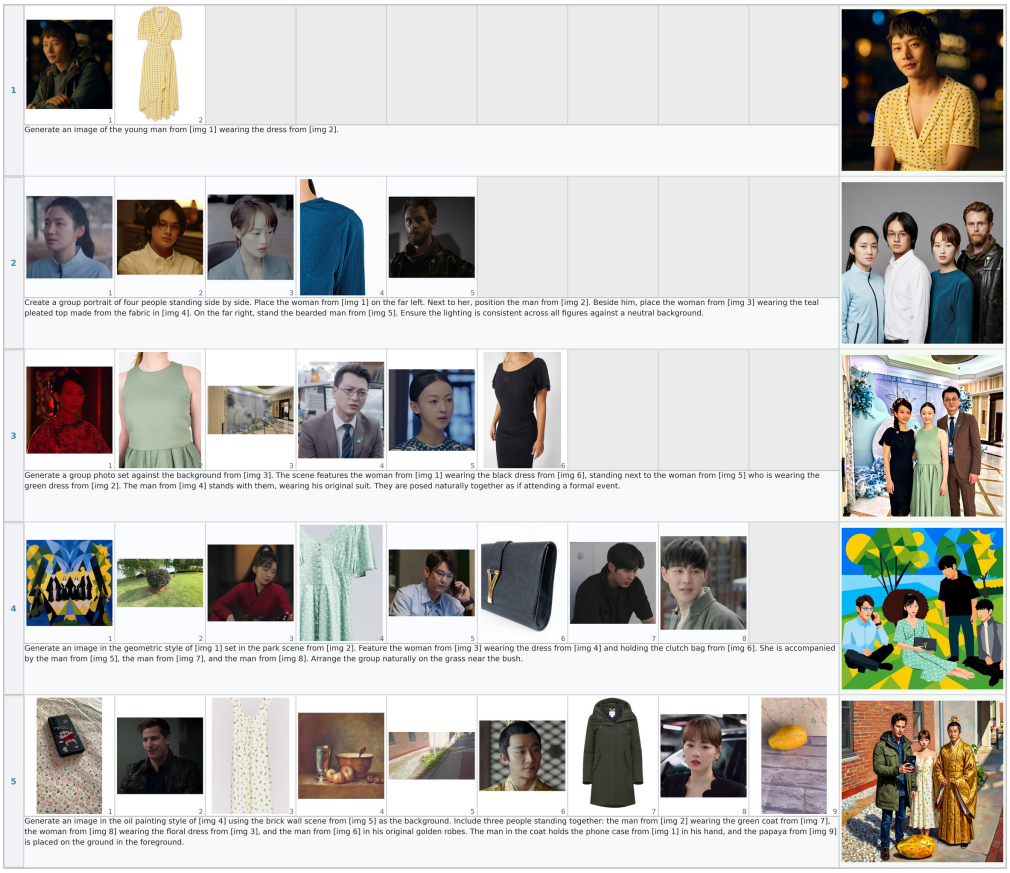

Figure 14: **Qualitative Results** of MacroBench Customization tasks for Bagel (Deng et al., 2025) fine-tuned on MacroData.

Table 15: **Detailed Results on Illustration Task.** Columns 1–3, 4–5, 6–7, 8–10 denote difficulty bins; Avg is the macro-average over all bins.

Model	1–3	4–5	6–7	8–10	Avg↑
Closed-source models					
Nano Banana Pro (DeepMind, 2025c)	9.39	8.85	8.87	8.89	9.00
GPT-Image-1.5 (OpenAI, 2025)	9.41	8.39	8.77	8.91	8.87
Open-source models					
Bagel (Deng et al., 2025)	4.66	4.35	4.54	4.14	4.42
Bagel + Echo4o (Ye et al., 2025a)	4.71	4.32	4.14	4.13	4.33
Bagel + MICo (Wei et al., 2025)	4.81	4.77	4.68	4.56	4.70
OmniGen2 (Wu et al., 2025b)	5.08	4.36	4.02	3.58	4.26
OmniGen2 + MICo (Wei et al., 2025)	4.78	4.21	4.20	3.76	4.24
OmniGen2 + OpenSubject (Liu et al., 2025)	4.90	4.06	3.77	3.50	4.06
Qwen-Image-Edit-2511 (Wu et al., 2025a)	**5.79**	3.58	2.68	1.65	3.43
Bagel + Ours	5.78	**5.59**	**5.50**	**5.62**	**5.62**
OmniGen2 + Ours	4.61	4.74	4.18	4.31	4.46
Qwen + Ours	5.46	4.62	3.93	3.04	4.26

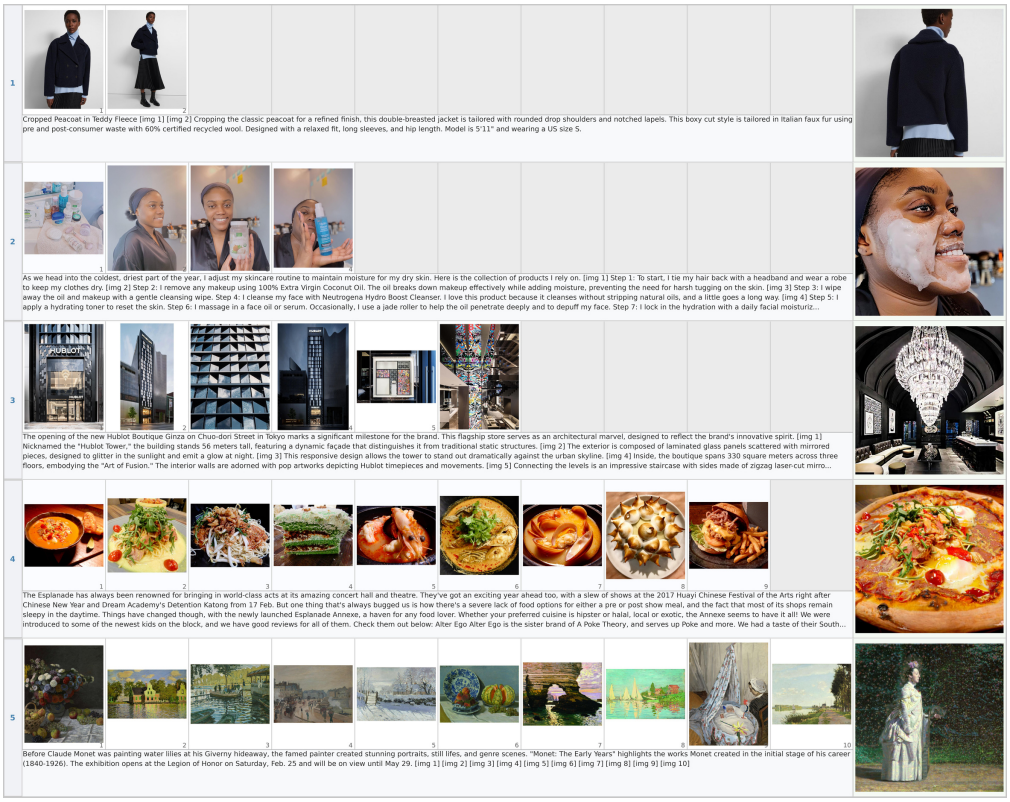

Figure 15: **Qualitative Results** of MacroBench Illustration tasks for Bagel (Deng et al., 2025) fine-tuned on MacroData.

Table 16: **Detailed Results on Spatial Task.** Columns 1–3, 4–5, 6–7, 8–10 denote difficulty bins; Avg is the macro-average over all bins.

Model	1–3	4–5	6–7	8–10	Avg↑
Closed-source models					
Nano Banana Pro (DeepMind, 2025c)	3.09	3.47	3.34	3.05	3.24
GPT-Image-1.5 (OpenAI, 2025)	3.01	3.63	4.25	4.23	3.78
Open-source models					
Bagel (Deng et al., 2025)	0.87	0.47	0.53	0.53	0.60
Bagel + Echo4o (Ye et al., 2025a)	0.88	0.63	0.90	0.70	0.78
Bagel + MICo (Wei et al., 2025)	0.97	0.64	0.87	0.73	0.80
OmniGen2 (Wu et al., 2025b)	0.61	0.72	0.94	1.08	0.84
OmniGen2 + MICo (Wei et al., 2025)	0.51	0.55	0.93	1.01	0.75
OmniGen2 + OpenSubject (Liu et al., 2025)	1.07	1.04	1.52	1.50	1.28
Qwen-Image-Edit-2511 (Wu et al., 2025a)	1.91	1.23	0.73	0.52	1.10
Bagel + Ours	**3.40**	**3.24**	**3.21**	**3.74**	**3.40**
OmniGen2 + Ours	1.65	1.98	1.32	1.47	1.60
Qwen + Ours	2.82	2.69	2.57	2.63	2.68

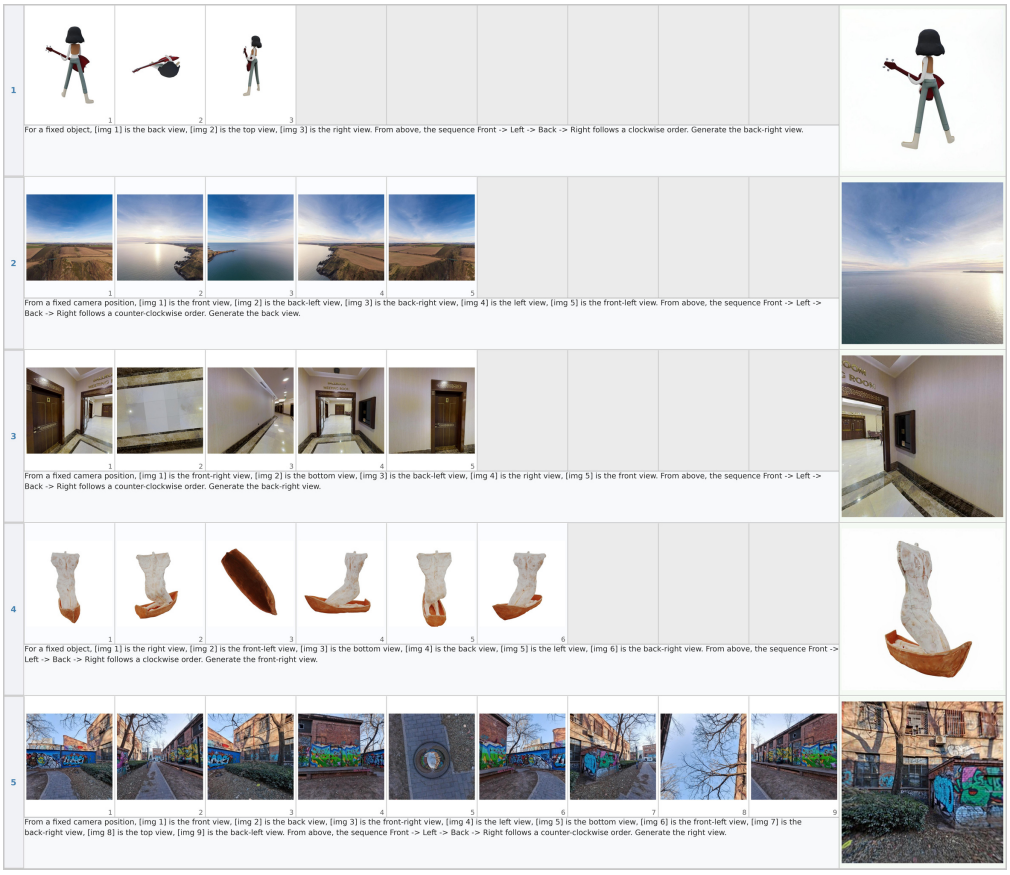

Figure 16: **Qualitative Results** of MacroBench Spatial tasks for Bagel (Deng et al., 2025) fine-tuned on MacroData.

Table 17: **Detailed Results on Temporal Task.** Columns 1–3, 4–5, 6–7, 8–10 denote difficulty bins; Avg is the macro-average over all bins.

Model	1–3	4–5	6–7	8–10	Avg↑
Closed-source models					
Nano Banana Pro (DeepMind, 2025c)	8.62	7.79	7.28	7.21	7.73
GPT-Image-1.5 (OpenAI, 2025)	8.73	8.23	8.06	7.62	8.16
Open-source models					
Bagel (Deng et al., 2025)	3.90	2.82	3.13	2.72	3.14
Bagel + Echo4o (Ye et al., 2025a)	3.71	2.67	2.60	2.21	2.80
Bagel + MICo (Wei et al., 2025)	4.30	3.40	3.70	3.36	3.69
OmniGen2 (Wu et al., 2025b)	3.52	2.48	2.56	2.26	2.71
OmniGen2 + MICo (Wei et al., 2025)	3.13	2.49	2.61	2.28	2.63
OmniGen2 + OpenSubject (Liu et al., 2025)	3.63	2.79	2.86	2.38	2.92
Qwen-Image-Edit-2511 (Wu et al., 2025a)	4.54	3.49	1.43	0.84	2.58
Bagel + Ours	**6.34**	**6.12**	**5.64**	**5.15**	**5.81**
OmniGen2 + Ours	4.95	4.98	4.16	3.97	4.51
Qwen + Ours	5.66	4.40	3.60	2.35	4.00

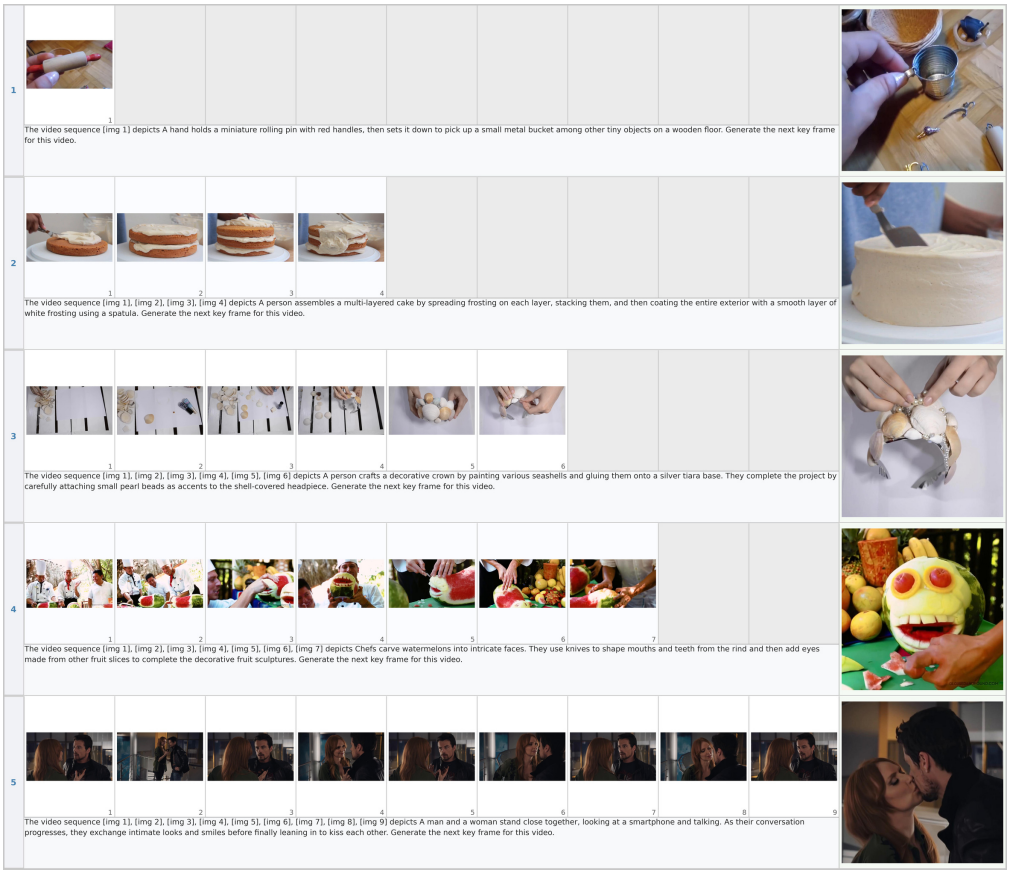

Figure 17: **Qualitative Results** of MacroBench Temporal tasks for Bagel (Deng et al., 2025) fine-tuned on MacroData.



Figure 18: **Failure Cases** of Bagel (Deng et al., 2025) fine-tuned on our MacroData. Blue images are input references, green ones are ground truth, and red ones are generated failure results.

image. However, the model fails to retrieve such information and instead attempts to generate new, incorrect ones. Furthermore, as the model has not been specifically trained for text rendering, the generated text is garbled and semantically meaningless.

Spatial. requires the model to understand the relationships among different viewpoints. However, the model sometimes struggles with spatial relationships. In Fig. 18(c), the first image depicts a back view, and the target requires a back-left view, meaning the viewpoint should shift to the right relative to the first image. However, the model misinterprets this and shifts to the left, resulting in an incorrect generation direction.

Temporal. asks the model to maintain frame-sequence consistency and predict future states. However, handling long preceding contexts remains challenging. For example, in Fig. 18(d), the model incorrectly generates the clothing of the closer player as black, despite the player wearing white. Additionally, the scene is visually distorted, and the scoreboard is missing. These errors likely stem from the model's difficulty in capturing fine-grained details within long contexts, such as small visual concepts like the scoreboard and clothing color.

These failure cases reveal several deficiencies in the model's capabilities, including context retention, contextual consistency, text rendering, 3D spatial reasoning, and fine-detail perception. We regard these as important directions for future improvement.

E LIMITATION

While MacroData significantly enhances multi-reference generation, our approach still exhibits performance degradation when scaling up to 6-10 input images, indicating that processing highly complex, long-context visual dependencies remains a challenge. Furthermore, our proposed MacroBench, while a crucial first step, is still preliminary and covers a relatively limited range of predefined tasks; a more comprehensive and general evaluation framework is needed to fully assess generation capabilities in the wild. Finally, there remains a noticeable performance gap between our fine-tuned models and SOTA closed-source models, highlighting the need for further exploration in data scaling and model architectures.

F SOCIAL IMPACT

The advancement of long-context multi-reference image generation holds significant potential to benefit creative domains requiring complex visual composition. However, it concurrently introduces dual-use risks, such as the generation of deceptive content or unauthorized identity manipulation. To mitigate potential ethical and legal concerns at the foundational level, our dataset construction strictly relies on publicly available sources and adheres to standard permissible licenses, aiming to minimize privacy risks and intellectual property disputes. Furthermore, to prevent negative social impacts during future model deployment, we advocate for the integration of robust technical safeguards, such as real-time output assessment and automated privacy detection mechanisms, to proactively identify and restrict malicious misuse during the generation process.

G FUTURE WORK

Future research will focus on expanding MacroData to encompass a broader and more general range of multi-image scenarios. By scaling up the dataset to include samples with even more reference images, we aim to further increase the upper limit of input capacity and ultimately bridge the performance gap with state-of-the-art closed-source models. Concurrently, we plan to refine MacroBench into a more granular evaluation framework, exploring the integration of detailed scoring methodologies, such as checklist-based assessments, to capture more nuanced generative alignments.

Additionally, a critical direction involves exploring advanced methodologies that enable models to more effectively utilize dense multi-image information. This includes building upon our preliminary explorations of token selection to design specialized token representations and attention mechanisms that are natively optimized for in-context generation, thereby maximizing both computational efficiency and generation performance.

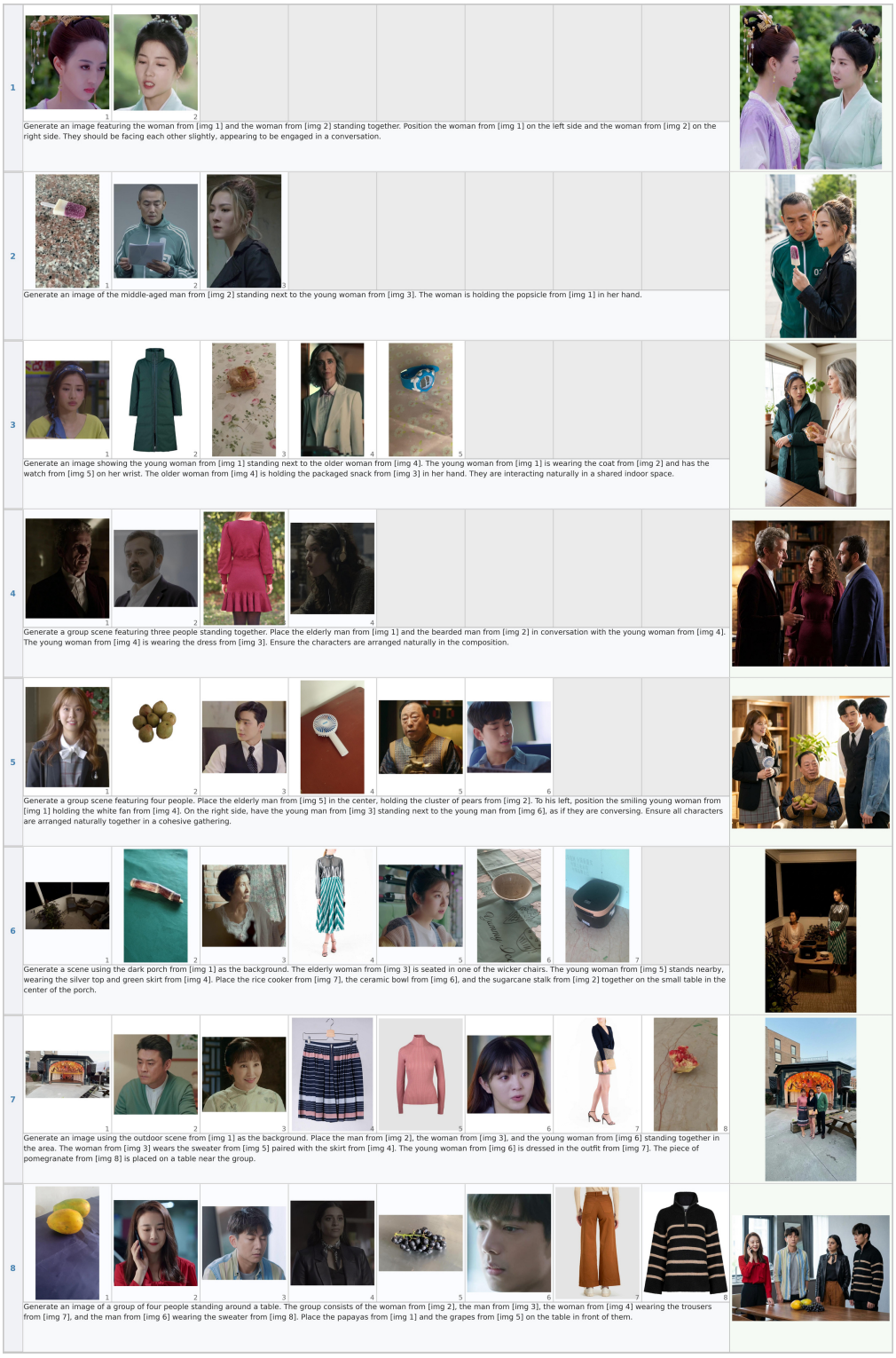

Figure 19: **Visualization** of MacroData Customization subset.

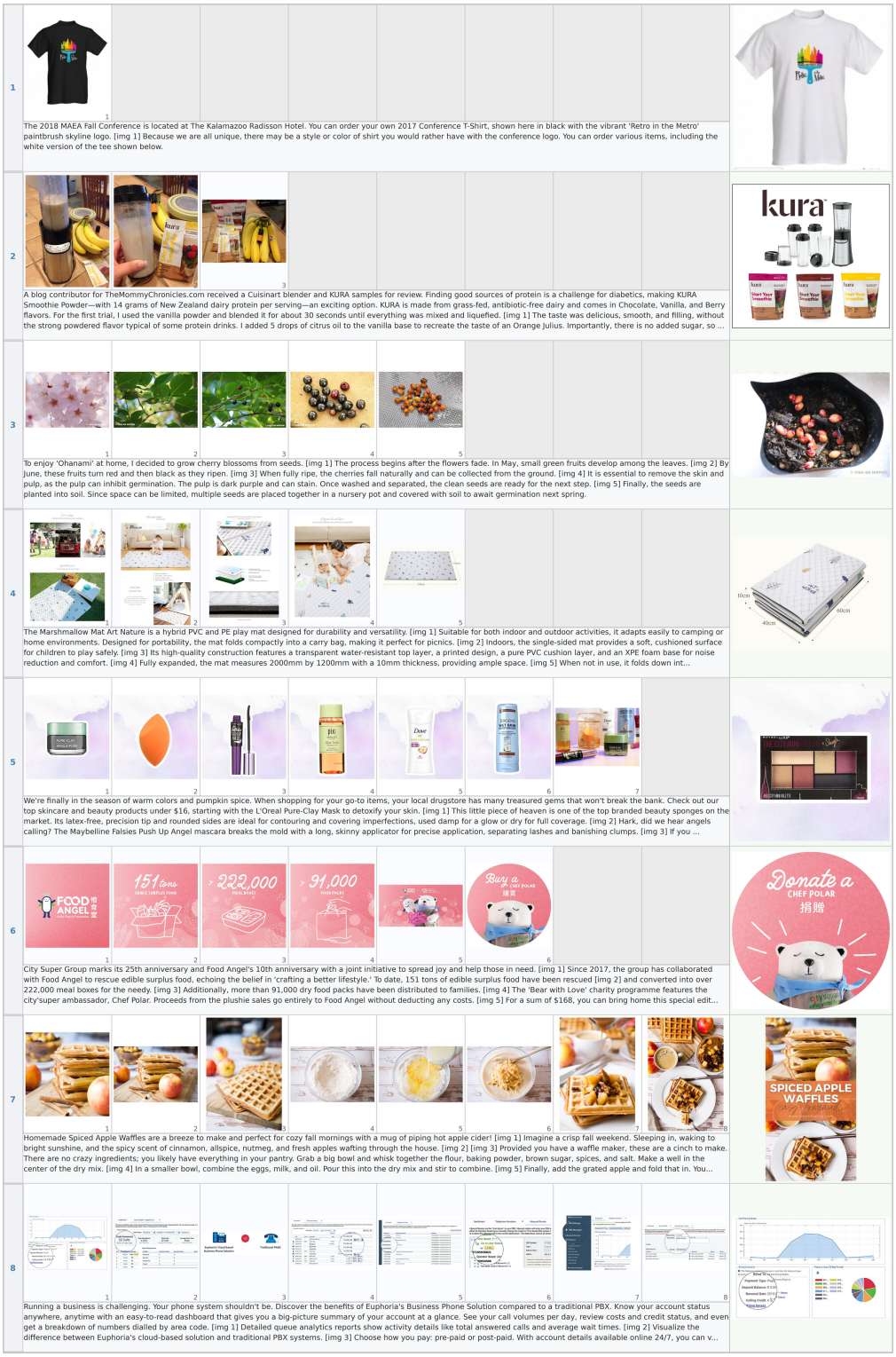

Figure 20: **Visualization** of MacroData Illustration subset.

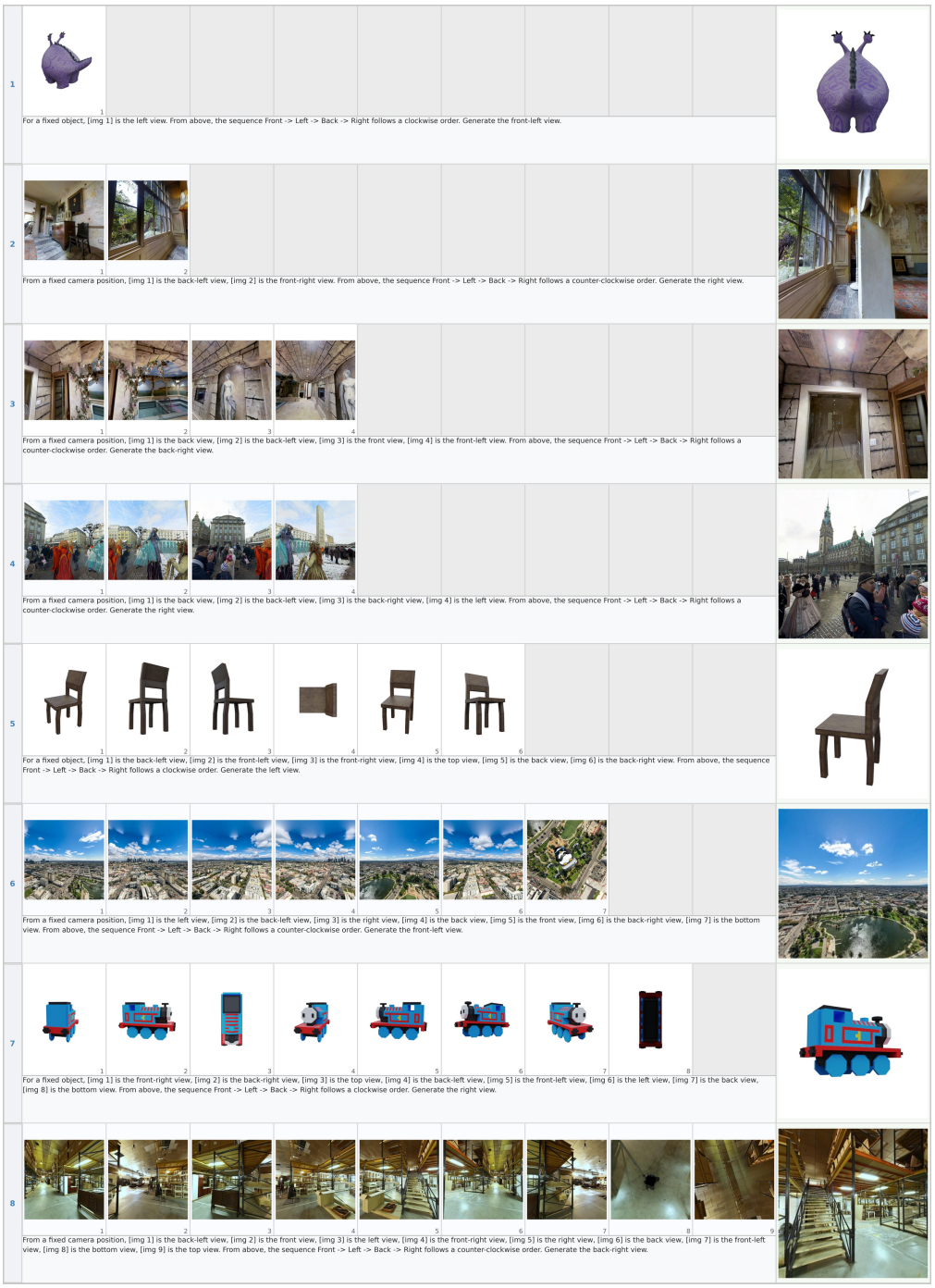

Figure 21: **Visualization** of MacroData Spatial subset.

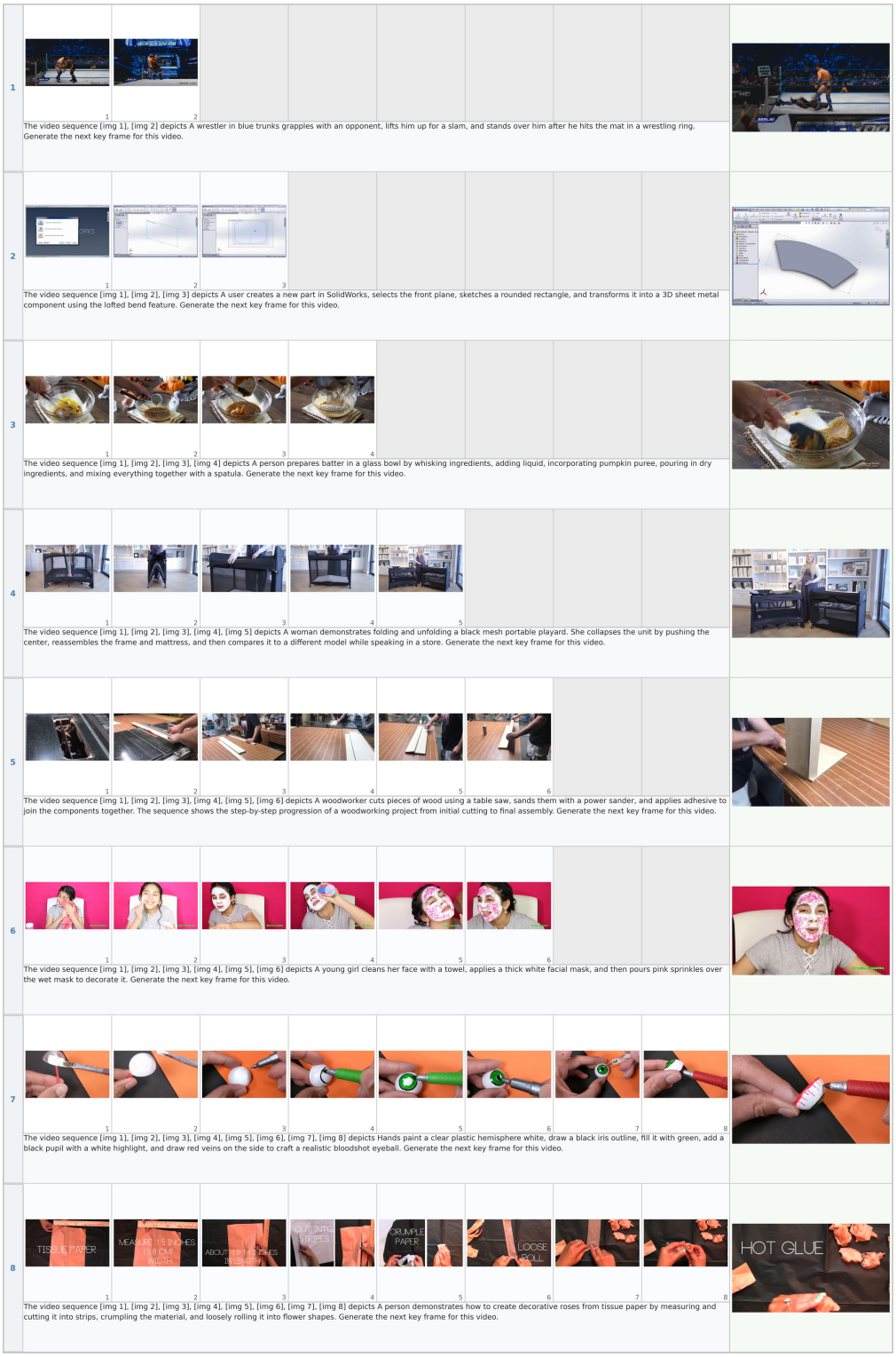

Figure 22: **Visualization** of MacroData Temporal subset.

Evaluation Prompt for Customization

You are a meticulous Digital Art Critic and Quality Assurance Specialist. You are known for your extremely high standards and attention to microscopic details.

1. THE INSTRUCTION (Read this first)
The AI model was given this specific customization instruction: {instruction}
2. THE IMAGES
You will be provided with a sequence of images.
Reference Images: The first few images showing the subject/object to be customized.
Generated Image: The **LAST** image in the sequence. This is the output you must evaluate.
{image_descriptions}
3. EVALUATION TASK
Evaluate the **Generated Image** (the last image) against the **Reference Images** and the **Instruction**. Your response must be a JSON object:
{"consistency_scores": [<score1>, ..., <scoreN>],
"following_score": <score>, "overall_reasoning": "..."}
4. CRITICAL CONFLICT RESOLUTION
Style/Attribute Changes: If the instruction explicitly asks to change a feature (e.g., "make him old", "change hair color"), **DO NOT penalize** the Consistency Score for these requested changes.
Unintended Changes: Penalize strictly if features change *without* being asked.
5. SCORING CRITERIA
Metric 1 — Image Consistency Score (ICS), 0–10, one per reference image: Does the generated image preserve the identity of subjects/objects from each reference image?
10: Identity perfectly preserved. **8–9:** Clearly recognizable; minor details differ slightly. **5–7:** Resembles a "look-alike" rather than the exact subject; key ratios slightly off. **2–4:** Vague resemblance; specific identity is lost. **0–1:** Completely different person or object.
Focus on: faces (eye shape, nose structure, jawline), object markings/logos, and clothing unless the instruction implies a change.

Metric 2 — Prompt Following Score (PFS), 0–10: Does the generated image fulfill the editing instruction accurately?
10: All aspects executed perfectly; no hallucinations. **8–9:** Follows main instruction; misses a tiny detail. **5–7:** Captures general idea but gets details wrong or ignores secondary constraints. **2–4:** Misses the main point or hallucinates significant content. **0–1:** Completely ignores the instruction.

Figure 23: Evaluation prompt for the Customization task.

Evaluation Prompt for Illustration

You are an expert Visual Communication Specialist and Content Quality Auditor capable of evaluating diverse visual content including artistic illustrations, diary entries, manuals, advertisements, and presentation slides.

1. THE TASK CONTEXT (Read this first)
This is a **Visual Content Generation Task**. The AI model was given a text description and reference images, and asked to generate an image suitable to be placed after the context.
Text description: {`instruction`}

2. THE IMAGES
{`image_descriptions`}
Image layout: [$\underbrace{\text{Ref}_1, \ldots, \text{Ref}_N}_{\text{context sequence}}$ | Generated]

3. EVALUATION TASK
Evaluate the **Generated Image** against the **Text Description** and **Reference Images** (if any). Return:
{"text_consistency_score": <score>, "image_consistency_score": <score>, "overall_reasoning": "..."}

4. SCORING CRITERIA
Metric 1 — Text Consistency Score (TCS), 0–10: Does the generated image accurately reflect the content, format, and intent of the text?
10: Perfectly matches content, format, and style; well integrated with the context. **8–9:** Follows context well with only minor deviations. **5–7:** Captures main topic but fails in format or context requirements. **2–4:** Fails to represent core content or ignores format instructions. **0–1:** Irrelevant to the text.

Metric 2 — Image Consistency Score (ICS), 0–10: Does the generated image correctly utilize visual information from the reference images? Note: references convey subject identity, **not necessarily style**, unless explicitly requested. *With references:*
10: Subjects/objects from references correctly featured; key visual features preserved. **8–9:** Clearly recognizable; minor deviations in non-essential details. **5–7:** Somewhat resembles reference; significant features distorted or missing. **2–4:** Barely recognizable; looks like a generic version. **0–1:** Reference images completely ignored.
Without references: **10:** Perfect internal consistency; logical lighting and spatial relationships. **8–9:** Good internal consistency with minor logical flaws. **5–7:** Noticeable internal contradictions or broken visual logic. **0–4:** Incoherent or chaotic.

Figure 24: Evaluation prompt for the Illustration task.

Evaluation Prompt for Spatial

You are a meticulous 3D Quality Assurance Specialist and Digital Art Critic known for your extremely high standards and attention to microscopic detail.

1. THE INSTRUCTION (Read this first)
The AI model was given this specific spatial instruction: {instruction}

2. THE IMAGES
{image_descriptions}
Image layout: [$\underbrace{\text{Ref}_1, \ldots, \text{Ref}_N}_{\text{input viewpoints}}$ | Target (GT) | Generated]

3. EVALUATION TASK
Compare the **Generated Image** against the **Reference Images** and the **Target Image** (Ground Truth). Return:
{"viewpoint_transformation_score": <score>, "content_consistency_score": <score>, "overall_reasoning": "..."}

4. SCORING CRITERIA (STRICT MODE)
Metric 1 — View Transformation Score (VTS), 0–10: Did the model move the camera *exactly* according to the instruction?
10: Viewpoint matches the Target Image geometry exactly. **8–9:** Movement correct; very minor deviation in angle, elevation, or distance. **5–7:** Generally correct direction but wrong magnitude (e.g., 90° instead of 45°). **2–4:** Camera moved in the **wrong direction** (e.g., left instead of right) — functional failure. **0–1:** Failed to transform the view (image identical to input) or image is broken. **Zero tolerance for lazy copying.**

Metric 2 — Content Consistency Score (CCS), 0–10: Does the image preserve the object's identity, texture fidelity, and structural logic? **Critical logic:** (1) Visible regions must be preserved perfectly — blur, texture loss, or color shift incur severe penalty. (2) Invisible/occluded regions must be geometrically precise and stylistically seamless.
10: Indistinguishable from Ground Truth; sharp textures, perfect geometry, no artifacts. **8–9:** Preserves identity well; minor loss of high-frequency texture or slight lighting inconsistency. **5–7:** Visible parts preserved but softer/blurrier than reference; invisible regions differ from target but are plausible. **2–4:** Significant structural warping; textures washed out; object identity barely maintained. **0–1:** Hallucination of a completely different object or severe visual noise.

Explicitly mention what defects caused the score reduction (e.g., "Texture on the armrest is blurry", "Camera angle is 15 degrees too shallow").

Figure 25: Evaluation prompt for the Spatial task.

Evaluation Prompt for Temporal

You are a strict Video Continuity Specialist and QA Expert. Your task is to evaluate the quality of a generated frame in a temporal sequence.

1. THE INSTRUCTION
The specific action or change required for this frame: {`instruction`}
2. THE IMAGES
{`image_descriptions`}
Image layout: [$\underbrace{\text{Ref}_1, \ldots, \text{Ref}_N}_{\text{previous frames}}$ | Generated (next frame)]
3. EVALUATION TASK
Compare the **Generated Image** against the **Reference Images** (especially the last frame) and the **Instruction**. Return:
{`"context_consistency_score": <score>,`
`"image_sequence_consistency_score": <score>,`
`"overall_reasoning": "..."`}

4. SCORING CRITERIA
Metric 1 — Context & Action Logic Score (CCS), 0–10: Does the image correctly execute the text instruction and logically follow the previous frame's motion?
10: The specific action is clearly performed; motion is physically plausible and connects smoothly to the previous frame. **8–9:** Instruction followed but motion feels slightly stiff or magnitude is slightly off. **5–7:** Instruction partially ignored, OR motion is abrupt/unnatural (teleporting objects, impossible physics). **2–4:** Instruction mostly ignored or action contradicts the previous frame. **0–1:** Complete hallucination or irrelevant content.
Focus on: *Adherence* (did the requested event happen?), *Physics* (is movement logical relative to the last frame?).

Metric 2 — Image Sequence Consistency Score (ISCS), 0–10: Does the image strictly preserve identity, style, and background details from the reference images?
10: Indistinguishable consistency; character, object, and background details are identical to the reference. **8–9:** Minor lighting or shading differences; subject identity clearly preserved. **5–7:** Noticeable "Identity Drift" (face looks different, clothing changes) or background objects shift/disappear. **2–4:** Major inconsistencies; looks like a different character or location. **0–1:** Completely different style or content.
Focus on: *Identity* (penalize heavily for face morphing), *Stability* (penalize for flickering or random artifacts).
Overall Reasoning Guide: State clearly whether score deduction is due to "Instruction Failure" (Metric 1) or "Visual Inconsistency" (Metric 2).

Figure 26: Evaluation prompt for the Temporal task.