INTREAL: A REAL-WORLD DATASET AND BENCHMARK FOR MULTIMODAL INTERLEAVED GENERATION

Anonymous authors
Paper under double-blind review

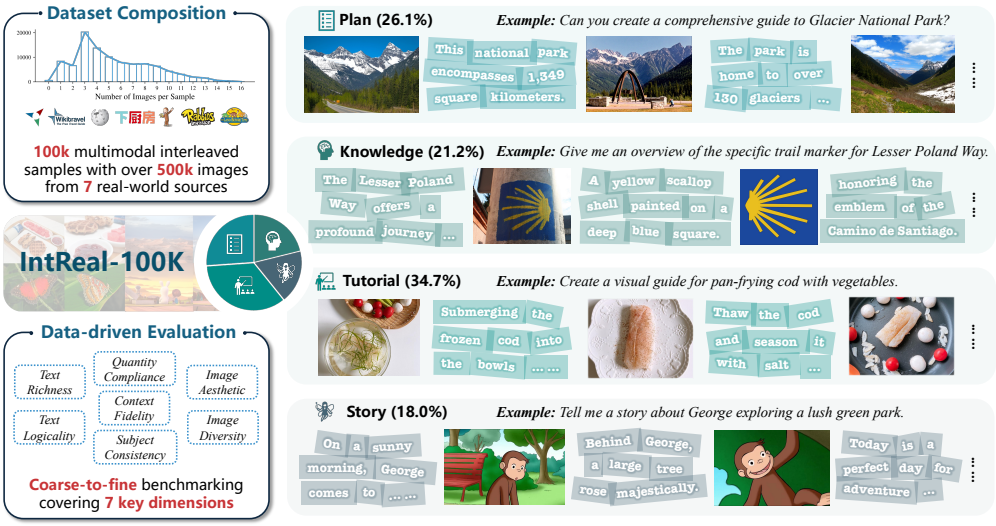


Figure 1: **Overview of IntReal.** IntReal-100K contains 100,000 multimodal interleaved samples with 538,452 images, curated from 7 real-world sources and organized into 4 categories. The IntReal Evaluation Suite performs coarse-to-fine diagnosis via 7 fine-grained dimensions.

ABSTRACT

Humans routinely consume multimodal content, where textual and visual information is naturally interleaved. Although Unified Multimodal Models (UMMs) can jointly generate text and images, multimodal interleaved generation remains largely underexplored. Progress in this area is hindered by two key limitations: (1) the lack of well-defined interleaved generation tasks and corresponding training data, and (2) the absence of fine-grained evaluation protocols that can reliably diagnose model performance across different capability dimensions. To this end, we present **IntReal**, comprising two complementary components: **IntReal-100K**, a real-world dataset of multimodal interleaved samples spanning four major categories, and the **IntReal Evaluation Suite**, which evaluates model performance across seven complementary dimensions using specialized expert models, enabling coarse-to-fine diagnosis. Extensive experiments validate our evaluation suite through low inter-metric redundancy (max pairwise absolute Spearman's $\rho \leq 0.42$), high agreement with human judgments (83.3% accuracy), and strong consistency across different expert evaluators (Spearman's $\rho > 0.9$), supporting the complementarity and reliability of our metrics. By benchmarking four families of UMMs, we show that IntReal Evaluation Suite reveals distinct strengths and weaknesses across models. Post-training experiments demonstrate the utility of IntReal-100K for improving interleaved generation and further show that **training unimodal generation can also benefit interleaved generation**. These findings suggest that interleaved generation can benefit from modality-specific optimization, rather than requiring joint text–image training at every stage. *We will release IntReal-100K, the fine-tuned models, and the evaluation code.*

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.

Table 1: **Comparison with open-source interleaved image-text datasets.** IntReal combines understanding- and generation-oriented filtering with a taxonomy of four real-world scenarios and curated web and video sources.

Dataset	Data Filtering		Taxonomy	Data Source
	For Und.	For Gen.		
MMC4	✓	✗	✗	Web
OBELICS	✓	✗	✗	Web
OmniCorpus	✓	✗	✗	Web
CoMM	✓	✓	✗	Web & Vid.
IntReal	✓	✓	✓	Web & Vid.

Table 2: **Comparison with open-source multimodal interleaved generation benchmarks.** IntReal Evaluation Suite employs VLMs and expert models to provide coarse-to-fine evaluation.

Benchmark	Evaluation Models		Evaluation Granularity	
	VLM	Expert Model	Holistic	Modality-Specific
OpenING	✓	✗	✓	✗
InterleavedEval	✓	✗	✓	✗
ISG-Bench	✓	✗	✓	✓
UEval	✓	✗	✗	✓
IntReal	✓	✓	✓	✓

1 INTRODUCTION

Flip through a book or browse a webpage, and text and images are often tightly interleaved. Encyclopedic articles interleave text and images to communicate knowledge, storybooks integrate narratives with vivid illustrations, tutorials pair step-by-step instructions with corresponding visual demonstrations, and travel guides present itineraries alongside maps and photographs. Despite the emergence of Unified Multimodal Models (UMMs) capable of generating both text and images, multimodal interleaved generation remains largely treated as a by-product of model unification rather than a fundamental generative task in its own right (Deng et al., 2025; Cui et al., 2025; Diao et al., 2026).

A major reason for this underexploration is the lack of datasets and benchmarks specifically designed for this task. On the data side, as shown in table 1, large-scale interleaved datasets such as MMC4 (Zhu et al., 2023), OBELICS (Laurençon et al., 2023) and OmniCorpus (Li et al., 2025b) are primarily collected for multimodal understanding rather than generation, and therefore do not provide generation-oriented quality control. CoMM (Chen et al., 2025c) aggregates interleaved data sources, but does not organize data collection around a taxonomy of real-world scenarios, leaving the task space of multimodal interleaved generation insufficiently characterized. This field thus lacks a high-quality interleaved generation dataset constructed under a principled taxonomy. On the evaluation side, as shown in table 2, existing benchmarks often rely on vision-language models (VLMs) to holistically assess long interleaved responses (Zhou et al., 2025b; Liu et al., 2024), which is inherently misaligned with the fine-grained evaluation required to characterize complex generative capabilities (Huang et al., 2024). Recent studies improve evaluation granularity through modality-specific decomposition or structured verification (Li et al., 2026b). For example, ISG-Bench (Chen et al., 2025a) decomposes responses into four levels of structural granularity, but this decomposition does not map onto distinct capability dimensions. Moreover, it relies solely on a VLM evaluator across all levels, which may lack the specialized competence some dimensions require.

To address these limitations, we introduce IntReal, a real-world dataset and benchmark tailored to multimodal interleaved generation. We first identify four core generation scenarios from real-world interleaved content and organize them into a taxonomy. Guided by this taxonomy, we apply both understanding-oriented and generation-oriented filtering to a million-scale pool of raw interleaved data, resulting in **IntReal-100K**, a high-quality real-world dataset containing 538K images. Building on IntReal-100K, we further develop the **IntReal Evaluation Suite**, an evaluation framework grounded in real-world scenarios. Specifically, we construct an evaluation set through stringent automated filtering and expert curation. We then evaluate interleaved generation in two complementary settings: (1) holistic generation of complete interleaved responses to assess open-ended capabilities and (2) next-image generation conditioned on interleaved contexts for controlled diagnosis. Across both settings, we integrate specialized expert evaluators, such as image quality assessment (IQA) models, into seven metrics, enabling fine-grained diagnosis from diverse perspectives.

We conduct extensive experiments to assess IntReal Evaluation Suite and the utility of IntReal-100K for training UMMs. First, we examine three properties of the suite. Its seven metrics show low redundancy (maximum pairwise absolute Spearman correlation below 0.42), suggesting that they capture distinct capability dimensions. Automatic evaluations agree with human judgments in 83.3% of cases, while results remain consistent across different expert evaluators (Spearman correla-

tion above 0.9). Second, supervised fine-tuning (SFT) on IntReal-100K improves text generation in models such as SenseNova-U1 (Diao et al., 2026), but degrades aspects of image generation, echoing the well-known tension between optimizing text and image generation in UMMs. Given the limited exploration of reinforcement learning for interleaved generation, we apply Flow-GRPO (Liu et al., 2026a) to next-image generation conditioned on interleaved contexts, while testing whether our metrics can serve as effective reward signals. Flow-GRPO recovers the degraded image generation capability and achieves the highest overall score among the open-source UMMs we evaluate on IntReal Evaluation Suite. These results show that training a single generation modality can improve interleaved generation, opening up optimization paths beyond conventional SFT on multimodal interleaved data. Together, these experiments show that IntReal-100K supports effective training of UMMs for interleaved generation, while IntReal Evaluation Suite accurately and reliably identifies strengths and weaknesses across multiple capability dimensions.

Overall, our main contributions are summarized as follows:

1. **Real-world dataset.** We introduce IntReal-100K, a taxonomy-guided dataset for multimodal interleaved generation containing 100K high-quality, real-world interleaved sequences. Experiments demonstrate its utility for improving the interleaved generation capabilities of UMMs.
2. **Fine-grained data-driven benchmark.** We develop IntReal Evaluation Suite, a data-driven benchmark that assesses interleaved generation across seven capability dimensions through coarse-to-fine evaluation. Three validation studies show that its metrics have low redundancy, align well with human judgments, and remain consistent across different expert evaluators.
3. **Modality-specific gains transfer to interleaved generation.** We show that optimizing image generation under interleaved contexts can improve overall multimodal interleaved generation capability. This finding expands the optimization space beyond joint SFT on interleaved sequences.

2 Real-World Interleaved Dataset Construction

2.1 Characterizing Real-World Interleaved Content

Web pages and videos are among the most abundant sources of naturally interleaved multimodal content. We collect a million-scale raw corpus of interleaved text-image content from seven sources that permit public collection and redistribution. Based on the functional roles and collaboration patterns between text and image, we analyze the collected data and identify four categories:

- **Planning:** *Decision making under user-specific goals and constraints.* Text provides detailed information under given objectives, preferences, and constraints, while images visualize the corresponding spatio-temporal changes. We collect content from *WikiVoyage* and *WikiTravel*.
- **Knowledge:** *Explaining concepts through real-world knowledge.* Text conveys concepts and explanations, while images provide corresponding visual representations and help ground abstract knowledge. We collect such encyclopedia content from *Wikipedia*.
- **Tutorial:** *Executing tasks through temporally ordered instructions.* Text describes procedural steps, while images illustrate intermediate states and temporal transitions to align instructions with execution, such as cooking tutorials. This category includes data collected from *Xiachufang*.
- **Story:** *Expressing ideas through narrative-driven content creation.* Text drives narrative progression, while images construct characters and scenes. Inspired by SEED-Story (Yang et al., 2025), this category includes data collected from *The Land Before Time*, *Rabbids*, and *Curious George*.

As illustrated in fig. 1, we visualize representative examples from the four representative categories identified above. Licensing details for each source are discussed in Appendix A.1.

2.2 Data Collection and Curation

High-quality training data for interleaved generation requires both individual modality quality and strong cross-modal alignment. To this end, we design a multi-stage curation pipeline that progressively improves data quality from low-cost filtering to high-cost semantic polishing, producing interleaved data for training UMMs. Specifically, we first apply low-cost visual quality filters to

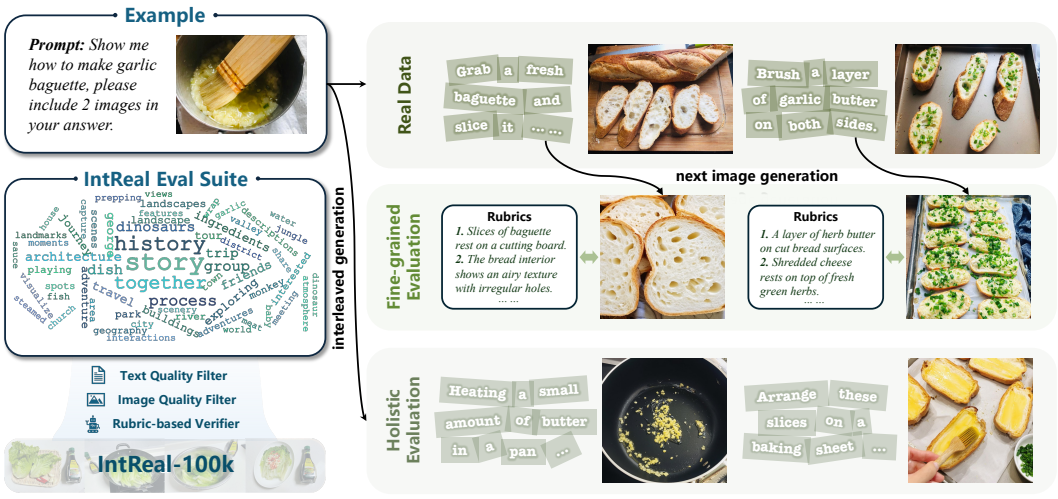


Figure 2: **Overview of the IntReal Evaluation Suite.** We select 600 high-quality samples from IntReal-100K to construct this suite. Each sample is evaluated at two granularities: holistic evaluation generates the complete interleaved sequence at once, while fine-grained evaluation generates the next image conditioned on the preceding interleaved context to enable controllable diagnosis.

remove noisy and low-quality samples. We then leverage VLMs to polish textual descriptions while preserving their original factual content, improving readability and cross-modal consistency. The overall pipeline consists of the following stages (see Appendix A.2 for detailed filtering criteria):

Step 1: Metadata-based Filtering. We first filter images by file metadata, removing those with abnormal aspect ratios, low resolution, or unsupported formats to ensure basic visual quality.

Step 2: Representation-based Filtering. We employ representation-based specialized visual models to identify and remove undesirable content. Specifically, we use the RALI (Zhao et al., 2026) Image Quality Assessment (IQA) model to filter low-quality images, and finetune SigLIP2 (Tschannen et al., 2025) classifiers on expert-annotated in-domain data to detect watermarks and human faces. We further perform intra-sequence deduplication using DINOv3 (Siméoni et al., 2025) to remove near-duplicate images, ensuring both visual quality and diversity within each sample.

Step 3: VLM-based Refinement. We employ Qwen3.5-397B-A17B (Qwen Team, 2026a) to generate image captions and polish the surrounding text, following a principle that prioritizes original text over generated captions to preserve factual authenticity while improving text-image consistency.

Step 4: LLM-based Annotation. We use Qwen3.5-397B-A17B to generate prompts and reasoning processes conditioned on the original content and its category, transforming each sample into structured multimodal interleaved data that can be directly used to fine-tune UMMs.

After the above curation process, we perform category-balanced sampling to ensure a basic balance across four categories. Only about 10% of the collected samples are ultimately retained, as any defective segment within a long interleaved sample leads to its removal. This process results in a high-quality interleaved dataset, IntReal-100K, with category statistics provided in Appendix A.3.

3 Real-World Data-Driven Evaluation

As shown in fig. 2, we build our evaluation suite on real-world interleaved samples from IntReal-100K, combining a coarse-to-fine protocol with expert-guided, multi-dimensional assessment.

3.1 Evaluation Set Construction

To construct a reliable evaluation set, we adopt a semi-automatic curation strategy to select high-quality samples from IntReal-100K (see Appendix B.1 for details). Specifically, we first build an automatic candidate pool by evaluating three key dimensions of interleaved sample quality:

- **Text Quality:** We leverage Qwen3.5-397B-A17B to assess textual quality from three perspectives: prompt specificity, response richness, and topic informativeness.
- **Image Quality:** We use RALI to retain samples with high-quality images.
- **Text-Image Alignment:** We adopt a context-aware self-verification strategy using Qwen3.5-397B-A17B. For each image, the model first derives visual rubrics specifying what the image should depict from the preceding text-image context, without accessing the target image. It then checks the target image against rubrics to filter out samples with cross-modal inconsistencies.

We adjust the filtering thresholds to obtain a high-quality candidate pool of manageable size, then manually remove samples that fail to meet the quality criteria. Finally, we perform stratified sampling by category to construct a balanced evaluation set of 600 samples, with 150 per category. We opt for a relatively compact evaluation set, since our coarse-to-fine protocol already requires generating over 6,000 images interleaved with text per model. Scaling further would substantially increase evaluation cost across multiple model families without commensurate benefit.

3.2 Expert Model-based Coarse-to-Fine Benchmarking

To evaluate UMMs in a coarse-to-fine manner, we design two complementary settings for interleaved generation. Holistic evaluation assesses complete interleaved responses generated from scratch. Fine-grained evaluation tests next-image generation under fixed real-world interleaved contexts. These shared contexts enable controlled, fine-grained comparisons across models. The two settings are illustrated in fig. 2. We design seven metrics to assess distinct capabilities of interleaved generation, where H and F denote metrics used for holistic and fine-grained evaluation, respectively:

- **Text Richness (H):** We employ an LLM as a text modality expert to compare generated textual responses with corresponding real-world textual responses through pairwise ranking, evaluating the richness and informativeness of generated texts.
- **Text Logicality (H):** Using the same LLM-based text expert, we assess whether generated texts contain factual errors, contradictions, or logical inconsistencies.
- **Image Aesthetic (H):** We employ an IQA model as a visual modality expert to measure the average aesthetic quality of generated images.
- **Image Diversity (H):** We identify visual repetition as a common failure mode, where images generated at consecutive positions may become overly similar and fail to reflect the progression of the context. To quantify this issue, we compute the DINO feature similarity between adjacent generated images, where extremely high similarity indicates potential visual repetition.
- **Quantity Compliance (H):** We use the number of images in the corresponding real-world sample as the target quantity and require the model to generate the same number, thereby evaluating instruction following with an image budget known to be sufficient for conveying the content.
- **Context Fidelity (F):** We assess whether generated next images are consistent with their preceding interleaved context. We employ a VLM as a multimodal expert to verify whether each generated image satisfies the visual rubrics constructed during evaluation set curation in section 3.1.
- **Subject Consistency (F):** We compute DINO similarity between generated image and their corresponding reference image at the same positions in interleaved samples to measure whether key visual subjects remain consistent, e.g., story characters or cooking ingredients.

Detailed formulations of metrics are provided in Appendix B.2. All metrics are evaluated across all categories, except Subject Consistency, which is computed only for Tutorial and Story. This is because Knowledge and Planning do not necessarily require consistent subjects across images.

Table 3: **IntReal evaluation result per dimension**. We evaluate four UMM families across seven dimensions. All scores are scaled to the range of $[0, 100]$, with higher values indicating better performance. The Δ vs. base rows report relative changes after post-training with respect to the corresponding base model, where green and red indicate improvements and degradations, respectively. Gray rows denote the closed-source API system, and Δ vs. OSS Best reports its relative difference with respect to the best open-source result in each dimension.

Model	Holistic Evaluation					Fine-grained Evaluation		Mean
	Text Richness	Text Logicality	Image Aesthetic	Image Diversity	Quantity Compliance	Context Fidelity	Subject Consistency	
Emu3.5	14.02	**75.17**	81.20	75.51	39.70	65.60	61.46	58.95
U1-A3B-MoT	24.67	69.17	80.70	88.70	72.50	86.81	63.45	69.43
Bagel+*SFT*	17.50	63.50	76.82	91.35	62.83	74.19	60.61	63.83
Bagel+*SFT+RL*	16.67	62.50	**81.94**	90.31	62.50	76.04	65.43	65.06
U1-8B-MoT	21.83	67.33	79.60	88.36	**95.33**	**87.29**	63.38	71.88
+*SFT*	**28.17**	71.83	75.10	87.05	93.50	83.02	**67.52**	72.31
Δ vs. base (%)	*+29.04%*	*+6.68%*	*−5.65%*	*−1.48%*	*−1.92%*	*−4.89%*	*+6.53%*	*+0.60%*
+*SFT+RL*	27.83	70.50	78.53	**93.31**	93.83	85.57	67.11	**73.81**
Δ vs. base (%)	*+27.49%*	*+4.71%*	*−1.34%*	*+5.60%*	*−1.57%*	*−1.97%*	*+5.89%*	*+2.69%*
NanoBanana2	90.80	97.49	82.32	83.56	95.32	76.70	67.65	84.84
Δ vs. OSS Best (%)	+222.33%	+29.69%	+0.46%	-10.45%	-0.01%	-12.13%	+0.19%	+14.94%

4 Experiment

4.1 Experimental Setup

We evaluate four families of UMMs with varying levels of support for interleaved generation and fine-tuning: **(1) Bagel** (Deng et al., 2025), which provides official fine-tuning scripts but whose released checkpoint does not natively support interleaved generation. We therefore report only its performance after fine-tuning on IntReal-100K. **(2) Emu3.5** (Cui et al., 2025), which supports native interleaved inference but does not release fine-tuning methods. We thus evaluate only its original checkpoint. **(3) SenseNova-U1** (Diao et al., 2026), which supports both fine-tuning and interleaved inference, allowing evaluation before and after fine-tuning on IntReal-100K. SenseNova-U1 provides two variants, 8B-MoT and A3B-MoT. Both use an 8B generation branch, while their understanding components adopt an 8B and a 30B-A3B MoE architecture, respectively. For 8B-MoT, the authors additionally release *SenseNova-U1-8B-MoT-Interleaved*, a checkpoint specifically optimized for this task. To ensure a strong baseline, we adopt this checkpoint for our evaluation and fine-tuning. No corresponding specialized checkpoint is released for A3B-MoT, for which we therefore evaluate the standard model. Due to computational constraints, we report fine-tuned results only for the 8B-MoT variant. **(4) NanoBanana2 (Gemini3.1-Flash-Image)**, which is a closed-source API supporting interleaved generation. Since the system is not publicly disclosed, we cannot determine whether this capability is native or implemented through an agentic pipeline. Other models are excluded if they are unavailable or lack publicly released implementations required for our finetuning and evaluation such as LLaDA2.0-Uni (AI et al., 2026). In addition, Anole (Chern et al., 2024) and Show-o2 (Xie et al., 2026) are excluded from our evaluation because their released models are limited in context length and cannot support the long-context interleaved generation setting.

For evaluation, we use Gemma4 (Team et al., 2026a) for both LLM- and VLM-based judging due to its strong performance, open-source availability, and modest computational requirements. For fine-tuning, we use the official training framework of each model for SFT, while all RL experiments are implemented with Flow-Factory (Ping et al., 2026). All post-training data are drawn from the 99.4K samples in IntReal-100K after excluding the 600 samples used in the IntReal Evaluation Suite, ensuring that the training and evaluation sets are strictly disjoint. Additional evaluation details are provided in Appendix B, and further training and inference details are provided in Appendix C.

4.2 Benchmarking UMMs on IntReal

Table 3 presents the per-dimension IntReal evaluation results, while fig. 3 further reports the results across different categories. From these results, we make the following observations.

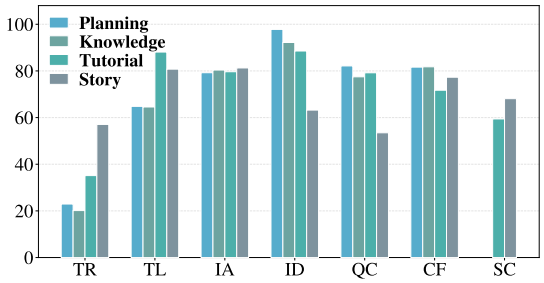


Figure 3: **IntReal evaluation results per category.** We report the score of each metric for each category, computed by averaging over UMMs.

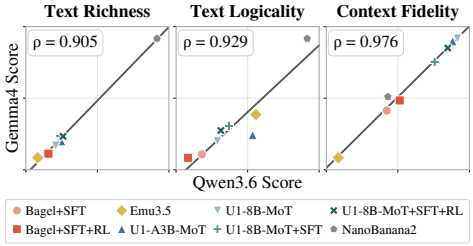


Figure 4: **Spearman correlations across metrics with different VLM evaluators.** The consistent model rankings demonstrate the robustness of our evaluation method.

Different UMMs exhibit distinct and complementary capabilities. No single model performs consistently well across all evaluation dimensions. Among open-source models, Emu3.5 achieves the strongest Text Logicality, while U1-8B-MoT performs substantially better in Quantity Compliance and Context Fidelity. The closed-source NanoBanana2 also does not uniformly outperform open-source models, showing advantages on unimodal dimensions such as Text Richness, Text Logicality, and Image Aesthetic, while remaining below the best open-source results on interleaved-specific dimensions such as Image Diversity and Context Fidelity. These results suggest that strong unimodal capabilities are important for interleaved generation, but are not sufficient on their own, as interleaved generation also requires optimization for interleaved-specific dimensions.

Different categories emphasize distinct capabilities. Planning and Knowledge are more demanding in Text Richness and Text Logicality, while Tutorial is particularly challenging under fine-grained evaluation, with lower Context Fidelity and Subject Consistency. Story often requires longer interleaved sequences, leading to greater challenges in Quantity Compliance and Image Diversity.

OSS UMMs exhibit imbalanced multimodal capabilities. The largest gap between open-source UMMs and the closed-source API lies in the text modality, whereas the gap in image modality is substantially smaller. This asymmetry is consistent with recent findings on scaling UMMs (Tong et al., 2026; Wang et al., 2026), where language is more parameter-hungry while vision is more data-hungry. Given that current OSS UMMs remain much smaller than today's flagship LLMs, scaling model capacity may be one of the most important steps toward stronger interleaved generation.

4.3 Post-training on IntReal-100K

Compared with the original U1-8B-MoT, SFT improves Text Richness, Text Logicality, and Subject Consistency by 29.04%, 6.68%, and 6.53%. However, Image Aesthetic and Context Fidelity decrease by 5.65% and 4.89%, revealing a trade-off between different capabilities. This degradation may arise from the inherent optimization conflict between understanding and generation in UMMs. We therefore investigate whether Reinforcement Learning (RL) can alleviate this trade-off.

To close the loop, we validate whether the metrics introduced in section 3.2 can serve as effective learning signals. We first select 10K high-quality RL samples and annotate them with rubrics following the same procedure in section 3.1. We then apply FlowGRPO (Liu et al., 2026a) using IQA, DINO, and rubric-based rewards for optimization. For U1-8B-MoT, RL substantially mitigates the degradation in Image Aesthetic and Context Fidelity introduced by SFT, while further improving Image Diversity. A similar trend holds for Bagel before and after RL in Image Aesthetic and Context Fidelity, while its already-high Image Diversity remains stable. Overall, U1-8B-MoT with SFT and RL achieves the highest mean score among all OSS models, indicating that improvements in image-generation capability can effectively transfer to interleaved generation. A case study of U1-8B-MoT under different post-training settings is shown in fig. 7.

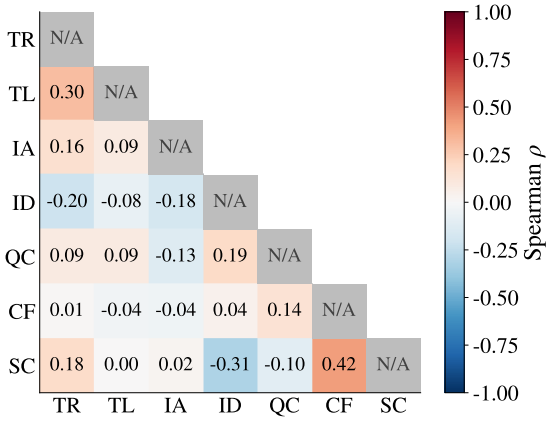


Figure 5: **Pairwise sample-level Spearman correlation between different evaluation metrics.** Low absolute correlations indicate limited redundancy.

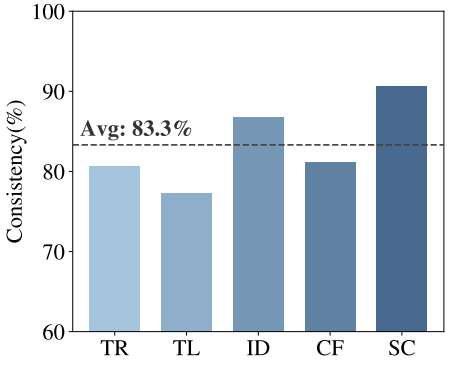


Figure 6: **Agreement between automatic and human annotation for metrics.** Metrics with direct measurements (Quantity Compliance) or human-aligned evaluation models (Image Aesthetic) are excluded.

4.4 Reliability of the IntReal Evaluation

To validate the reliability of our proposed evaluation method, we conduct three complementary analyses. We select Emu3.5, U1-A3B-MoT, Bagel-SFT, U1-8B-MoT and NanoBanana2, avoiding additional fine-tuned variants to prevent repeatedly evaluating highly similar models.

Metric Orthogonality. We first investigate the numerical redundancy and orthogonality among the seven metrics. For each selected model, we compute pairwise sample-level Spearman rank correlations between all metric pairs and average the correlations across models. As shown in fig. 5, the absolute Spearman correlation between any pair of metrics remains below 0.42, indicating that the proposed metrics provide complementary measurements with limited redundancy.

Human-Automatic Agreement. We focus on five subjective metrics, excluding Image Aesthetic and Quantity Compliance, since the former is evaluated by a human-aligned IQA model and the latter is objectively computed. For representative models, we construct all $\binom{5}{2} = 10$ model pairs and sample 15 comparisons for each pair, resulting in 150 pairwise comparisons per metric and 750 expert-annotated comparisons in total. The automatic evaluation agrees with expert judgments in 83.3% of comparisons on average, demonstrating strong human-automatic consistency.

Robustness to expert evaluators. We assess the robustness of our VLM-based metrics by comparing Gemma4 with Qwen3.6-35B-A3B (Qwen Team, 2026b), an open-source model of similar scale. For Text Richness, Text Logicality, and Context Fidelity, we compute the Spearman correlation between all model rankings produced by the two evaluators. As shown in fig. 4, all correlations exceed 0.9, indicating that our metrics maintain stable model rankings across different evaluators.

5 Related Work

Unified Multimodal Models (UMMs). Existing research on UMMs has primarily explored two complementary directions. The first seeks to unify text and image generation within a single model (Ramesh et al., 2021; Sun et al., 2024; Wu et al., 2025a; Zhou et al., 2025a; Chen et al., 2025b; Zhang et al., 2026; Yang et al., 2026; Dong et al., 2024; Team et al., 2026b; Li et al., 2026c). Representative studies combine multimodal understanding with text-to-image generation, while subsequent works further incorporate image editing and other any-to-image tasks to enable image generation conditioned on multimodal contexts (Xie et al., 2025; Wu et al., 2025b; Liu et al., 2026c; Team, 2025). The second treats interleaved text-image sequences as a native form of multimodal pretraining rather than a collection of separate downstream tasks (Cui et al., 2025; Tong et al., 2026; Han et al., 2026; Tian et al., 2024; Kou et al., 2024; AI et al., 2026). Chameleon (Team, 2024) exem-



Figure 7: **Case study on SenseNova-U1-8B-MoT.** SFT produces more detailed text but degrades visual details (distorted brush), while RL preserves the text improvements and restores image quality.

plifies this formulation through mixed-modal sequence modeling, while Bagel (Deng et al., 2025) provide s a large-scale open-source realization based on interleaved multimodal data.

Multimodal Interleaved Datasets and Benchmarks. Early interleaved datasets were primarily developed for multimodal understanding and pretraining, constructed through text-based synthesis or web-scale collection (Zhu et al., 2023; Laurençon et al., 2023; Li et al., 2025b; Awadalla et al., 2024). CoMM (Chen et al., 2025c) further introduces generation-oriented filtering, making interleaved data more suitable for generation tasks. Recent studies have shifted toward enabling models to learn multimodal reasoning traces (Gu et al., 2026; Li et al., 2026a). On the evaluation side, early benchmarks mainly define diverse downstream tasks for interleaved generation and commonly rely on VLMs for holistic assessment (Zhou et al., 2025b; Liu et al., 2024). UEval (Li et al., 2026b) and ISG-Bench (Chen et al., 2025a) take steps toward fine-grained evaluation by decoupling interleaved responses into more structured evaluation targets. Meanwhile, Uni-MMMU (Zou et al., 2026) and RealUnify (Shi et al., 2026) benchmark the bidirectional synergy between understanding and generation through tasks that explicitly couple textual reasoning and visual generation. In contrast, WEAVE (Chow et al., 2026) evaluates UMMs over multi-turn interleaved text-image contexts, with each turn producing a single-modality response. Beyond UMMs, UniM (Li et al., 2026d) evaluates omni-modal models under arbitrary combinations of inputs and outputs across multiple modalities.

6 Conclusion

In this work, we systematically study multimodal interleaved generation from the perspectives of application scenarios, data construction, evaluation, and post-training. We first characterize four representative real-world scenarios and develop a taxonomy-guided data curation pipeline to construct IntReal-100K from real-world content. Building upon these data, we introduce the IntReal Evaluation Suite, which follows a coarse-to-fine evaluation principle and integrates specialized expert models to provide complementary measurements across seven dimensions. Our experiments reveal that SFT on IntReal-100K brings gains but also introduces trade-offs across different capabilities. Through Flow-GRPO, we further show that the proposed metrics can serve as effective learning signals, and that improving unimodal generation capability can generalize to multimodal interleaved generation. These findings suggest that modality-specific optimization can improve interleaved generation, expanding the training options beyond joint optimization of text and image generation .IntReal also has several limitations. Constructing high-quality real-world interleaved data requires aggressive filtering and thus yields a relatively low retention rate. In addition, we do not explore mixed post-training with unimodal generation data to maintain a clean experimental setting. We leave synthetic data augmentation and mixed-data post-training for future work.

AI USE STATEMENT

Generative AI tools were used as part of the research methodology for dataset construction and evaluation, as described in detail in the main paper and appendix. During manuscript preparation, generative AI tools were used to assist with language polishing and translation. Generative AI tools were also used to assist with the implementation of non-critical code components, all of which were manually reviewed and verified by the authors. All human annotation reported in the paper involved actual human participants and were not replaced by generative AI.

ETHICS STATEMENT

IntReal is constructed from publicly accessible real-world sources that permit data collection and redistribution, with source-specific licensing information provided in Appendix A.1. During data curation, we apply dedicated filters to remove human faces, watermarks, and low-quality visual content to reduce privacy and data-quality concerns. Nevertheless, IntReal may inherit biases, factual errors, or other undesirable content from its original sources or from model-assisted refinement. We therefore encourage users to consider these limitations when training or deploying models using the dataset. The dataset and accompanying resources will be released with documentation of their sources, intended use, and relevant licensing information.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the data sources and curation pipeline in Section 2 and Appendix A, and document the construction, metric definitions, evaluation protocols, and human validation of the IntReal Evaluation Suite in Section 3 and Appendix B. Model selection, inference settings, and post-training configurations are provided in Section 4 and Appendix C. We ensure that all post-training data are strictly disjoint from the evaluation set. The complete IntReal-100K dataset and evaluation code will be publicly released to facilitate reproduction and future research.

REFERENCES

Inclusion AI, Tiwei Bie, Haoxing Chen, Tieyuan Chen, Zhenglin Cheng, Long Cui, Kai Gan, Zhicheng Huang, Zhenzhong Lan, Haoquan Li, et al. Llada2. 0-uni: Unifying multimodal understanding and generation with diffusion large language model. *arXiv preprint arXiv:2604.20796*, 2026.

Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Guha, Matt Jordan, Sheng Shen, Mohamed Awadalla, Silvio Savarese, et al. Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens. *Advances in Neural Information Processing Systems*, 37:36805–36828, 2024.

Dongping Chen, Ruoxi Chen, Shu Pu, Zhaoyi Liu, Yanru Wu, Caixi Chen, Benlin Liu, Yue Huang, Yao Wan, Pan Zhou, and Ranjay Krishna. Interleaved scene graphs for interleaved text-and-image generation assessment. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025b.

Wei Chen, Lin Li, Yongqi Yang, Bin Wen, Fan Yang, Tingting Gao, Yu Wu, and Long Chen. Comm: A coherent interleaved image-text dataset for multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8073–8082. IEEE, 2025c.

Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024.

Wei Chow, Jiachun Pan, Yongyuan Liang, Mingze Zhou, Xue Song, Liyu Jia, Saining Zhang, Siliang Tang, Juncheng Li, Fengda Zhang, et al. Weave: Unleashing and benchmarking the in-context

interleaved comprehension and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15343–15353, 2026.

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.

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.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, et al. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026.

Runpei Dong, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, et al. Dreamllm: Synergistic multimodal comprehension and creation. In *International Conference on Learning Representations*, volume 2024, pp. 6666–6702, 2024.

Jiawei Gu, Yunzhuo Hao, Huichen Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 141405–141447, 2026.

Junlin Han, Shengbang Tong, David Fan, Minghao Chen, Philip Torr, Filippos Kokkinos, and Mike Lewis. Towards physics of multimodal pretraining: Knowledge flow, modality synergy, early unification, and recipes. *arXiv preprint arXiv:2608.05000*, 2026.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Siqi Kou, Jiachun Jin, Zhihong Liu, Chang Liu, Ye Ma, Jian Jia, Quan Chen, Peng Jiang, and Zhijie Deng. Orthus: Autoregressive interleaved image-text generation with modality-specific heads. *arXiv preprint arXiv:2412.00127*, 2024.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023.

Ang Li, Charles L. Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, Tom Goldstein, and Micah Goldblum. Zebra-cot: A dataset for interleaved vision-language reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Bo Li, Yida Yin, Wenhao Chai, Xingyu Fu, and Zhuang Liu. Ueval: A benchmark for unified multimodal generation. *arXiv preprint arXiv:2601.22155*, 2026b.

Junzhe Li, Yutao Cui, Tao Huang, Weijie Kong, Yiming Cheng, Chuxuan Zeng, Yinping Ma, Chun Fan, Miles Yang, Zhao Zhong, et al. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025a.

Lijiang Li, Zuwei Long, Yunhang Shen, Heting Gao, Haoyu Cao, Xing Sun, Caifeng Shan, Ran He, and Chaoyou Fu. Omni-diffusion: Unified multimodal understanding and generation with masked discrete diffusion. In *Forty-third International Conference on Machine Learning*, 2026c.

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. In *International Conference on Learning Representations*, volume 2025, pp. 13647–13689, 2025b.

Yanlin Li, Minghui Guo, Kaiwen Zhang, Shize Zhang, Yiran Zhao, Haodong Li, Congyue Zhou, Weijie Zheng, Yushen Yan, Shengqiong Wu, et al. Unim: A unified any-to-any interleaved multimodal benchmark. *arXiv preprint arXiv:2603.05075*, 2026d.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2026a.

Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. Holistic evaluation for interleaved text-and-image generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22002–22016, 2024.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, et al. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *arXiv preprint arXiv:2601.05242*, 2026b.

Zhiheng Liu, Weiming Ren, Haozhe Liu, Zijian Zhou, Shoufa Chen, Haonan Qiu, Xiaoke Huang, Zhaochong An, Fanny Yang, Aditya Patel, et al. Tuna: Taming unified visual representations for native unified multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15740–15751, 2026c.

Bowen Ping, Chengyou Jia, Minnan Luo, Hangwei Qian, and Ivor Tsang. Flow-factory: A unified framework for reinforcement learning in flow-matching models. *arXiv preprint arXiv:2602.12529*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026b.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Yang Shi, Yuhao Dong, Yue Ding, Yuran Wang, Xuanyu Zhu, Sheng Zhou, Wenting Liu, Haochen Tian, Rundong Wang, Huanqian Wang, et al. Realunify: Do unified models truly benefit from unification? a comprehensive benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22488–22497, 2026.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *International conference on learning representations*, volume 2024, pp. 12352–12380, 2024.

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

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026a.

Meituan LongCat Team, Bin Xiao, Chao Wang, Chengjiang Li, Chi Zhang, Chong Peng, Hang Yu, Hao Yang, Haonan Yan, Haoze Sun, et al. Longcat-next: Lexicalizing modalities as discrete tokens. *arXiv preprint arXiv:2603.27538*, 2026b.

Tencent Hunyuan Foundation Model Team. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, et al. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *arXiv preprint arXiv:2401.10208*, 2024.

Shengbang Tong, David Fan, John Nguyen, Ellis L Brown II, Gaoyue Zhou, Shengyi Qian, Boyang Zheng, Théophane Vallaeys, Rob Fergus, Naila Murray, Marjan Ghazvininejad, Mike Lewis, Nicolas Ballas, Amir Bar, Michael Rabbat, Jakob Verbeek, Luke Zettlemoyer, Koustuv Sinha, Yann LeCun, and Saining Xie. Beyond language modeling: An exploration of multimodal pretraining. In *Forty-third International Conference on Machine Learning*, 2026.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Feng Wang and Zihao Yu. Coefficients-preserving sampling for reinforcement learning with flow matching. *arXiv preprint arXiv:2509.05952*, 2025.

Haifeng Wang, Hua Wu, Tian Wu, Yu Sun, Jing Liu, Dianhai Yu, Yanjun Ma, Jingzhou He, Zhongjun He, Dou Hong, et al. Ernie 5.0 technical report. *arXiv preprint arXiv:2602.04705*, 2026.

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

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 93620–93638, 2025b.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025.

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

Yixian Xu, Kaiyuan Gao, Yuxiang Chen, Yilei Chen, Zecheng Tang, Zihao Liu, Zikai Zhou, Deqing Li, Hao Meng, Kuan Cao, et al. Qwen-image-2.0-rl technical report. *arXiv preprint arXiv:2606.27608*, 2026.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *Advances in Neural Information Processing Systems*, 38:138867–138907, 2026.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Ying-Cong Chen. Seed-story: Multimodal long story generation with large language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1871–1881, 2025.

Yichen Zhang, Da Peng, Zonghao Guo, Zijian Zhang, Xuesong Yang, Tong Sun, Shichu Sun, Yidan Zhang, Yanghao Li, Haiyan Zhao, et al. Cheers: Decoupling patch details from semantic representations enables unified multimodal comprehension and generation. *arXiv preprint arXiv:2603.12793*, 2026.

Shijie Zhao, Xuanyu Zhang, Weiqi Li, Junlin Li, Li Zhang, Tianfan Xue, and Jian Zhang. Reasoning as representation: Rethinking visual reinforcement learning in image quality assessment. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

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. In *International Conference on Learning Representations*, volume 2025, pp. 6446–6469, 2025a.

Pengfei Zhou, Xiaopeng Peng, Jiajun Song, Chuanhao Li, Zhaopan Xu, Yue Yang, Ziyao Guo, Hao Zhang, Yuqi Lin, Yefei He, et al. Opening: A comprehensive benchmark for judging open-ended interleaved image-text generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 56–66, 2025b.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36:8958–8974, 2023.

Kai Zou, Ziqi Huang, Yuhao Dong, Shulin Tian, Dian Zheng, Hongbo Liu, Jingwen He, Bin Liu, Yu Qiao, and Ziwei Liu. Uni-mmmu: A massive multi-discipline multimodal unified benchmark. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 908–924, 2026.

Appendix

A Dataset Details . . . 15
A.1 Data Sources and Licensing . . . 15
A.2 Detailed Filtering Pipeline . . . 16
A.3 Dataset Statistics . . . 17
B Benchmark Details . . . 17
B.1 Evaluation Set Construction . . . 17
B.2 Metric Definitions . . . 18
B.3 Evaluation Prompts . . . 20
B.4 Metric Calibration Cases . . . 21
C Experiment Details . . . 24
C.1 Model Training Settings . . . 24
C.2 Model Inference Settings . . . 25
C.3 Detailed Evaluation Results . . . 25
D Qualitative Examples . . . 27

A Dataset Details

A.1 Data Sources and Licensing

Wikivoyage. Wikivoyage is a collaborative, community-edited travel guide hosted by the Wikimedia Foundation. It provides destination-oriented articles covering attractions, transportation, accommodation, local customs, and suggested itineraries, making it a natural source of planning-oriented interleaved content. Textual content on Wikivoyage is released under the Creative Commons Attribution-ShareAlike license, which permits copying, redistribution, and adaptation provided that appropriate attribution is given and derivative works are distributed under the same or a compatible license. Images and other media may be subject to file-specific licenses, and we retain the corresponding provenance and licensing information when collecting the data. Our use and release of Wikivoyage-derived content follow the applicable attribution and share-alike requirements.

Wikitravel. Wikitravel is a collaboratively written travel guide containing destination descriptions, and itinerary-oriented content contributed by its community. Its structured combination of descriptive text and travel-related images provides another source of planning-oriented multimodal sequences in IntReal. Wikitravel requires its textual contributions to be licensed under the Creative Commons Attribution-ShareAlike 3.0 license or a compatible license, allowing redistribution and modification with appropriate attribution and share-alike requirements. Uploaded images and other media are governed by their respective file-specific licenses. Therefore, we inspect the licensing metadata associated with each sample and filter out content whose licensing terms do not permit redistribution. All Wikitravel-derived data released in IntReal follow these licensing conditions.

Wikipedia. Wikipedia is a collaboratively maintained online encyclopedia hosted by the Wikimedia Foundation, covering a broad range of real-world concepts, entities, events, and scientific topics. Its articles naturally integrate explanatory text with illustrative diagrams and photographs, making Wikipedia the primary source of knowledge-oriented interleaved content in IntReal. Wikipedia textual content is distributed under Creative Commons licenses that permit reuse, redistribution, and adaptation subject to attribution and share-alike requirements. Images and other media are individually licensed, including Creative Commons and other compatible licenses, and their licensing

information is provided on the corresponding file pages. During data construction, we preserve source and license metadata and only redistribute media whose individual licenses permit such use.

Xiachufang. Xiachufang is a community-based cooking platform where users share recipes and cooking instructions, providing naturally aligned procedural text-image sequences for the Tutorial in IntReal. According to Xiachufang's community guidelines, the copyright of recipes, images, and textual content published by users remains with the corresponding content creators. The platform permits third parties to reproduce user-generated content for non-commercial purposes provided that the original author and Xiachufang are clearly attributed, the original source link is retained. Such permission is subject to individual users' copyright declarations. During data curation, we inspect the available copyright and attribution metadata and filter out samples for which the user explicitly prohibits redistribution or whose licensing conditions are incompatible with our intended research use. For each retained image, we preserve its original source URL, and the corresponding original textual content is also retained for provenance and attribution. The VLM-polished text introduced during data curation is used only as training supervision. All retained data are used solely for non-commercial academic research.

Curious George. Curious George is an educational animated series centered on the adventures of George, a curious monkey who explores everyday problems through observation, experimentation, and problem solving. The series is designed to introduce young audiences to concepts in science, engineering, and mathematics, and is distributed through PBS KIDS. Curious George is a production of Imagine, WGBH, and Universal. The characters created by Margret and H.A. Rey are copyrighted and trademarked by Houghton Mifflin Harcourt and used under license from Universal Studios Licensing LLC. The corresponding PBS terms of use permit downloading materials for non-commercial, personal use with copyright notices retained, and additionally allow limited reproduction of website materials by schools and libraries for educational purposes.

Rabbids Invasion. Rabbids Invasion is an animated comedy series featuring the Rabbids, a group of mischievous rabbit-like characters whose episodic adventures form short, visually driven stories. We use episodes from the series as a source of naturally occurring interleaved story content. The series is currently distributed by Animaj, whose official page lists Rabbids Invasion as a four-season children's animated series and identifies its role as distribution. The audiovisual content remains subject to the copyright and intellectual-property rights of its respective rights holders. We use the collected material exclusively for non-commercial academic research and retain provenance information linking each sample to its original source.

The Land Before Time. The Land Before Time is a long-running animated franchise following a group of young dinosaurs, including Littlefoot, Cera, Ducky, Petrie, and Spike, through narrative adventures centered on friendship, exploration, and cooperation. We derive story samples from episodes and films in the franchise to represent long-form narrative interleaved content. The franchise is distributed by Universal Pictures, and the corresponding official releases are copyrighted by Universal Studios with all rights reserved. We use the collected material only for non-commercial academic research and maintain the provenance of each derived sample.

For the three animated series above, we follow the data organization and release practices of StoryStream (Yang et al., 2025), which has publicly released image–text samples derived from the same sources. We sincerely thank the creators, contributors, and maintainers of all data sources used in IntReal. The provenance and applicable licensing information for each source are described above. All data collected in IntReal are intended solely for non-commercial academic research, and we respect the intellectual property rights of the original content creators. For openly licensed content, we follow the corresponding license requirements, including attribution and redistribution conditions where applicable. Users of IntReal are responsible for complying with the copyright, licensing terms, and usage policies of the respective original data sources.

A.2 Detailed Filtering Pipeline

We apply a unified multi-stage filtering pipeline to all collected sources, while adapting several filtering criteria to the characteristics of individual domains. After filtering, we use VLMs to caption the remaining high-quality visual content, followed by LLM-based refinement and annotation grounded

in the original real-world samples. This process improves text-image consistency and converts the collected content into trainable supervision while preserving its factual authenticity.

Filtering Rules. For metadata-based filtering, we retain images with aspect ratios between 1:2 and 2:1 and exclude unsupported formats, including GIF and WebP. We further impose source-specific minimum resolution requirements on the shorter side of each image. Specifically, we require a minimum shorter-side resolution of 512 pixels for Wikivoyage, Wikitravel, Wikipedia, and Xiachufang, while using a threshold of 384 pixels for the three story sources. The relaxed threshold for story data accounts for the relatively low native resolution of older animated content, whose shorter side frequently falls below 512 pixels.

For representation-based filtering, we adapt the activation of different filters according to the characteristics of each source. Since Xiachufang primarily consists of user-uploaded content, we apply both face and watermark filtering to remove potentially undesirable user-generated images. We fine-tune one SigLIP2-based (Tschannen et al., 2025) binary classifiers for these tasks using 5000 manually annotated Xiachufang samples, achieving an F1 score of 0.82 on the held-out test set. In contrast, we do not apply face filtering to the three Wiki-based sources, since human faces often constitute semantically necessary content, such as portraits of notable individuals, while these sources already undergo platform-level content moderation and rarely contain intrusive watermarks. Similarly, face filtering is not applied to the three story sources, where recurring characters constitute essential visual subjects and should be preserved.

Across all sources, we additionally apply RALI-based (Zhao et al., 2026) image quality filtering and DINO-based intra-sequence deduplication. For RALI, we require each individual image to obtain a score above 3.0 and the average score of all images within a sample to exceed 3.5. For DINO-based deduplication, adjacent images are required to have a feature similarity below 0.85, which removes highly repetitive visual sequences while preserving meaningful visual continuity.

Content Refinement and Annotation. After filtering, all retained samples undergo a four-stage refinement and annotation pipeline to improve text-image alignment while preserving the factual content of the original sources. First, we generate detailed captions for each image to recover visual information that may be absent from the surrounding text. Second, we refine the original textual content by incorporating complementary information from the generated captions, while explicitly preserving the facts and semantics of the source content. Third, we construct hierarchical annotations at both the sequence and image levels: for each interleaved sample, we annotate its user prompt and generate a sequence-level reasoning trace, while each individual image is additionally associated with an image-specific reasoning trace describing its role and generation rationale within the sequence. Finally, we jointly refine the complete prompt and interleaved content to further improve their global consistency and coherence. Through this process, the collected web content is transformed into training-ready interleaved data with improved multimodal alignment, structured supervision, and preserved factual fidelity.

A.3 Dataset Statistics

Table 4 presents the statistics of IntReal-100K across different categories and data sources. Different categories exhibit distinct structural characteristics. Planning data are relatively concise, with fewer images and shorter text per sample, while Knowledge and Tutorial data contain richer multimodal content. Xiachufang constitutes the largest individual source, whereas Story samples generally contain the longest interleaved sequences with around eight images per sample. At the category level, the dataset maintains a relatively balanced coverage of diverse real-world content. Overall, IntReal-100K contains 100K samples, 538K images, and 146.6M text tokens.

B Benchmark Details

B.1 Evaluation Set Construction

Initial Filtering. We first apply several basic constraints to remove samples that are unsuitable for evaluation. Specifically, we retain samples with an average image aesthetic score of at least 3.8 and between 3 and 10 output images. This ensures sufficient visual quality while avoiding samples with

Table 4: **Statistics of IntReal-100K across different data sources.** We report the number of samples, number of images, average number of images per sample, total number of text tokens, and average number of text tokens per sample. Text tokens are computed using the Qwen2.5 tokenizer.

Category	Source	# Samples	# Images	Images / Sample	# Text Tokens	Text Tokens / Sample
Planning	Wikivoyage	21,055	56,527	2.68	21.5M	1,021.13
	Wikitravel	5,055	12,254	2.42	5.2M	1,020.53
Knowledge	Wikipedia	21,138	100,198	4.74	32.2M	1,522.34
Tutorial	Xiachufang	34,771	213,818	6.15	52.3M	1,504.77
Story	Curious George	7,298	63,088	8.64	14.2M	1,941.43
	Rabbids Invasion	3,949	34,284	8.68	7.8M	1,965.36
	The Land Before Time	6,734	58,283	8.66	13.5M	2,007.22
Total	**7 sources**	**100,000**	**538,452**	**5.38**	**146.6M**	**1,466.06**

too few images to reflect interleaved generation or excessively long sequences that may introduce disproportionate evaluation cost.

Content Quality Assessment. For the remaining samples, we use Qwen3.5-397B-A17B to assess content quality along three dimensions on a 1–5 scale: *prompt specificity*, *response richness*, and *topic informativeness*. We retain samples with prompt specificity ≥ 3, response richness ≥ 4, and topic informativeness ≥ 4. Prompt specificity ensures that the prompt provides a sufficiently clear generation goal, while response richness encourages outputs with substantive and informative content. The topic informativeness criterion is used to reduce highly idiosyncratic or sparse examples, such as unusual user-created recipes or obscure encyclopedic topics, which may be difficult to evaluate reliably.

Context-Aware Rubric Verification. We further verify whether each image is sufficiently supported by its preceding interleaved context. For each target image, a VLM generates a set of visual rubrics based only on the preceding context. We require each rubric to be atomic, focusing on a single verifiable visual property without combining multiple conditions, and self-contained, such that it can be understood without referring back to the original context. In particular, ambiguous pronouns or context-dependent entity references are avoided.

We then use Qwen3.5-397B-A17B to independently verify each rubric against the corresponding image. A sample is retained only when every image satisfies at least one generated rubric and the overall rubric pass rate across the sample is at least 50%. We do not require all rubrics to be satisfied, since naturally occurring interleaved content may contain imperfect image–text correspondence. This relaxed criterion preserves meaningful contextual consistency while retaining a moderate level of difficulty in the evaluation set.

Human Validation and Sampling. The automatic filtering pipeline produces approximately 1.5K candidate samples. The authors then manually review these candidates according to the same criteria and remove remaining failure cases. Finally, we perform stratified sampling across the four categories and different output image counts to obtain a balanced evaluation set of 600 samples.

B.2 Metric Definitions

We provide detailed definitions of the seven metrics used in IntReal. Unless otherwise specified, each metric is normalized to $[0, 1]$ and multiplied by 100 for reporting, with the convention that higher values consistently indicate better performance across all metrics. We denote the model-generated and reference texts of a sample as T^m and T^g, and the generated and reference images at position i as I_i^m and I_i^g, respectively.

Text Richness (TR). Text Richness measures whether the generated textual response is at least as informative and complete as the corresponding real-world response. An LLM compares T^m and T^g through pairwise ranking, considering informativeness, completeness, and unnecessary redundancy. To reduce positional bias, we alternate the order of the generated and reference responses during

evaluation. We treat both a model win and a tie as successful, since the objective is to determine whether the generated text reaches the quality of real-world content:

$$\mathrm{TR} = \begin{cases} 1, & T^m \text{ wins or ties}, \\ 0, & T^g \text{ wins}. \end{cases} \tag{1}$$

Text Logicality (TL). Text Logicality evaluates the internal coherence and factual plausibility of the generated interleaved response. The LLM checks whether the response follows the entities and constraints specified in the prompt, avoids contextual contradictions or abrupt logical transitions, and remains consistent with common knowledge for non-creative tasks. The score is defined as

$$\mathrm{TL} = \begin{cases} 1, & T^m \text{ is judged coherent}, \\ 0, & \text{otherwise}. \end{cases} \tag{2}$$

Image Aesthetic (IA). Image Aesthetic measures the perceptual quality of generated images using RALI (Zhao et al., 2026). Given a raw aesthetic score $a_i \in [1, 5]$ for each generated image, we first normalize it to $[0, 1]$ and then average over all generated images:

$$\mathrm{IA} = \frac{1}{N} \sum_{i=1}^{N} \frac{a_i - 1}{4}. \tag{3}$$

Image Diversity (ID). Image Diversity measures visual repetition between consecutive generated images. We extract normalized DINOv3 (Siméoni et al., 2025) features and compute the cosine similarity s_i between each pair of adjacent images. Rather than rewarding arbitrary visual differences, we only penalize excessive similarity beyond a threshold $\tau = 0.85$. The score is determined by the most repetitive adjacent pair:

$$\mathrm{ID} = \begin{cases} 1, & s_{\max} \le \tau, \\ \dfrac{1 - s_{\max}}{1 - \tau}, & s_{\max} > \tau, \end{cases} \qquad s_{\max} = \max_i \cos\left(f(I_i^m), f(I_{i+1}^m)\right). \tag{4}$$

Quantity Compliance (QC). Quantity Compliance evaluates whether the model generates the expected number of images. Let N_m and N_g denote the numbers of generated and reference images, respectively. We adopt strict exact matching:

$$\mathrm{QC} = \begin{cases} 1, & N_m = N_g, \\ 0, & N_m \neq N_g. \end{cases} \tag{5}$$

Context Fidelity (CF). Context Fidelity measures whether a generated image satisfies the visual requirements implied by its preceding interleaved context. For each position i, we reuse the context-derived visual rubrics constructed during evaluation-set curation and ask a VLM to verify them against the generated image. Let R_i denote the number of rubrics and K_i the number of satisfied rubrics. We define the position-level score as:

$$\mathrm{CF}_i = \begin{cases} 0, & K_i = 0, \\ 0.5 + \dfrac{K_i - 1}{R_i - 1} \times 0.5, & K_i \ge 1. \end{cases} \tag{6}$$

The sample-level score is then

$$\mathrm{CF} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{CF}_i. \tag{7}$$

This scoring function explicitly distinguishes complete contextual failure from partial satisfaction: satisfying none of the expected visual properties receives zero, while satisfying at least one establishes partial contextual grounding, with the score increasing toward one as additional rubrics are fulfilled. In practice, $R_i \geq 2$ for all positions in the evaluation set, so eq. (6) is always well-defined.

Subject Consistency (SC). Subject Consistency evaluates whether the subject generated at each position resembles the corresponding subject in the real-world next image. This metric is evaluated under the fine-grained next-image generation setting, where the model and the real-world sample share the same preceding interleaved context. We therefore use the position-aligned real image as the visual reference and compute the cosine similarity between DINOv3 (Siméoni et al., 2025) features of the generated and reference images:

$$\mathrm{SC} = \frac{1}{N} \sum_i \cos\left(f(I_i^m), f(I_i^g)\right). \tag{8}$$

A higher score indicates that the generated subject better preserves the visual identity or appearance of the subject that naturally occurs at the corresponding position in the real-world sequence. We apply this metric to Story and Tutorial, where persistent or semantically corresponding visual subjects are meaningful.

Overall Score. For each metric, we first average scores over all samples for which the metric is valid. The overall IntReal score is the equally weighted mean of the seven metric scores:

$$\mathrm{Overall} = \frac{\mathrm{TR} + \mathrm{TL} + \mathrm{IA} + \mathrm{ID} + \mathrm{QC} + \mathrm{CF} + \mathrm{SC}}{7}. \tag{9}$$

We emphasize that this overall score is intended primarily as a convenient reference for coarse-grained comparison across models, rather than as a definitive measure of generation quality. Since interleaved generation is a highly multi-faced and open-ended task, we recommend examining performance along individual dimensions to better understand a model's specific strengths and weaknesses, rather than relying solely on the aggregated score.

B.3 Evaluation Prompts

Prompt for Text Richness Evaluation

You are an expert evaluator comparing two text responses to the same prompt.
Task: Given a PROMPT and two RESPONSE texts, evaluate which response is better in terms of informativeness, completeness, and conciseness.
Evaluation Criteria:
1. *Informativeness:* Does the response provide meaningful and useful content that genuinely helps the user?
2. *Completeness:* Does the response adequately address the full scope of the prompt without leaving key aspects unaddressed?
3. *Conciseness:* Does the response avoid unnecessary repetition, filler content, or verbosity that dilutes the core information?
Important Rules:
- Evaluate ONLY the textual components. Ignore images or image placeholders.
- Compare the two responses directly against each other.
- Do NOT let response length bias your judgment. A shorter response can be better if it is more precise.
- You MUST pick a winner (1 or 2) unless the two responses are truly indistinguishable in quality. Only output winner: 0 as a last resort when no meaningful difference exists.
Output Format:
{"winner": 1} if RESPONSE1 is better.
{"winner": 2} if RESPONSE2 is better.
{"winner": 0} if the two are roughly equal in quality.

Prompt for Text Logicality Evaluation

You are an evaluator for contextual consistency in interleaved text-image generation tasks.
Task: Determine whether the MODEL RESPONSE is logically consistent, coherent, and free from hallucinations or logical leaps based on the PROMPT and contextual common sense.
Evaluation Criteria:
1. *Context consistency:* The response must strictly follow the context, entities, facts, and constraints provided in the prompt. It should not contradict prior information.
2. *Logical coherence:* The response must be a smooth and sensible continuation. It should focus on the same topic/event without abrupt topic shifts or sudden logical leaps.
3. *Real-world Factuality:* For creative/fictional tasks (e.g., storytelling, sci-fi), allow for imaginative elements, judging only by internal consistency. For non-creative/factual tasks, the response must not severely contradict common sense or real-world logic.
Important: The model response is extracted from an interleaved text-image sequence. It may contain placeholders like '[image]' or other special tokens. You should IGNORE these placeholders and evaluate ONLY the logical coherence and factuality of the actual text content.
Output Format: You must output ONLY a single valid JSON object and nothing else. No explanations, no analysis, no markdown.
{"coherent": true} if the model response is consistent.
{"coherent": false} if the model response is inconsistent.

Prompt for Context Fidelity Evaluation

You are an image checker. Given an image and a numbered checklist of rubric items, verify each item against what you actually see in the image.
For each rubric item:
- Mark true if the described object, action, state, or property is satisfied.
- Mark false if the item is clearly absent or completely contradicted by the image.
- Err on the side of false when in doubt.
Output EXACTLY: {"passes": [true, false, true, ...]} The boolean array MUST have the same length as the number of rubric items (one boolean per item, in order).
CRITICAL RULES:
- Output ONLY the JSON object. No markdown, no code fences, no preamble, no extra text.
- The "passes" array must contain only booleans (true/false, not strings).
- Array length MUST exactly equal the number of rubric items.

B.4 Metric Calibration Cases

We present representative low-scoring examples from different models to illustrate the failure patterns captured by each metric.

Prompt: *Can you give me a detailed guide on the Cetonia aurata beetle? I'd love to learn about its physical features, how to tell it apart from similar species, and the science behind its metallic color.*

The *Cetonia aurata*, commonly known as the goldored chafer, is a magnificent member of the Scarabaeidae family, renowned for its dazzling appearance. This beetle is a true jewel of the insect world, boasting a glossy, metallic exoskeleton that shifts through shades of copper, bronze, and gold depending on the angle of the light...
Identifying the goldored chafer can be tricky because it is frequently confused with the *Cetonia aurata* var. *rubrolineata*. However, keen observers can distinguish them by specific markings. The var. *rubrolineata* typically features a distinct red line running along the side of the body, a subtle but crucial detail...

Why Low Text Richness: The specific color characteristics of the beetle (such as a green body) were not clearly stated. When comparing with similar insects, only one aspect was examined.

Figure 8: **Calibration case for Text Richness.** A low-scoring sample from **Bagel+SFT**, where the response fails to provide sufficiently specific insect characteristics and offers only limited detail when distinguishing the target insect from similar species, leading to a low **Text Richness** score.

Prompt: *Can you put together a travel guide for Zarasai, Lithuania? I'd love to learn about its famous lakes, the ancient Stelmužė Oak, and other key landmarks in the area.*

... Your journey begins in the heart of the town, where the history of the region is palpable. The centerpiece is the legendary Stelmužė Oak, a massive, gnarled tree estimated to be over 1,000 years old. Its sprawling canopy dominates the square, standing as a symbol of the town's resilience and longevity amidst the traditional wooden houses... ...Our final destination takes us away from the town and into the pristine wilderness of the Zarasai lakes. The journey culminates at Lake Luodis, a place of unparalleled beauty. From the peaceful observation bridge, you can take in the crystal-clear waters and the lush greenery of the surrounding forests, offering a sense of natural harmony and tranquility.

Why Low Text Logicality: It incorrectly places the Stelmužė Oak in central Zarasai, and conflates Lake Luodis with the Zarasai observation bridge, which actually overlooks Lake Zarasas.

Figure 9: **Calibration case for Text Logicality.** A low-scoring sample from **U1-8B-MoT**, where spatial and factual inconsistencies in the description lead to a low **Text Logicality** score.

Prompt: *I'd love to learn how to make clay pot rice with cured meats. Could you walk me through the whole process from prepping the ingredients to the final dish, showing me what each step looks like along the way?*



Figure 10: **Calibration case for Image Aesthetic.** A low-scoring sample from **U1-8B-MoT-SFT**, where image quality degrades over steps, with later images exhibiting excessively strong saturation and contrast and appearing less photorealistic, leading to a low **Image Aesthetic** score.

Prompt: *I'd love to learn how to make clay pot rice with cured meats. Could you walk me through the whole process from prepping the ingredients to the final dish, showing me what each step looks like along the way?*



Figure 11: **Calibration case for Image Diversity.** A low-scoring sample from **Emu3.5**, where Images 1&2, 3&4, and 5&6 are excessively similar to each other, resulting in limited visual variation and a low **Image Diversity** score.

Prompt: *Could you walk me through how to make pineapple roasted chicken wings? I'd love to see the process from prepping the ingredients to the final plating. I need **8** images in your answer.*



Figure 12: **Calibration case for Quantity Compliance.** A low-scoring sample from **Bagel+SFT+RL**, where the prompt requests eight images but the response generates only six, leading to a low **Quantity Compliance** score.

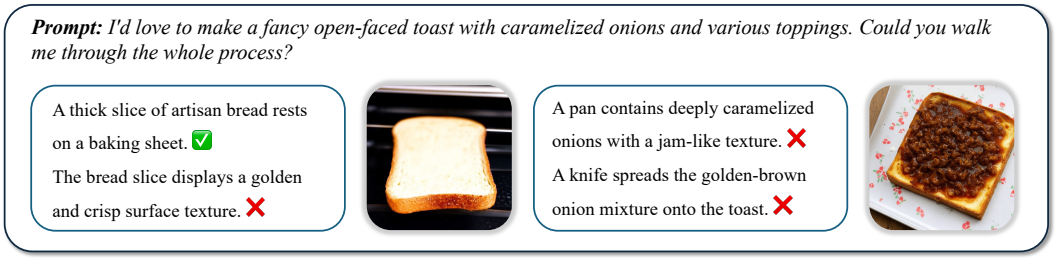


Figure 13: **Calibration case for Context Fidelity.** A low-scoring sample from **U1-8B-MoT+SFT+RL**, where the generated images fail to faithfully follow the described cooking process: the bread is insufficiently toasted in the first image, while the second image omits the preparation and spreading of the caramelized onions, leading to a low **Context Fidelity** score.



Figure 14: **Calibration case for Subject Consistency.** A low-scoring sample from **U1-A3B-MoT**, where the main subjects are not consistently preserved across images. For example, the green dinosaur differs across Images 1, 4, and 5, while the teddy bear changes color between Images 3 and 6. Since the reference sequence maintains a stable subject appearance throughout, such internal inconsistency in the generated images necessarily manifests as low similarity to the reference at the corresponding positions, leading to a low Subject Consistency score.

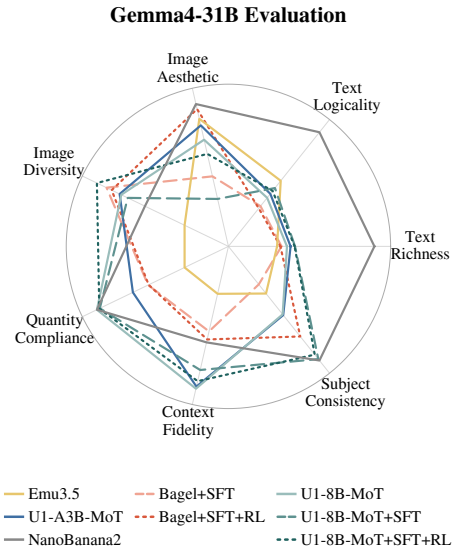


Figure 15: **IntReal evaluation results with Gemma4-31B as the evaluator.** We compare models across seven evaluation dimensions.

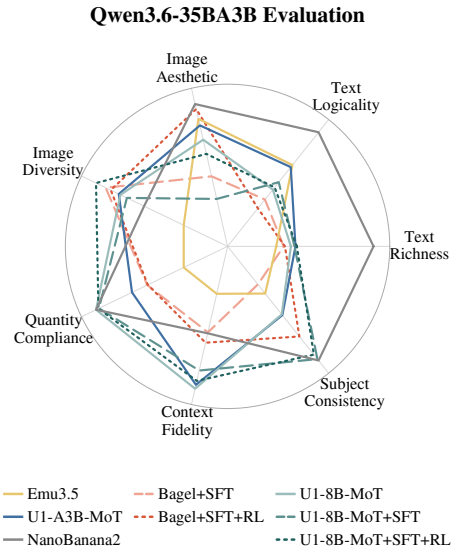


Figure 16: **IntReal evaluation results with an alternative evaluator.** We use Qwen3.6-35B-A3B in place of Gemma4-31B as the evaluator.

Table 5: **SFT configurations for Bagel and SenseNova-U1.** We report the main optimization, batching, loss, and distributed training hyperparameters used in our post-training experiments.

Hyperparameter	**Bagel**	**SenseNova-U1**
Initialization	BAGEL-7B-MoT	SenseNova-U1-8B-MoT-Interleaved
Training steps	6,700	7,000
Optimizer	AdamW	AdamW
Learning rate	2×10^{-5}	2×10^{-5}
LR schedule	Cosine	Cosine
Warmup steps	670	700
Adam β_1/β_2	0.9 / 0.95	0.9 / 0.95
Adam ϵ	1×10^{-15}	1×10^{-8}
Weight decay	0	0
Gradient clipping norm	1.0	1.0
Gradient accumulation steps	4	4
Microbatch size / GPU	1 packed sequence	1 packed sequence
Packed Sequence Length	36,000	18,000
Text / Image Loss Weights	1.0 / 1.0	1.0 / 1.0
EMA decay	0.995	0.999
Distributed configuration	FSDP, 8 shards	ISP, WP=8, TP=1, PP=1, ZeRO-1

Table 6: **RL configurations for Bagel and SenseNova-U1.** We report the main rollout, stochastic sampling, reward, and evaluation configurations used in our reinforcement learning experiments.

Hyperparameter	**Bagel**	**SenseNova-U1**
Training resolution	1024	1024
Rollout denoising steps	10	20
Stochastic scheduler	Flow-CPS	Flow-SDE
Noise level	0.7	0.4
Candidate stochastic step indices	$\{1, 2, 3\}$	$\{1, 2, 3\}$
Selected stochastic steps / rollout	1	1
Rollout text CFG	1.0	4.0
Rollout image CFG	1.0	1.0
Reward weights (RALI / DINO / Rubric)	1.5 / 0.5 / 1.0	1.0 / – / –
Evaluation denoising steps	50	50

C Experiment Details

C.1 Model Training Settings

We build on the official codebases of Bagel and SenseNova-U1 for SFT, with their hyperparameters summarized in Table 5. The two models use different visual representations and consequently organize their visual tokens in the KV cache differently. In Bagel, reference images in the prompt are represented by both VAE latent tokens and ViT embeddings, whereas generated images use only VAE latent tokens. SenseNova-U1 instead performs visual understanding and generation in pixel space, using representations derived from image patches. To accommodate their different visual token budgets, we adopt model-specific smart resizing strategies that preserve aspect ratios and align spatial dimensions to the required grids. For Bagel, the VAE pathway uses a maximum-edge limit of 768 pixels and a target minimum short edge of 512 pixels, with dimensions aligned to multiples of 16; the ViT pathway uses corresponding settings of 518, 224, and 14. For SenseNova-U1, native-resolution resizing constrains image areas to 512^2–1024^2 pixels for understanding and 768^2–1024^2 pixels for generation, with dimensions aligned to the required patch grid.

Detailed RL hyperparameters are provided in Table 6. We initialize Bagel and SenseNova-U1 from their SFT checkpoints at steps 4,500 and 6,900, respectively, using non-EMA weights. Given the limited SFT data and training steps, we found that EMA weights lagged behind the updated language branch, resulting in underfitting. We then perform RL post-training using Flow-Factory. For Bagel,

Table 7: **IntReal evaluation results with an alternative evaluator.** We report results following the protocol in table 3, using Qwen3.6-35B-A3B in place of Gemma4 as the evaluator.

Model	Holistic Evaluation					Fine-grained Evaluation		Mean
	Text Richness	Text Logicality	Image Aesthetic	Image Diversity	Quantity Compliance	Context Fidelity	Subject Consistency	
Emu3.5	13.34	**69.93**	81.20	75.51	39.70	71.06	61.46	58.89
U1-A3B-MoT	29.50	69.00	80.70	88.70	72.50	89.11	63.45	70.42
Bagel+*SFT*	19.67	54.33	76.82	91.35	62.83	78.71	60.61	63.48
Bagel+*SFT+RL*	20.67	50.33	**81.94**	90.31	62.50	80.76	65.43	64.56
U1-8B-MoT	25.67	58.83	79.60	88.36	**95.33**	**89.88**	63.38	71.58
+*SFT*	28.67	62.17	75.10	87.05	93.50	86.28	**67.52**	71.47
Δ vs. base (%)	*+11.69%*	*+5.68%*	*−5.65%*	*−1.48%*	*−1.92%*	*−4.11%*	*+6.53%*	*−0.15%*
+*SFT+RL*	**30.67**	59.83	78.53	**93.31**	93.83	88.34	67.11	**73.09**
Δ vs. base (%)	*+19.48%*	*+1.70%*	*−1.34%*	*+5.60%*	*−1.57%*	*−1.82%*	*+5.89%*	*+2.11%*
NanoBanana2	94.98	84.78	82.32	83.56	95.32	78.89	67.65	83.93
Δ vs. OSS Best (%)	+209.68%	+21.24%	+0.46%	-10.45%	-0.01%	-12.32%	+0.19%	+14.83%

we adopt Flow-CPS Wang & Yu (2025) as the stochastic sampler. For SenseNova-U1, we retain Flow-SDE, since Flow-CPS is not implemented for U1 in our training setup and its applicability to pixel-space diffusion remains unverified. We set the stochastic noise level α to 0.7 and 0.4 for Bagel and SenseNova-U1, respectively. Following MixGRPO (Li et al., 2025a), we accelerate training by randomly selecting a single step among the first three denoising steps for an SDE transition and policy optimization, while using deterministic transitions elsewhere. For stability, Bagel uses no CFG during rollout or policy updates. For SenseNova-U1, we adopt Hybrid CFG following Qwen-Image-2.0-RL (Xu et al., 2026), which we found more effective in our preliminary experiments. We emphasize aesthetic rewards for both models, motivated by our observation that improved visual quality and more faithful object rendering are accompanied by better text-image alignment and subject consistency. For experiments involving multiple reward signals, we adopt GDPO (Liu et al., 2026b) to balance and optimize multiple reward objectives.

SFT experiments are conducted on a single node of 8×H800-80G GPUs. RL experiments are conducted on two nodes of 8×H800-80G GPUs (16 GPUs in total) to accelerate policy optimization.

C.2 Model Inference Settings

During inference, models generate text autoregressively and switch to image generation when the special `<image>` token is produced. We follow the reasoning and generation template used by Emu3.5 (Cui et al., 2025) for interleaved generation. Unless otherwise specified, all generated images are produced at a resolution of 1024×1024 for both holistic and fine-grained evaluation. We recommend 1024×1024 as the default inference resolution, as it provides sufficient visual fidelity while avoiding excessive growth in visual tokens and the resulting interleaved context length.

Most models are successfully evaluated on all 600 samples in the evaluation set. However, for Emu3.5, eight samples exceed the maximum context length or memory limits even on 8×H800-80G GPUs, so we report its results on the remaining 592 samples. For NanoBanana2, two samples repeatedly triggered the API's built-in safety guardrails and could not be processed, so we report its results on the remaining 598 samples. Manual inspection revealed no apparent safety concerns in these two samples, suggesting that the guardrails may have been falsely triggered.

C.3 Detailed Evaluation Results

Evaluation results with Qwen3.6. We repeat the evaluation in Table 3, replacing Gemma4 with Qwen3.6-35B-A3B as the evaluator. The resulting scores are reported in Table 7. Across most dimensions, the two evaluators produce nearly identical scores, and model rankings remain highly consistent, corroborating the robustness analysis in section 4.4. The one notable exception is Text Logicality, which shows larger absolute discrepancies between the two evaluators (e.g., a 12.71-point gap for NanoBanana2). We attribute this to the binary nature of the Text Logicality judgment: unlike Text Richness, which relies on a direct pairwise comparison against a reference response, Text Logicality requires each evaluator to independently determine whether a response is coherent,

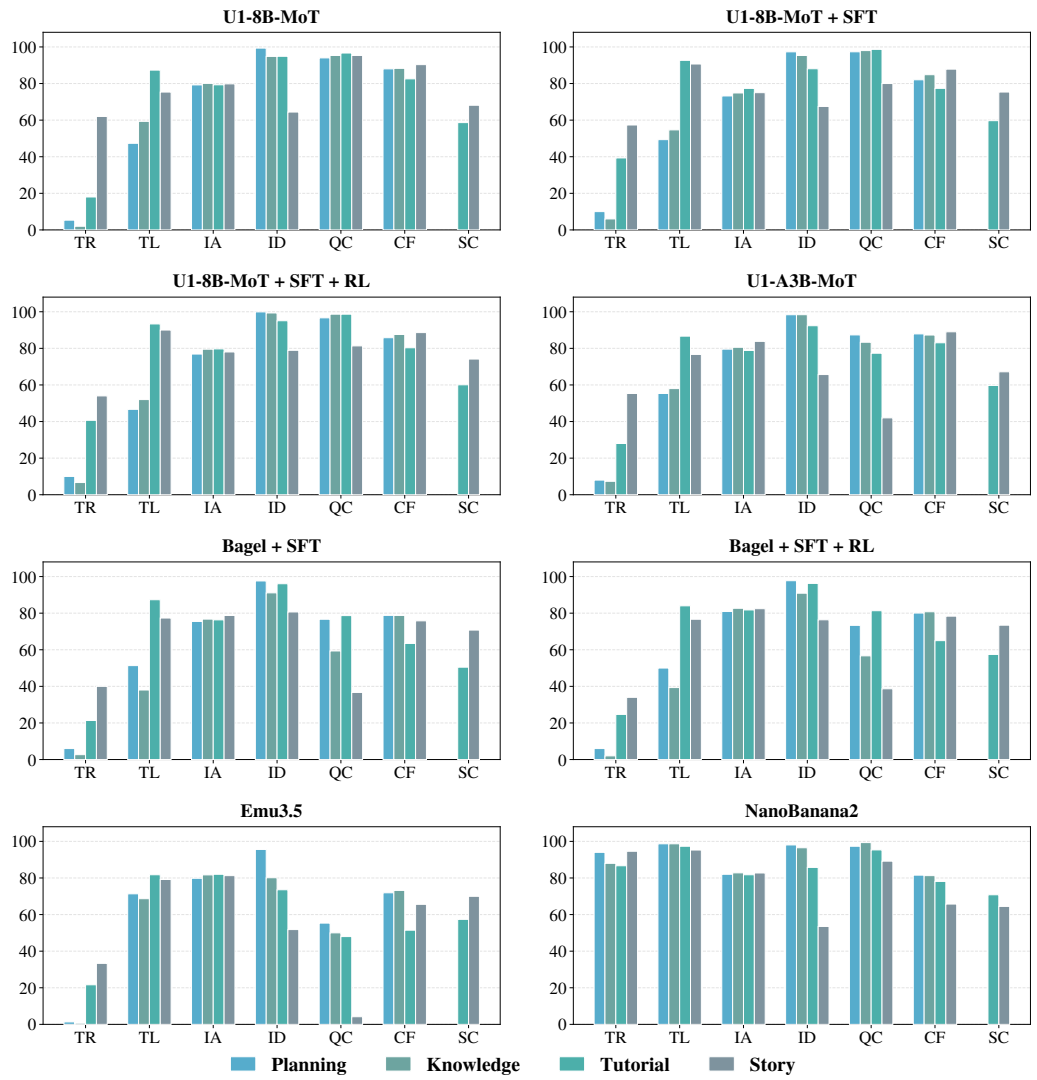


Figure 17: **IntReal evaluation results per category for individual models.** We report the score of each metric for each category separately for each model.

without an explicit comparative anchor. As a result, each evaluator's own internal calibration of coherence introduces a systematic offset in absolute scores. Despite this discrepancy in absolute values, the relative ranking of models under Text Logicality remains largely preserved across the two evaluators, indicating that our conclusions are not sensitive to the specific evaluator model used. The radar plots in figs. 15 and 16 further show highly similar performance profiles across the two evaluators, providing visual evidence of the robustness of our evaluation.

Category-wise Evaluation Results. Figure 3 summarizes the category-wise results of 5 representative models (U1-8B-MoT, Bagel+SFT, U1-A3B-MoT, Emu3.5, and NanoBanana2), while Figure 17 provides a detailed breakdown for each model. Open-source models generally exhibit category-wise patterns consistent with the aggregate results: Planning and Knowledge pose greater challenges in Text Richness and Text Logicality, Tutorial is more challenging in Context Fidelity and Subject Consistency, and Story presents greater difficulties in Quantity Compliance and Image Diversity. The closed-source NanoBanana2 exhibits different patterns on some dimensions, particularly Text Richness and Text Logicality, where its consistently strong performance across categories suggests reduced sensitivity to these category-specific challenges.

D QUALITATIVE EXAMPLES

Prompt: Can you give me an overview of the geography and history of Mayfair and Marylebone in London? I'd love to learn about their key landmarks, how they developed over time, and the best ways to get around using the Underground. It would be great if you could include some visuals that capture the classical architecture on the busy streets, the historic feel of Grosvenor Square, and the peaceful atmosphere of Regent's Park. Please provide 3 images in your answer.

Model:

Mayfair and Marylebone are two of London's most sophisticated and historically rich districts, situated in the West End. Here is an overview of their geography, history, and key landmarks, along with visuals of their unique character.



Mayfair is bounded by some of London's iconic landmarks: Oxford Street to the north, Regent Street to the east, Piccadilly and Green Park to the south, and Park Lane and Hyde Park to the west. It has long been synonymous with luxury and exclusivity.**Marylebone** (pronounced *Mar-lee-bone*) lies just north of Mayfair, separated by the hustle of Oxford Street. Its boundaries are Marylebone Road to the north, Great Portland Street to the east, Oxford Street to the south, and Edgware Road to the west. This area blends high-end residential streets with bustling medical districts and cultural institutions.

Both districts are defined by stunning classical architecture, which sets them apart from the more modern developments in other parts of London. The visual journey through these areas is one of looking upward at detailed facades while navigating vibrant streets. This juxtaposition is captured in the image below, which shows a typical view near Oxford Street, where the wide thoroughfare is framed by magnificent 18th and 19th-century buildings.

Mayfair's name derives from the annual fortnight-long "May Fair" that took place in Shepherd Market from 1686 until it was banned in 1764. The development of the area began in earnest in the late 17th century under the Grosvenor family (who still own much of the freehold). They transformed it into London's premier aristocratic residential district. Grosvenor Square is the historic heart of Mayfair. This square, developed in the 1720s, was long considered the most prestigious address in London, home to numerous dukes and statesmen.



Historic Landmark: Grosvenor Square: The square retains its historic, upscale feel, characterized by the elegant, uniform red-brick and stone architecture that the Grosvenor Estate established. For many decades, it was also famous as the location of the US Embassy, which gave it a unique diplomatic character. The image below captures this prestigious atmosphere:



Marylebone developed slightly later than Mayfair, primarily in the 18th century. It was also formed by great estates, notably the Portman Estate and the Portland Estate (now the Howard de Walden Estate). This structured, grid-like development is visible in the layout of Harley Street (famous for private medicine) and Wimpole Street

… …

Figure 18: **Qualitative example from the Planning category generated by Nanobanana2.** The response interleaves text and images to introduce the geography, history, and landmarks of Mayfair and Marylebone.

Prompt: Can you give me a detailed overview of the copper demoiselle damselfly, covering its appearance, mating habits, and how scientists study them? I'd love to see some pictures illustrating the different life stages and behaviors you mention. Please provide 4 images in your answer.

Model:

The sun rose over the wetland, casting a golden glow on the still water. A copper demoiselle damselfly, its wings shimmering like polished copper and its body a vibrant emerald, perched on a cattail. Tiny black eyes scanned the surface, searching for a mate.



The female, slightly larger, joined him. A delicate dance began as they twined their bodies in a synchronized display. The male's vibrant colors flashed against the reeds, a bold assertion of his fitness. Their movements were fluid and precise—a testament to millions of years of evolution.





Together, they selected a submerged stem of pondweed. The female inserted her ovipositor and laid her eggs, one by one, into the soft tissue. The male hovered protectively nearby, ready to chase away any threats. Their work complete, they parted ways, leaving their offspring to begin their aquatic journey.

Weeks turned into months. Beneath the water, the nymph grew, shedding its exoskeleton several times as it matured. When the time was right, it clung to a reed, its final skin splitting open. A newly emerged adult, its wings still folded and soft, began to pump hemolymph through its delicate veins, expanding and drying them in the sunlight. The copper demoiselle damselfly had completed its transformation. It rested, its wings gleaming like burnished copper, ready to take flight and join the vibrant tapestry of life in the wetland.



Figure 19: **Qualitative example from the Knowledge category generated by Emu3.5.** The response interleaves text and images to describe the appearance, reproductive behavior, and development of the copper demoiselle damselfly.

Prompt: Can you show me how to roast cherry tomatoes with garlic and rosemary, including how to arrange them and serve with bread? Make sure to include 5 images in your answer.

Model:



Roasting cherry tomatoes with garlic and rosemary is a simple yet transformative technique that concentrates their natural sweetness and creates a savory, umami-rich side dish. To begin, you'll want to prepare your baking sheet by arranging the ingredients in a single, even layer. Place halved cherry tomatoes cut-side up so their interiors are exposed to the heat. Scatter whole, unpeeled garlic cloves and fresh sprigs of rosemary among the tomatoes. This specific arrangement ensures that the garlic infuses the oil without burning and the rosemary distributes its aromatic oils evenly throughout the dish.



Once your tray is prepped, slide it into an oven preheated to 180°C (350°F). The roasting process typically takes about 30 minutes, though you may want to adjust the time slightly depending on the size of your tomatoes and your specific oven. As they roast, the tomatoes will release their juices, soften, and eventually develop slightly wrinkled skins, while the garlic cloves will turn a beautiful golden brown and become tender enough to squeeze right out of their skins.



When you remove the tray from the oven, the dish is almost ready, but a few final touches elevate it from good to exceptional. Immediately season the hot tomatoes and garlic with a generous pinch of flaky sea salt and freshly cracked black pepper. Then, drizzle a liberal amount of high-quality olive oil over the mixture. The heat from the vegetables will warm the oil, helping it penetrate the softened tomatoes and creating a glossy, flavorful coating that ties all the herbs and aromatics together.



To serve, this dish shines when paired with crusty bread, such as baguette slices or chunks of ciabatta. The bread acts as the perfect vehicle for the rich, savory juices released during roasting. Simply place the warm tomatoes and garlic on a serving platter alongside the bread, allowing the flavors to meld at the table. The rustic texture of the bread complements the soft, jammy consistency of the roasted tomatoes perfectly.



The final result is a delightful interplay of textures and flavors. As you take a bite, you'll experience the burst of sweet, concentrated tomato juice, the mellow pungency of the roasted garlic, and the earthy note of rosemary, all absorbed by the crisp, airy crumb of the bread. It's a simple combination that feels both elegant and comforting, ideal for appetizers, bruschetta, or a vibrant side dish.

Enjoy your homemade roasted cherry tomatoes while they are still warm for the best flavor experience. This dish proves that with just a few fresh ingredients and a little patience, you can create something truly delicious.

Figure 20: **Qualitative example from the Tutorial category generated by U1-8B-MoT+SFT+RL.** The response interleaves text and images to explain how to roast cherry tomatoes with garlic and rosemary and serve them with bread.

Prompt: Tell me a whimsical story about a group of dinosaurs exploring a lush jungle, and include several illustrations along the way to bring their adventure to life. I need 8 images in your answer.

Model:



Once upon a time, in a jungle so lush it seemed to breathe with life, a diverse group of dinosaurs gathered on a sun-dappled dirt path. It was a colorful assembly: a massive brown sauropod stood tall, a small purple theropod looked up with curiosity, and a green armored dinosaur carried a tiny yellow hatchling on its back. They weren't alone for long; a pink pterosaur with a vibrant red crest swooped down from the canopy, joining the circle. Together, they shared a moment of pure joy, their laughter echoing through the trees as they prepared for the day's adventure.



The group soon ventured deeper into the verdant wilderness, where the sunlight filtered through the leaves in golden beams. The large brown sauropod led the way, his long neck scanning the horizon, while the smaller purple theropod trotted alongside, eager to see what lay ahead. They passed a massive tree trunk and a mossy rock, the air thick with the scent of ferns and damp earth. The dense foliage created a serene, almost magical atmosphere, making the jungle feel like a living storybook come to life.



Suddenly, the peaceful stroll turned into a lively chase! The small purple dinosaur, full of energy, darted ahead, its eyes wide with excitement. The large brown sauropod gave chase, his long legs pumping with determination, while the green armored dinosaur and the pink pterosaur flanked them, keeping pace. They ran through the vibrant greenery, their movements a blur of color and motion. Even a tiny yellow dinosaur peeked out from behind a fern, watching the fun with a grin, as the group moved as one united team through the wild.



After the excitement of the chase, the group decided to take a well-deserved break. The large brown sauropod found a perfect spot on a grassy patch, lying down to rest. The little purple dinosaur, tired from the run, climbed onto the giant's back, nuzzling its neck in a display of trust and affection. Nearby, the green armored dinosaur stood guard, while the pink pterosaur hovered gently above, keeping a watchful eye. It was a moment of perfect harmony, where the big and the small coexisted in peace under the bright jungle sky.



Not far away, the adventure continued for others. A small purple dinosaur with a distinctive red eye and a white snout was having the time of its life. It ran energetically across a grassy clearing, its mouth open in a joyful roar, chasing a fluttering butterfly. The lush green forest surrounded it, with tall trees and dense foliage creating a vibrant backdrop. The little dinosaur's dynamic movement and cheerful expression captured the pure spirit of play, reminding everyone that fun is the most important part of any journey.

... ...

Figure 21: **Qualitative example from the Story category generated by Bagel+SFT+RL.** The response interleaves text and images to narrate a group of dinosaurs exploring and playing together in a lush jungle.