INTERGRPO: GLOBAL-LOCAL CREDIT ASSIGNMENT FOR MULTI-UNIT INTERLEAVED TEXT–IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-unit interleaved text-image generation aims to produce coherent sequences of alternating textual and visual content, such as visual storytelling generation, requiring semantic and visual consistency across units as well as coherence of the complete sequence. Yet models trained primarily through supervised fine-tuning have limited ability to improve these qualities from generation feedback. Although recent reinforcement learning methods offer a potential remedy, existing methods primarily optimize a single image or an atomic text-image unit. Extending them to multi-unit outputs poses two challenges: a trajectory-level reward cannot identify which visual units should be reinforced, and uncertain quality judgments can provide unreliable optimization signals. To address this challenge, we introduce **InterGRPO**, an online group-based policy optimization framework tailored to multi-unit interleaved text-image generation. InterGRPO employs global-local credit assignment to optimize individual visual units while preserving sequence-level coherence. However, scoring hallucinations is widespread in general vision-language models for image quality evaluation, leading to inaccurate quality scores. To address this issue, we further develop an **U**ncertainty-**C**alibrated **Reward (UC-Reward)** model, which distills calibrated score distributions from multiple task-specific teachers. Experiments on SIL-Bench and ISG-Bench show that InterGRPO outperforms the strongest baseline by 2.5% and 4.5%, respectively, while improving cross-unit consistency, logical coherence, and overall generation quality.

1 INTRODUCTION

Recently, Unified Multimodal Models (UMMs) Deng et al. (2025); Liao et al. (2025); Cui et al. (2025) have integrated multimodal understanding and image synthesis within a single architecture. This integration has enabled multimodal interleaved reasoning Huang et al. (2026); Zhang et al. (2026); Chen et al. (2026a) and interleaved text–image generation Wang et al. (2026a); Ye et al. (2026). In the latter, textual and visual information are generated as parts of a unified sequence. Given a single multimodal sequence input, the model can autonomously construct a coherent sequence of alternating text and visual units, rather than treating language generation and image synthesis as separate tasks. Unlike conventional unimodal generation paradigms, such as language modeling and text-to-image synthesis, interleaved generation treats the two modalities as interdependent throughout generation, allowing them to mutually contextualize and complement one another while progressively shaping the structure and semantics of the resulting content. This paradigm more closely resembles the collaborative nature of human content creation and better accommodates the compositional demands of real-world applications.

Despite substantially expanding the expressive capacity of multimodal generation, existing interleaved text–image generation models are still predominantly optimized through supervised fine-tuning (SFT) on offline datasets. For multi-unit interleaved generation, SFT can learn patterns of alternating text and image outputs through a sequence-level training objective, but its optimization signal is primarily derived from matching fixed reference sequences. Consequently, model behavior remains constrained by the content distribution and modality arrangements in the training data,

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.

making it difficult to adapt the generation policy based on output quality or to accommodate diverse user preferences in open-ended settings.

Given the limitations of supervised fine-tuning in preference alignment and adaptive optimization from generation feedback, reinforcement learning (RL) has been increasingly adopted across large language modeling and visual generation, giving rise to effective policy optimization frameworks such as Liu et al. (2026b); Xue et al. (2025); Li et al. (2025); Liu et al. (2026a); Zhao et al. (2026b). Despite their success across generative tasks, existing online reinforcement fine-tuning methods are primarily designed for single-modality outputs, such as language generation or text-to-image synthesis. Their reward formulations and credit assignment mechanisms do not account for the structure of interleaved generation, in which multiple text and image units must be locally aligned and collectively form a coherent sequence. This limits their direct applicability to interleaved text–image generation. UniGRPO Liu et al. (2026c) jointly optimizes textual reasoning and subsequent image synthesis, but considers only one text–image unit. It does not address dependencies across multiple alternating units, leaving semantic and visual consistency, as well as credit assignment between unit quality and sequence coherence, unresolved. Furthermore, we observe that large vision–language models (VLMs) vary in scoring accuracy across tasks. Directly scoring interleaved text–image outputs with a VLM can also yield score distributions with high entropy and variance, indicating uncertainty in the quality assessment. Such ambiguous reward signals may provide misleading feedback during policy optimization.

To address these challenges, we introduce **InterGRPO**, a Group Relative Policy Optimization (GRPO) framework specifically designed for multi-unit interleaved text-image generation. InterGRPO directly optimizes the interleaved generation policy through multimodal reward feedback and stable online policy updates. When a generated trajectory contains both high- and low-quality visual units, broadcasting a single sequence-level advantage assigns them the same optimization signal. As a result, low-quality images may be reinforced alongside successful ones. Therefore, we proposed a global-local advantage estimation mechanism that integrates unit-level feedback with sequence-level rewards. This mechanism reformulates the conventional trajectory-level GRPO objective into a fine-grained, visual-unit-level objective, enabling effective credit assignment across multiple generated images while preserving global supervision over the entire multimodal sequence. Second, we introduce UC-Reward, an uncertainty-calibrated reward model for interleaved multimodal generation. It combines task-aware score-distribution distillation, repeated evaluation of ambiguous samples, selective human calibration, and high-quality reference anchoring to reduce the impact of scoring uncertainty and capture fine-grained quality differences. By jointly optimizing local generation quality and global sequence-level coherence, InterGRPO substantially improves cross-unit content consistency, logical continuity, and the overall naturalness of long-horizon interleaved text-image generation.

Overall, the contributions of this article are as follows:

- We introduce **InterGRPO**, an online reinforcement fine-tuning framework that extends GRPO-based optimization to multi-unit interleaved text-image generation. Its global-local credit assignment combines sequence-level and image-unit advantages to optimize each visual unit while promoting coherence across the complete sequence.
- We develop an **U**ncertainty-**Calibrated Reward (UC-Reward)** model for interleaved generation, combining task-aware score-distributional distillation with multiple rounds of evaluation, selective human calibration, and high-quality reference anchoring. This design provides more reliable preference supervision for heterogeneous interleaved outputs.
- We construct a multi-task benchmark for evaluating multi-unit interleaved text–image generation. Quantitative evaluations across multiple benchmarks show that InterGRPO improves cross-unit consistency and overall generation quality. Ablation studies further validate the effectiveness of its individual components.

2 METHOD

We introduce InterGRPO for reinforcement learning of multi-unit interleaved text–image generation. Unlike multi-turn interaction, our setting takes a single input $I = [u_1, u_2, \dots], \quad u_i \in \{c_{\text{text}}, c_{\text{img}}\}$. Under the policy π_θ, we sample a group of K candidate trajectories $\tau_k{}_{k=1}^{K}$ for relative comparison.

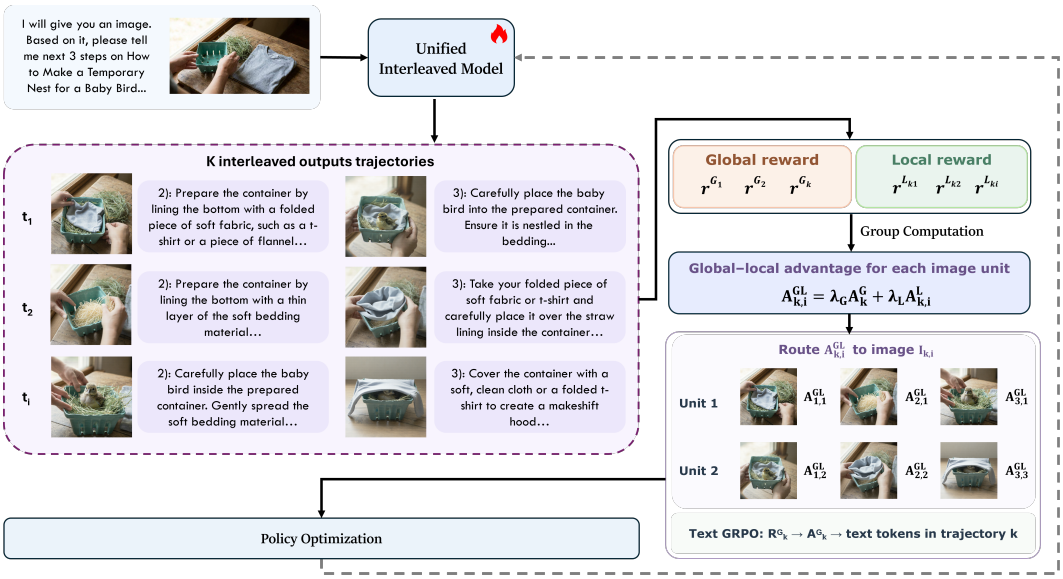


Figure 1: **Overall of InterGRPO.** Each visual unit uses its own $A_{k,i}^{\mathrm{GL}}$ for policy optimization, so that both sequence-level quality and the unit's local quality contribute to credit assignment.

Each trajectory contains U consecutively generated text-image units:

$$\tau_k = [(y_{k,i}, x_{k,i})]_{i=1}^{U}, \quad (y_{k,i}, x_{k,i}) \in \mathcal{C}, \quad \mathcal{C} = \{c_{\text{text}}, c_{\text{img}}\}$$

where τ_k denotes the k-th interleaved multimodal trajectory. The variables $y_{k,i}$ and $x_{k,i}$ denote the textual and visual outputs associated with its i-th unit, respectively. The order of text and image units is not fixed in advance. Given a single multimodal input sequence, we aim to improve the quality of the resulting multi-unit interleaved text–image output through online group-based policy optimization. Our method consists of three components. First, we formulate interleaved text-image generation as a Markov decision process, in which textual and visual units are generated sequentially within a shared multimodal context. Second, we introduce a global-local credit assignment mechanism that integrates sequence-level rewards with unit-level feedback, enabling fine-grained policy optimization while preserving global coherence. Third, to reduce unreliable quality scores from general VLMs, We develop UC-Reward by distilling teacher score distributions calibrated through a multi-stage process. We describe each component in detail below.

2.1 Interleaved Text-Image Genetation as a Markov Decision Process

We formulate interleaved text-image generation as a finite-horizon Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R)$, where the model sequentially generates multiple alternating textual and visual units from a single multimodal input sequence.

- **State Space $\mathcal{S}$:** At unit k, the state $s_k = (x, y_{<k}^{text}, y_{<k}^{img})$ comprises the input x and all previously generated textual $y_{<k}^{text}$ and visual $y_{<k}^{img}$ content .
- **Action Space $\mathcal{A}$:** An action corresponds to predicting the next text token during text generation or performing the next denoising transition during image generation.
- **Transition $\mathcal{P}$:** Given an action, the transition appends the predicted token to the textual context or advances the image latent to the next denoising step, progressively extending the interleaved sequence.
- **Reward $\mathcal{R}$:** The reward comprises a trajectory-level global score that evaluates overall sequence quality and coherence, together with image-unit-level local scores that assess the quality of individual visual units.

2.2 INTERGRPO

InterGRPO treats the generation of a multi-unit interleaved text-image sequence as a structured policy optimization problem. It evaluates each completed trajectory using a trajectory-level global reward R_k^{G} and unit-level local rewards $R_{k,i}^{\mathrm{L}}$. The mean and standard deviation of global reward for the same input are:

$$\mu^G = \frac{1}{K}\sum_{k=1}^{K} R_k^G, \quad \sigma^G = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(R_k^G - \mu^G\right)^2}. \tag{1}$$

Global advantage can be expressed as:

$$A_k^G = \frac{R_k^G - \mu^G}{\sigma^G + \varepsilon}. \tag{2}$$

For each image unit i, we independently normalize the scores across the K rollouts generated from the same input:

$$\mu_i^L = \frac{1}{K}\sum_{k=1}^{K} R_{k,i}^L, \quad \sigma_i^L = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(R_{k,i}^L - \mu_i^L\right)^2}. \tag{3}$$

The local advantage of the i-th image unit in the k-th rollout is defined as:

$$A_{k,i}^L = \frac{R_{k,i}^L - \mu_i^L}{\sigma_i^L + \varepsilon}. \tag{4}$$

The trajectory-level global advantage is broadcast to all image units within the trajectory and combined with each unit's corresponding local advantage through a weighted sum:

$$A_{k,i}^{GL} = \lambda_G A_k^G + \lambda_L A_{k,i}^L, \tag{5}$$

The first term captures the trajectory-level quality relative to the rollout group associated with input, whereas the second term measures the relative quality of the i-th image unit across the corresponding image positions. Their weighted combination provides each image unit with both sequence-level supervision and position-specific credit. In all experiments, we set $\lambda_G = \lambda_L = 0.5, \varepsilon = 10^{-4}$. For the i-th image unit in the k-th rollout and diffusion timestep t, we define the importance-sampling ratio between the current and old policies as

$$\rho_{k,i,t}(\theta) = \frac{\pi_\theta\left(z_{t-1}^{(k,i)} \mid z_t^{(k,i)}, c_{k,i}\right)}{\pi_{\theta_{\mathrm{old}}}\left(z_{t-1}^{(k,i)} \mid z_t^{(k,i)}, c_{k,i}\right)}. \tag{6}$$

where $z_t^{(k,i)}$ denotes the noisy latent of the i-th image unit at timestep t, and $c_{k,i}$ represents its conditioning context. Equivalently, the probability ratio can be computed in log space as

$$\rho_{k,i,t}(\theta) = \exp\left[\log \pi_\theta\left(z_{t-1}^{(k,i)} \mid z_t^{(k,i)}, c_{k,i}\right) - \log \pi_{\theta_{\mathrm{old}}}\left(z_{t-1}^{(k,i)} \mid z_t^{(k,i)}, c_{k,i}\right)\right], \tag{7}$$

Using the global-local advantage $A_{k,i}^{\mathrm{GL}}$, we formulate the clipped surrogate objective as

$$J_{\text{GL-GRPO}}(\theta) = \mathbb{E}_{I,k,i,t}\left[\min\left(\rho_{k,i,t}(\theta)A_{k,i}^{\mathrm{GL}}, \mathrm{clip}\left(\rho_{k,i,t}(\theta), 1-\epsilon_{\mathrm{low}}, 1+\epsilon_{\mathrm{high}}\right)A_{k,i}^{\mathrm{GL}}\right)\right], \tag{8}$$

To prevent the optimized policy from deviating excessively from the reference model, we further introduce a KL-divergence regularization term. Let N_k denote the number of image units in the k-th trajectory. The overall training objective is is then given by:

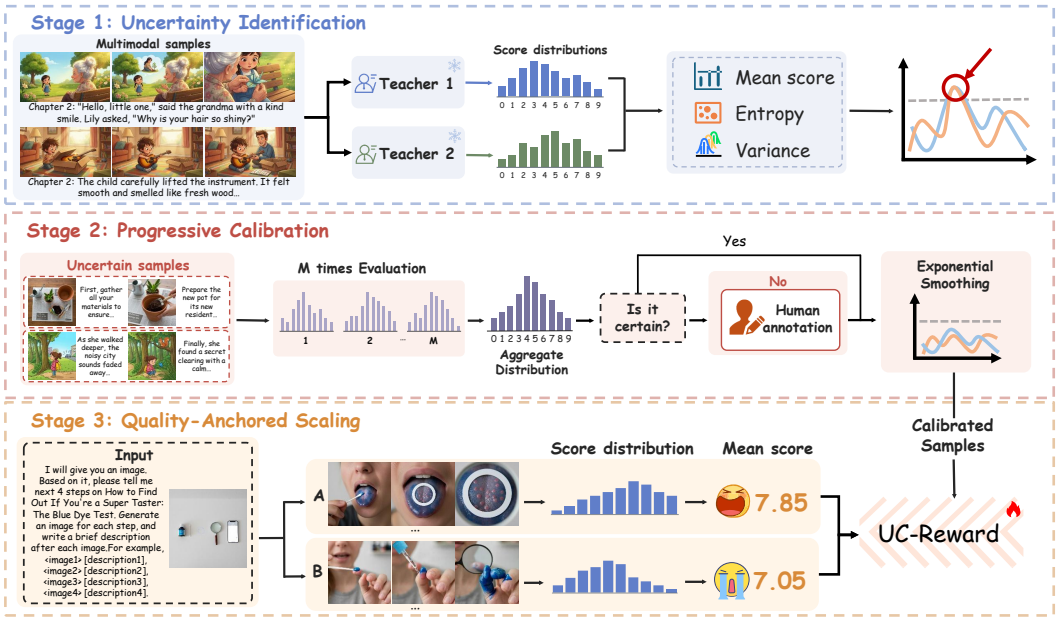


Figure 2: **Overall of UC-Reward.** The student model distills score distributions from multiple teachers. Training data are constructed through three stages: uncertainty identification, progressive calibration, and quality-anchored scaling. In the rightmost plots of Stages 1 and 2, the two curves represent entropy and variance, and the gray vertical dashed line marks the uncertainty threshold.

$$\mathcal{L}_{\text{GL-GRPO}} = \mathbb{E}_{I,k}\left[\frac{1}{N_k}\sum_{i=1}^{N_k}\left(\frac{1}{T}\sum_{t=1}^{T}\max\Big(-\rho_{k,i,t}(\theta)A_{k,i}^{\text{GL}}, -\text{clip}\left(\rho_{k,i,t}(\theta), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}\right)A_{k,i}^{\text{GL}}\Big) + \beta\mathcal{L}_{k,i}^{\text{KL}}\right)\right]. \tag{9}$$

For the textual modality, we adopt the standard GRPO objective. To avoid confounding effects caused by variations in the number and positions of text units across trajectories, we compute the advantage solely from the global reward, which evaluates the overall logical coherence of the generated text.

2.3 UC-Reward

Directly using VLMs evaluation can introduce substantial uncertainty. When the evaluator cannot confidently assess the quality of a sample, its predicted score distribution tends to exhibit high entropy and variance, causing the resulting scores to fluctuate between substantially different values (e.g., 3 and 7). Using such predictions directly as reward supervision may propagate evaluation noise into downstream reinforcement learning. To provide robust and reliable reward signals for downstream reinforcement learning of interleaved generation, we introduce **U**ncertainty-**C**alibrated **Reward** (UC-Reward), a reward modeling framework designed to reduce reward fluctuations arising from evaluation uncertainty and stochastic model predictions. Since evaluators exhibit complementary strengths across different generation tasks, UC-Reward adopts a task-aware multi-teacher distillation strategy. Specifically, for tasks such as visual storytelling and how-to generation, we select teacher models with task-specific evaluation strengths to produce supervisory signals and distill their complementary knowledge into a unified reward model, thereby improving its evaluation capability across diverse interleaved generation tasks. Rather than relying on a single scalar score, it distills the complete score distributions predicted by the teacher models, providing more fine-grained

supervision that improves the reward model's discriminative capability while preserving the continuity of its reward scores. To construct accurate and stable reward supervision, we further develop a three-stage data calibration pipeline.

First, we identify uncertainty in the score distributions produced by the teacher models. Specifically, we quantify sample-level evaluation uncertainty using the entropy and variance of the predicted score distributions, thereby identifying samples associated with ambiguous evaluations. Second, during Progressive Calibration, we perform multiple ($M = 20$ in our experiments) independent evaluations of each uncertain sample and aggregate the resulting predictions to mitigate noise arising from stochastic scoring. Samples that remain uncertain after repeated evaluations are further calibrated through selective human annotation. Finally, during Quality-Anchored Scaling, we incorporate a manually curated high-quality generation for each input into the reward supervision dataset as a reference anchor. Rather than treating this anchor as the unique ground-truth output, we use it to establish a stable quality reference, enabling the reward model to learn a well-calibrated scoring scale and more effectively discriminate among outputs of varying quality.

Together, these stages improve the stability and reliability of the reward supervision while mitigating the risks of reward-scale drift and reward hacking during reinforcement learning. Based on the data obtained after uncertainty identification, progressive calibration, and quality-anchored scaling, we perform offline score-distributional distillation to train the student reward model. Specifically, for each sample x_i, the student model extracts the hidden representation $h_\theta(x_i)$ at the position of the special token <|score|> and predicts a probability distribution over $L = 10$ discrete score bins:

$$p_{\theta,i}(\ell) = \mathrm{softmax}(W h_\theta(x_i) + b)_\ell\,, \qquad \hat{s}_i = \sum_{\ell=0}^{L-1} \ell\, p_{\theta,i}(\ell), \qquad \bar{s}_i = \sum_{\ell=0}^{L-1} \ell\, q_i(\ell), \tag{10}$$

where $p_{\theta,i}(\ell)$ is the student model's predicted probability that sample x_i receives score ℓ, and $q_i(\ell)$ is the corresponding target probability derived from multiple teachers or human annotations. Moreover, $\hat{s}_i$ and $\bar{s}_i$ are the expected scores under the student-predicted and target distributions, respectively. Compared with regressing a single scalar label, distributional supervision preserves the relative confidence, evaluation ambiguity, and fine-grained quality variations encoded in the teacher assessments. The overall training objective of the student reward model is defined as

$$\begin{aligned}\mathcal{L}_{UC-Reward} = & -\frac{1}{N}\sum_{i=1}^{N}\sum_{\ell=0}^{L-1} q_i(\ell)\log p_{\theta,i}(\ell) + \lambda_{\mathrm{corr}}\left[1-\rho\big(\{\hat{s}_i\}_{i=1}^{N},\{\bar{s}_i\}_{i=1}^{N}\big)\right] \\ & + \frac{\lambda_{\mathrm{rank}}}{|\mathcal{P}|}\sum_{(high,low)\in\mathcal{P}} \log[1+\exp(m-(\hat{s}_{high}-\hat{s}_{low}))]\,.\end{aligned} \tag{11}$$

Here, N is the batch size. The first term is a distributional distillation loss that uses soft-label cross-entropy to match the student prediction to the calibrated target distribution. The second term maximizes the Pearson correlation ρ between predicted and target expected scores across the batch, encouraging consistent quality ordering. The third term is a pairwise ranking loss over preference pairs $\mathcal{P}$, where $high$ and low denote the higher- and lower-quality candidates and m is the ranking margin. It encourages the preferred candidate to receive a higher predicted score.

Together, these stages enable the student to capture both the teachers' quality assessments and their score distributions. High-quality reference generations anchor the scoring scale without imposing a unique ground-truth output. The resulting model provides a calibrated, fine-grained reward signal for reinforcement learning.

3 Experiments

We use ILLUME-X Wang et al. (2026a) as the policy model for reinforcement fine-tuning. Its Mixture-of-Transformers (MoT) architecture supports multi-unit interleaved text–image generation through shared self-attention and modality-specific projections and feed-forward networks. The local reward is based on an aesthetic score for each generated image. For all experiments, we configure the generation process with 15 sampling steps and generate images at a resolution of 512×512. The training batch size is set to 6. For each input prompt, we sample 24 candidate outputs to construct the response group used for group-relative advantage estimation. We optimize the trainable

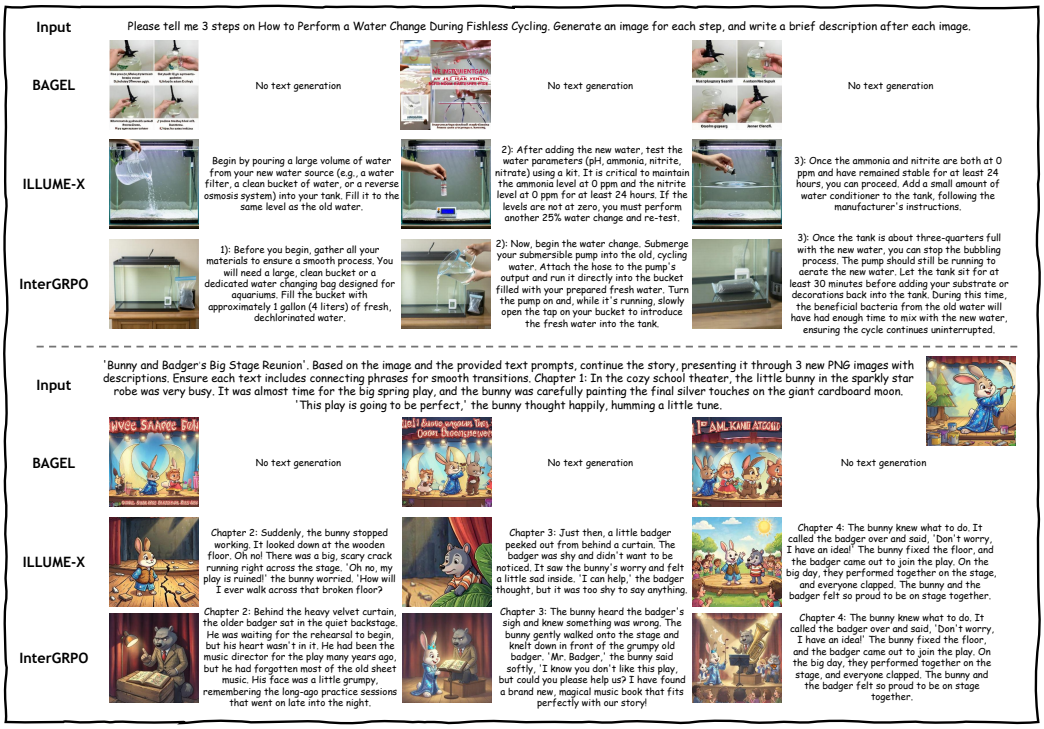


Figure 3: **Visual comparison of different model results on SIL Bench.** For BAGEL, we generate the response through multi-turn interaction.

parameters with a learning rate of 3×10^{-4}. During training, the classifier-free guidance (CFG) scale is fixed at 1.0, allowing the policy to explore diverse generations without introducing additional guidance amplification. Unless otherwise specified, these hyperparameters are kept consistent across all experiments to ensure a fair comparison between different methods and configurations.

3.1 Inteleaved Text-Image Generation

Our task allows users to specify the number and arrangement of interleaved generation. Therefore, we adopt ISG-Bench Chen et al. (2024) as one of our evaluation benchmarks. However, its reference images have limited visual quality, we so construct a **S**tructured **I**nter**L**eaved multimodal generation **Bench**mark (SIL-Bench) with high-quality visual content. SIL-Bench covers How-to and Visual Storytelling generation. These tasks inherently demand temporal reasoning and strong cross-modal complementarity, making them representative scenarios in which interleaved text-image generation is expected to offer substantial benefits. Following ILScore Wang et al. (2026a), SIL-Bench evaluate model performance along four dimensions: image-text accuracy, single image accuracy, image-sequence accuracy, and text accuracy.

SIL-Bench. We evaluate InterGRPO on SIL-Bench alongside two open-source baselines BAGEL Deng et al. (2025), ILLUME-X Wang et al. (2026a), and commercial model Gemini 3.0 Pro Image Preview Comanici et al. (2025). To reduce variability from individual judgments, we conduct multiple evaluation runs using Gemini 3.1 Pro Preview as the judge. Table 1 presents the results. Gemini 3.0 Pro Image Preview leads on most dimensions, while InterGRPO ranks second overall and outperforms both open-source baselines. Compared with ILLUME-X, InterGRPO improves the overall score by 0.14 and Image Sequence Accuracy by 0.23. On the Visual Storytelling task, InterGRPO also surpasses Gemini 3.0 pro Image Preview in image sequence accuracy. We attribute this result to the emphasis on image sequence consistency in our reward model.

ISG-Bench. For a fair comparison with the baselines, we follow the original ISG-Bench evaluation protocol and use ChatGPT-4o as the judge. Table 2 compares InterGRPO with several publicly

Table 1: **Quantitative evaluation of various models on SIL-Bench.** Bold and underlined values indicate the best and second-best overall performance in each category, respectively.

Dimension	Model	How-to	Visual Storytelling	Avg
Image-Text Accuracy	Gemini 3.0 Pro Image Preview Comanici et al. (2025)	**6.58**	**6.40**	**6.49**
	BAGEL Deng et al. (2025)	3.42	3.33	3.38
	ILLUME-X Wang et al. (2026a)	5.32	5.94	5.63
	InterGRPO	<u>5.19</u>	<u>6.03</u>	<u>5.61</u>
Single Image Accuracy	Gemini 3.0 Pro Image Preview	**6.44**	**6.82**	**6.63**
	BAGEL	3.73	4.43	4.08
	ILLUME-X	<u>5.16</u>	<u>6.08</u>	<u>5.62</u>
	InterGRPO	<u>5.16</u>	6.04	5.60
Image Sequence Accuracy	Gemini 3.0 Pro Image Preview	**6.12**	<u>5.63</u>	**5.88**
	BAGEL	3.85	4.59	4.22
	ILLUME-X	5.00	5.59	5.30
	InterGRPO	<u>5.63</u>	**5.71**	<u>5.67</u>
Text Accuracy	Gemini 3.0 Pro Image Preview	**6.76**	**6.90**	**6.83**
	BAGEL	3.46	3.00	3.23
	ILLUME-X	<u>6.00</u>	6.19	6.10
	InterGRPO	5.92	<u>6.74</u>	<u>6.33</u>
Overall	Gemini 3.0 Pro Image Preview	**6.47**	**6.44**	**6.46**
	BAGEL	3.61	3.84	3.73
	ILLUME-X	5.37	5.95	5.66
	InterGRPO	<u>5.47</u>	<u>6.13</u>	<u>5.80</u>

available models on ISG-Bench. InterGRPO attains the best overall performance among the evaluated models. Relative to ILLUME-X, it improves performance across all task categories except Image Decomposition and raises the overall score by 0.28. Notably, our training data cover only How-to and Visual Storytelling generation, with no examples from the other ISG-Bench task categories. The improvements on those unseen categories suggest that InterGRPO generalizes beyond its training tasks. We further observe that Gemini 3.0 Pro Image Preview receives a relatively low overall score, largely because it does not support the 3D-related tasks in the benchmark and often produces responses that do not address the VQA questions.

Table 2: **Quantitive comparison across different tasks on ISG-Bench Chen et al. (2024).** Bold and underlined values indicate the best and second-best overall performance in each category, respectively.

Model	Is unified model?	Style Transfer	Progressive	3D_Scene	Image Decomposition	Image-Text Complementation	Temporal Prediction	Visual StoryTelling	VQA	**AVG**
Show-o Xie et al. (2024)	✓	2.11	2.41	1.43	2.87	2.06	2.58	3.32	1.86	2.33
Anole Chern et al. (2024)	✓	2.93	2.76	1.85	1.49	3.21	2.58	2.97	4.70	2.81
Minigpt-5 Zheng et al. (2023)	✓	2.15	3.15	1.79	2.54	2.72	2.73	2.91	4.29	2.79
CoMM-Minigpt-5 Chen et al. (2025b)	✓	2.96	2.60	3.09	2.24	3.09	2.52	2.72	2.87	2.96
Seed-Llama-14b Ge et al. (2023)	✓	1.84	3.30	1.52	3.69	1.94	1.78	2.84	2.20	2.39
Gemini Team et al. (2024) & SD3 Esser et al. (2024)	✗	4.89	**6.59**	2.68	<u>7.26</u>	6.37	5.26	5.68	**7.89**	5.83
ISG-AGENT Chen et al. (2024)	✗	5.87	<u>6.46</u>	<u>4.89</u>	**7.58**	6.93	4.54	7.03	6.80	<u>6.26</u>
BAGEL Deng et al. (2025)	✓	3.90	3.15	3.59	3.10	4.19	3.24	4.07	5.88	3.89
ILLUME-X Wang et al. (2026a)	✓	<u>6.02</u>	5.75	3.96	6.97	<u>7.06</u>	<u>6.43</u>	<u>7.25</u>	6.67	<u>6.26</u>
Gemini 3.0 Pro Image Preview Comanici et al. (2025)	✓	4.29	5.20	2.54	4.54	5.87	5.09	5.86	2.12	4.44
InterGRPO (Ours)	✓	**6.47**	5.77	**5.58**	6.32	**7.18**	**6.51**	**7.46**	<u>7.05</u>	**6.54**

3.2 Ablation Study

As shown in Figure 4, we conduct ablation studies on the reward model and credit assignment strategy. When Qwen3-VL is used as the reward model, global-local credit assignment yields faster initial improvement than using the global reward alone. However, both configurations degrade substantially as training progresses, eventually falling below their initial reward levels. Replacing Qwen3-VL with UC-Reward leads to a more stable improvement even when only the global reward is used. Combining UC-Reward with global-local credit assignment achieves the best overall reward trajectory, demonstrating the complementary benefits of reliable reward supervision and fine-grained credit assignment. Figure 5 reports the proportion of reward scores that fall within different error tolerances of human ratings. UC-Reward, with 8B parameters, substantially outperforms Qwen3-VL 8B model across all tolerances. It also surpasses the Qwen-3.6 27B model and achieves agreement comparable to that of the Qwen-3.5 122B model.

4 Related Work

Interleaved Text-Image Generation. Recent advances in image synthesis have produced models capable of generating high-fidelity images from textual or multimodal inputs Wu et al. (2025c);

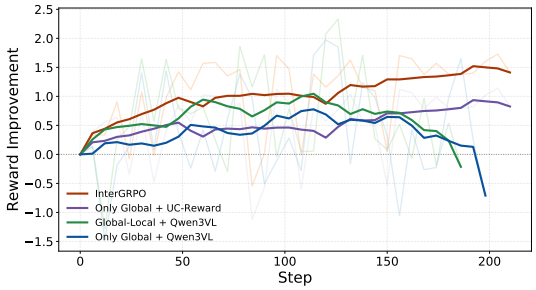


Figure 4: **Training reward curves for ablations of global–local credit assignment and UC-Reward.**

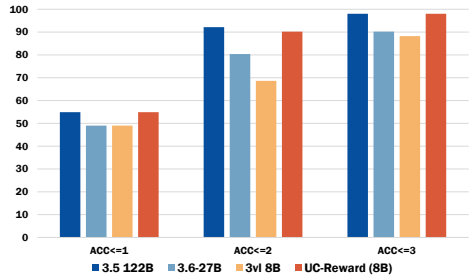


Figure 5: **Agreement between reward model scores and human ratings across error tolerances.** All compared models belong to the Qwen family.

Podell et al. (2024); Esser et al. (2024); Wu et al. (2025a); Wang et al. (2024); Cai et al. (2025); Zhao et al. (2026a). However, these models primarily produce visual outputs. Unified multimodal models Chen et al. (2025c); Team (2025); Chern et al. (2024); Chen et al. (2025a); Feng et al. (2026) can generate both text and images within a shared architecture, but many still assume a predefined output modality for each request Wu et al. (2025e); Chen et al. (2026b); Wu et al. (2025b). Several studies have explored interleaved text-image generation Deng et al. (2025); Xie et al. (2024); Liao et al. (2025); Cui et al. (2025), establishing the feasibility of producing alternating text and image units. Free-form N-to-M generation extends this setting by requiring models to follow user-specified output structures while maintaining coherence across units. Wang et al. (2026a); Ye et al. (2026) supports such customizable outputs, but visual inconsistency and error accumulation across successive units remain challenges Zhou et al. (2025).

Reinforcement Learning for Image Generation. Reinforcement learning has been widely used to align image generation models with human preferences. Recent studies have explored preference-based optimization of generative models using pairwise comparisons Wallace et al. (2024); Croitoru et al. (2025). Subsequent advances in online reward-guided policy optimization and reward modeling Liu et al. (2026b); Xue et al. (2025); Li et al. (2025); He et al. (2026); Zhou et al. (2026); Wu et al. (2025d); Xu et al. (2026); Wang & Yu (2025); Wang et al. (2026b); Wei et al. (2025) have yielded substantial improvements in prompt adherence, text rendering accuracy, and perceptual quality across image generation, video synthesis, and image editing. Liu et al. (2026c) combines GRPO with Flow-GRPO to jointly optimize textual reasoning and image synthesis using rewards derived from the final generated image. However, its evaluation is limited to a single atomic text-image unit. Practical applications often require multi-unit outputs that maintain text-image alignment and preserve consistent character identities, styles, and content throughout the sequence. In this setting, assigning the same trajectory-level advantage to all images obscures their individual quality differences, leading to inaccurate credit assignment across visual units.

5 Conclusion

In this work, we introduce InterGRPO, an online reinforcement fine-tuning framework for multi-unit interleaved text-image generation. Its global-local credit assignment optimizes individual visual units while accounting for sequence-level coherence. We further introduce UC-Reward, which calibrates uncertainty in the score distributions produced by vision-language models, providing more reliable reward feedback for interleaved text–image generation. We also construct SIL-Bench to evaluate interleaved generation under user-specified output structures and high-quality visual references. Experiments on SIL-Bench and ISG-Bench show that InterGRPO outperforms the evaluated baselines, while ablations confirm the contributions of global–local credit assignment and UC-Reward. Together, these results demonstrate the effectiveness of online preference optimization for multi-unit interleaved generation. Collectively, our work establishes a foundation for reliable preference optimization of long-form interleaved multimodal generation.

AI USE STATEMENT

In this work, we used generative AI tools (MLLM) for synthesizing training/evaluation data, generating and validating experimental results, drafting and editing portions of the manuscript text, and polishing the language of the paper, and the remaining required-disclosure tasks are not applicable to this work. All AI-assisted work has been reviewed: the AI-synthesized data and AI-generated experimental results were manually inspected and cross-checked against our experimental pipeline to verify their correctness and consistency, and the AI-assisted text was carefully read and revised by the authors to ensure accuracy, clarity, and alignment with the actual contributions of this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Dongping Chen, Ruoxi Chen, Shu Pu, Zhaoyi Liu, Yanru Wu, Caixi Chen, Benlin Liu, Yue Huang, Yao Wan, Pan Zhou, et al. Interleaved scene graphs for interleaved text-and-image generation assessment. *arXiv preprint arXiv:2411.17188*, 2024.

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

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 *CVPR*, 2025b.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling. *arXiv 2501.17811*, 2025c.

Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. *Advances in neural information processing systems*, 38:69110–69139, 2026a.

Zisheng Chen, Chunwei Wang, Runhui Huang, Hongbin Xu, Xiuwei Chen, Jun Zhou, Jianhua Han, Hang Xu, and Xiaodan Liang. Semhitok: A unified image tokenizer via semantic-guided hierarchical codebook for multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2026, pp. 44597–44620, 2026b.

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.

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

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, Nicu Sebe, and Mubarak Shah. Curriculum direct preference optimization for diffusion and consistency models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2824–2834. IEEE, 2025.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, Yueze Wang, Chengyuan Wang, Fan Zhang, Yingli Zhao, Ting Pan, Xianduo Li, Zecheng Hao, Wenxuan Ma, Zhuo Chen, Yulong Ao, Tiejun Huang, Zhongyuan Wang, and Xinlong Wang. Emu3.5: Native Multimodal Models are World Learners. *arXiv 2510.26583*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging Properties in Unified Multimodal Pretraining. *arXiv 2505.14683*, 2025.

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

Kailai Feng, Yuxiang Wei, Bo Chen, Yang Pan, Songwei Liu, Chenqian Yan, Yuan Gao, and Wangmeng Zuo. Dreamlite: A lightweight on-device unified model for image generation and editing. In *European Conference on Computer Vision*, pp. 214–232. Springer, 2026.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023.

Xiaoxuan He, Siming Fu, Yuke Zhao, Wanli Li, Jian Yang, Dacheng Yin, Fengyun Rao, and Bo Zhang. Tempflow-grpo: When timing matters for grpo in flow models. In *International Conference on Learning Representations*, volume 2026, pp. 132869–132895, 2026.

Wenxuan Huang, Shuang Chen, Zheyong Xie, Shaosheng Cao, Shixiang Tang, Yufan Shen, Qingyu Yin, Wenbo Hu, Xiaoman Wang, Yuntian Tang, et al. Interleaving reasoning for better text-to-image generation. In *International Conference on Learning Representations*, volume 2026, pp. 106153–106182, 2026.

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*, 2025.

Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An Omni Foundation Model for Interleaved Multi-Modal Generation. *arXiv 2505.05472*, 2025.

Henglin Liu, Huijuan Huang, Jing Wang, Chang Liu, Xiu Li, and Xiangyang Ji. Diversegrpo: Mitigating mode collapse in image generation via diversity-aware grpo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1864–1873, 2026a.

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, 2026b.

Jie Liu, Zilyu Ye, Linxiao Yuan, Shenhan Zhu, Yu Gao, Jie Wu, Kunchang Li, Xionghui Wang, Xiaonan Nie, Weilin Huang, et al. Unigrpo: Unified policy optimization for reasoning-driven visual generation. *arXiv preprint arXiv:2603.23500*, 2026c.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

Chameleon Team. Chameleon: Mixed-Modal Early-Fusion Foundation Models. *arXiv 2405.09818*, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: A family of highly capable multimodal models. arxiv 2023. *arXiv preprint arXiv:2312.11805*, 2024.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 2024.

Chonghuinan Wang, Zhikai Chen, Chunwei Wang, Yecong Wan, Junwei Yang, Zhixin Wang, Wei Zhang, Jiaqi Xu, Renjing Pei, Xiaohe Wu, et al. Illuminating unified multimodal model for free-form interleaved text-image generation. *arXiv preprint arXiv:2606.30054*, 2026a.

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

Jing Wang, Jiajun Liang, Jie Liu, Henglin Liu, Gongye Liu, Jun Zheng, Wanyuan Pang, Ao Ma, Zhenyu Xie, Xintao Wang, et al. Grpo-guard: Mitigating implicit over-optimization in flow matching via regulated clipping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5988–5998, 2026b.

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

Hongyang Wei, Baixin Xu, Hongbo Liu, Size Wu, Jie Liu, Yi Peng, Peiyu Wang, Zexiang Liu, Jingwen He, Yidan Xietian, et al. Skywork unipic 2.0: Building kontext model with online rl for unified multimodal model. *arXiv preprint arXiv:2509.04548*, 2025.

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

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

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. OmniGen2: Exploration to Advanced Multimodal Generation. *arXiv 2506.18871*, 2025c.

Jie Wu, Yu Gao, Zilyu Ye, Ming Li, Liang Li, Hanzhong Guo, Jie Liu, Zeyue Xue, Xiaoxia Hou, Wei Liu, et al. Rewarddance: Reward scaling in visual generation. *arXiv preprint arXiv:2509.08826*, 2025d.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17739–17750. IEEE, 2025e.

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

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.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Mingcheng Ye, Jiaming Liu, and Yiren Song. Loom: Diffusion-transformer for interleaved generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4582–4592, 2026.

Lei Zhang, Junjiao Tian, Zhipeng Fan, Kunpeng Li, Jialiang Wang, Weifeng Chen, Markos Georgopoulos, Felix Juefei-Xu, Manling Li, Julian McAuley, et al. Think in strokes, not pixels: Process-driven image generation via interleaved reasoning. In *European Conference on Computer Vision*, pp. 390–407. Springer, 2026.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, et al. Qwen-image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026a.

Yida Zhao, Kuan Li, Xixi Wu, Liwen Zhang, Ding-Chu Zhang, Baixuan Li, Maojia Song, Zhuo Chen, Chenxi Wang, Xinyu Wang, et al. Repurposing synthetic data for fine-grained search agent supervision. In *International Conference on Learning Representations*, volume 2026, pp. 14467–14492, 2026b.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*, 2023.

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

Yujie Zhou, Pengyang Ling, Jiazi Bu, Yibin Wang, Yuhang Zang, Jiaqi Wang, Li Niu, and Guangtao Zhai. Fine-grained grpo for precise preference alignment in flow models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20045–20054, 2026.

A Appendix

In this supplementary material, we provide additional explanation and experimental results to further support the main paper. The content is organized as follows: Sec. A.1 provides a more detailed experimental analysis; Sec. A.2 provides the detailed prompt templates used for scoring with UC-Reward; Sec. A.3 offers additional experimental result visualizations; Sec. A.4 presents the prompt templates used to construct SIL-Bench and provides some cases of SIL-Bench.

A.1 More Experimental Analysis

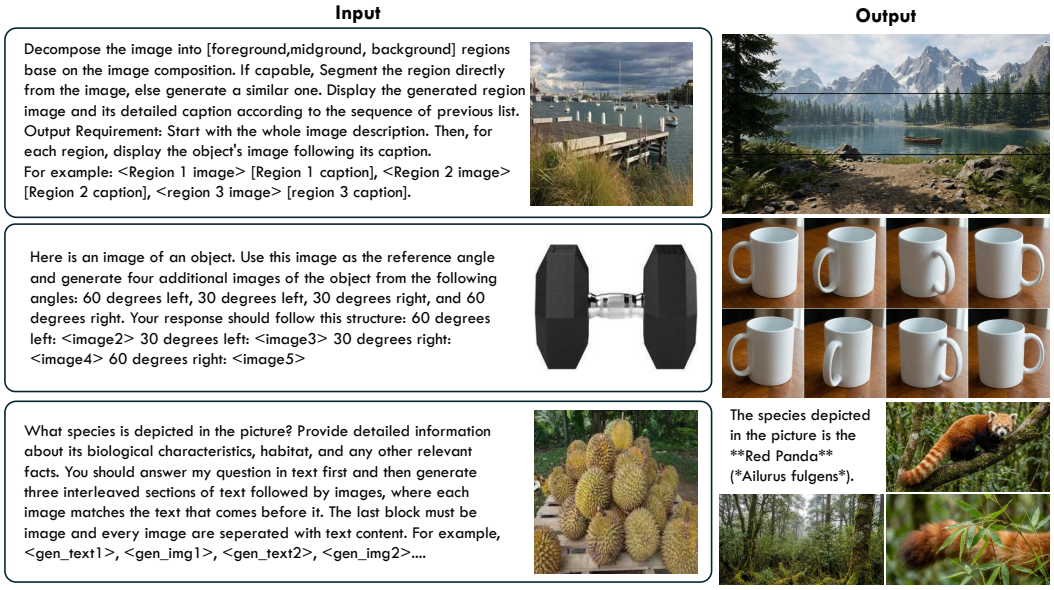


Figure 6: **Failure cases of Gemini 3.0 Pro Image Preview on ISG-Bench**

As shown in Table 2, Gemini 3.0 Pro Image Preview performs unexpectedly poorly, particularly on Image Decomposition, 3D Scene, and VQA. Inspection of its outputs reveals cases in which the generated content is unrelated to the input instruction or image; representative failures are shown in Figure 6. On Image Decomposition, the model draws separation lines over the original image instead of generating an image of the target object with its components separated, as required by ISG-Bench. When asked to generate multiple images, it also tends to combine the requested content into a single image. On VQA, it frequently produces incorrect or irrelevant answers. We hypothesize

that these failures may be partly associated with the limited image quality of ISG-Bench, which is one of the issues SIL-Bench was designed to address.

A.2 UC-REWARD SCORING PROMPTS

Figure 7 presents the prompt templates used to obtain quality scores with UC-Reward, providing the details needed to reproduce the scoring procedure. Figure 8 presents the prompt specifying the scoring rubric used by UC-Reward.

A.3 MORE RESULTS

Figure 9 to 18 provides additional visualizations of the generated results, illustrating the performance of InterGRPO across interleaved text-image generation tasks.

A.4 MORE OF SIL-BENCH

Figure 19 and 20 presents the prompt templates used to construct SIL-Bench, along with figure 21 and 22 representative benchmark examples.

You are an impartial evaluator for multimodal interleaved generation.
You are given:
(1) a multimodal query consisting of text and one or more images;
(2) a multimodal answer consisting of interleaved text and images.
Evaluate the overall quality of the answer according to the following criteria:
1. Instruction Following
- Whether the answer correctly understands and fulfills the user's request.
- Whether all requested content is provided.
2. Text-Image Consistency
- Whether each image faithfully supports the surrounding text.
- Whether the text accurately describes the images.
- Whether text and images complement each other to form a coherent answer.
3. Inter-image Consistency
- Whether multiple generated images are mutually consistent.
- Maintain consistent subjects, identities, attributes, style, color, viewpoint and scene when appropriate.
- Images should not contradict each other.
4. Organization and Logicality
- Whether the answer is well organized.
- Whether the ordering of text and images is natural.
- Whether the reasoning or narration is coherent and easy to follow.
5. Image Quality
- Whether images are clear, natural and visually pleasing.
- Penalize blurry, distorted, low-quality or artifact-heavy images.
- Consider composition, rendering quality and overall aesthetics.
Judge the answer based on the overall quality rather than any single criterion.

Figure 7: **Scoring Prompts for UC-Reward.**

Select one integer score from 0 to 9 according to the overall quality.
Your final answer after reasoning must be a single integer from 0 to 9.
Only output the number.

Scoring guideline:
9 - Perfect: The output fully satisfies the request. All images accurately match their corresponding text and collectively provide a coherent and complete visual response. Every generated image correctly represents the described content, with strong consistency in identity, attributes, scene, style, and details across the entire sequence. The narrative or process is logically organized. Images are clear and visually high quality without noticeable artifacts.
7-8 - Excellent: The output strongly satisfies the request with only minor issues. Most or all images correctly correspond to the text and support the intended meaning, with only small mismatches or missing details. The image sequence maintains good consistency in identity, scene, and style. Any errors are limited and do not significantly affect understanding. Image quality is generally high.
5-6 - Acceptable: The output satisfies the main requirements but has noticeable limitations. Most images are relevant and correctly aligned with the text, but some images may contain incorrect details, missing key elements, or weak visual grounding. The image sequence is generally coherent but may have inconsistencies in identity, scene, or style. Image quality may be moderate, but visual quality alone should not compensate for semantic errors.
3-4 - Below Standard: The output has significant problems. Several images do not correctly match the corresponding text, or important visual information is missing. The connection between images and text is weak, and the generated sequence may contain noticeable inconsistencies in identity, scene, style, or logical progression. Even if some images are visually appealing, incorrect or irrelevant images substantially reduce the score.
1-2 - Poor: The output fails in most aspects. Most images are inconsistent with the text or fail to represent the intended content. The image sequence lacks continuity, with major identity, scene, or style changes. The output only partially satisfies the request and contains severe semantic or visual problems.
0 - Terrible: The output completely fails. Images are unrelated to the text or user request, and the generated content is unusable. There is little to no meaningful image-text alignment or cross-image consistency.

Important Rules:
- Semantic correctness is more important than visual quality. A visually high-quality image that does not match the text should receive a low score.
- If some images are correct but others are incorrect, judge based on the proportion and importance of incorrect images. A few minor errors may still receive a high score, but multiple incorrect images should significantly reduce the score.
- If most images correctly match the text but contain small inconsistencies, the score should remain moderate to high.
- If severe artifacts, obvious blur, strong distortion, or image corruption significantly affect usability, the score should not exceed 6.
- If there are major image-text mismatches or severe inconsistencies across images (e.g., identity or scene changes), the output should not receive a high score even if individual images are visually appealing.
- Judge the overall quality of the output rather than allowing a single minor flaw to dominate the score.

Figure 8: **Scoring Rubric Prompts for UC-Reward.**



Figure 9: **Results of How-to on SIL-Bench**

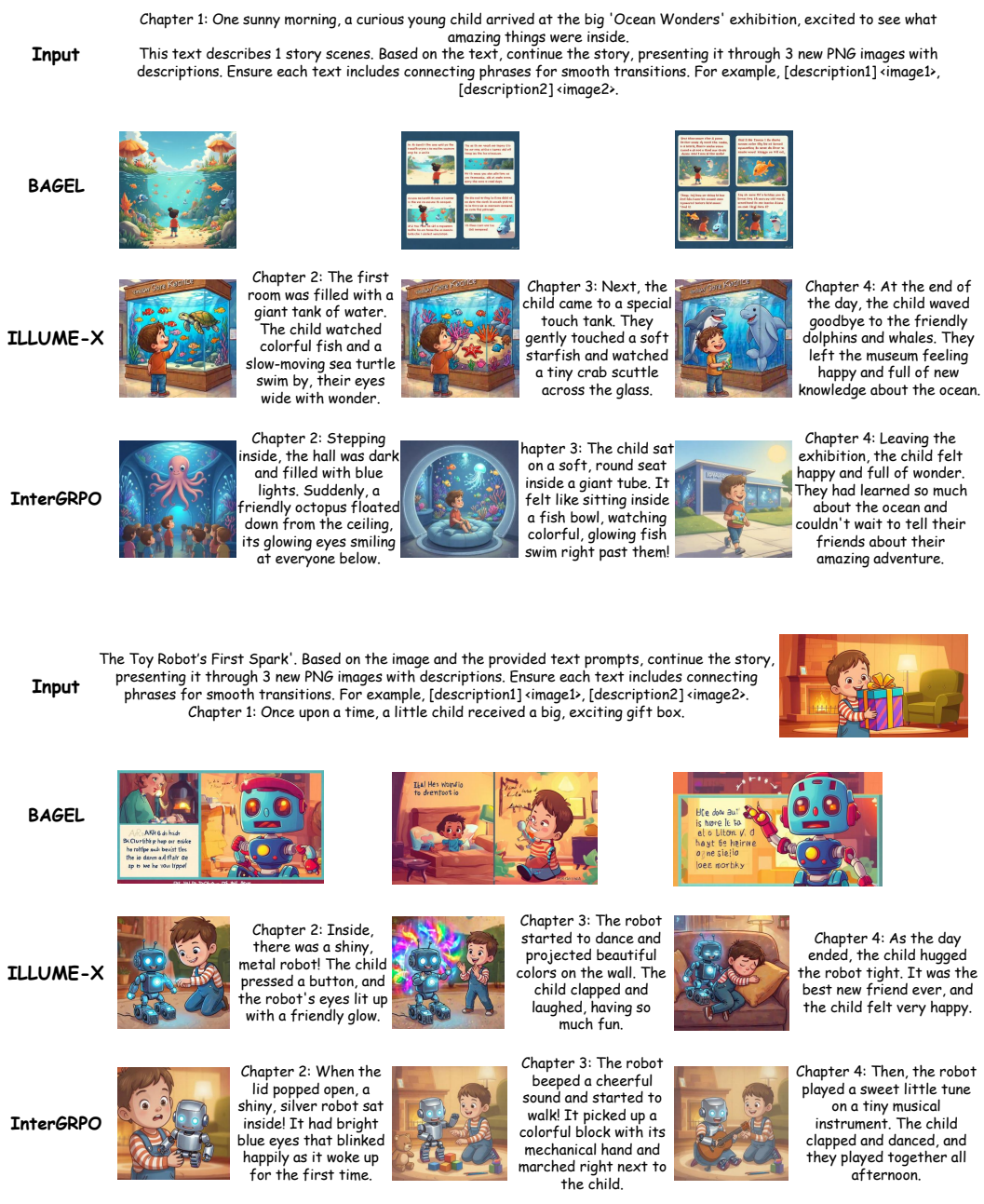


Figure 10: **Results of Visual Storytelling on SIL-Bench**

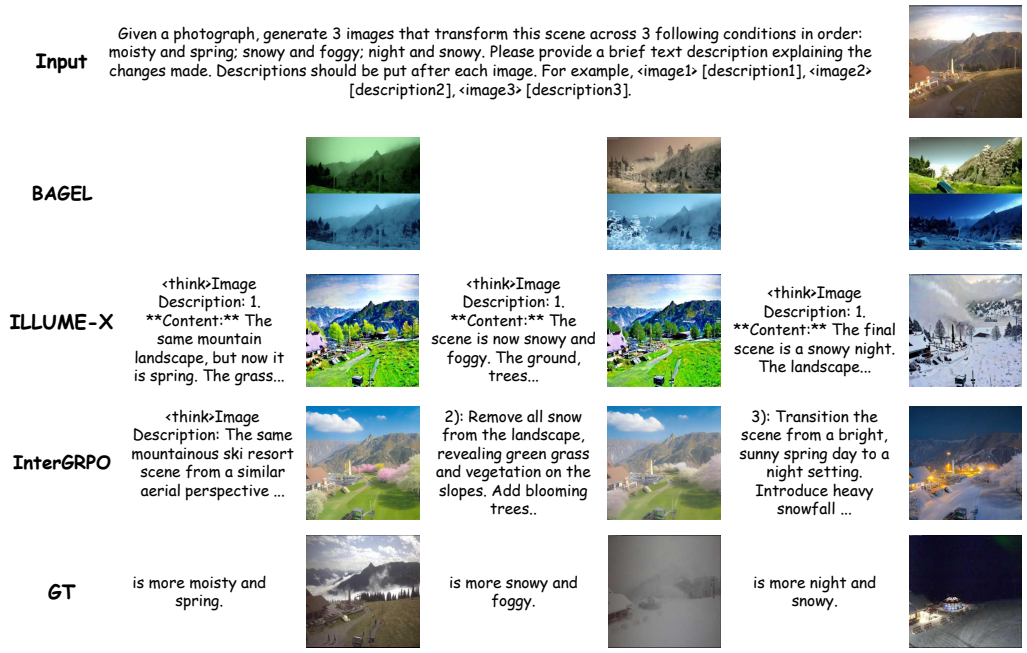


Figure 11: **Results of Style Transfer on ISG-Bench**

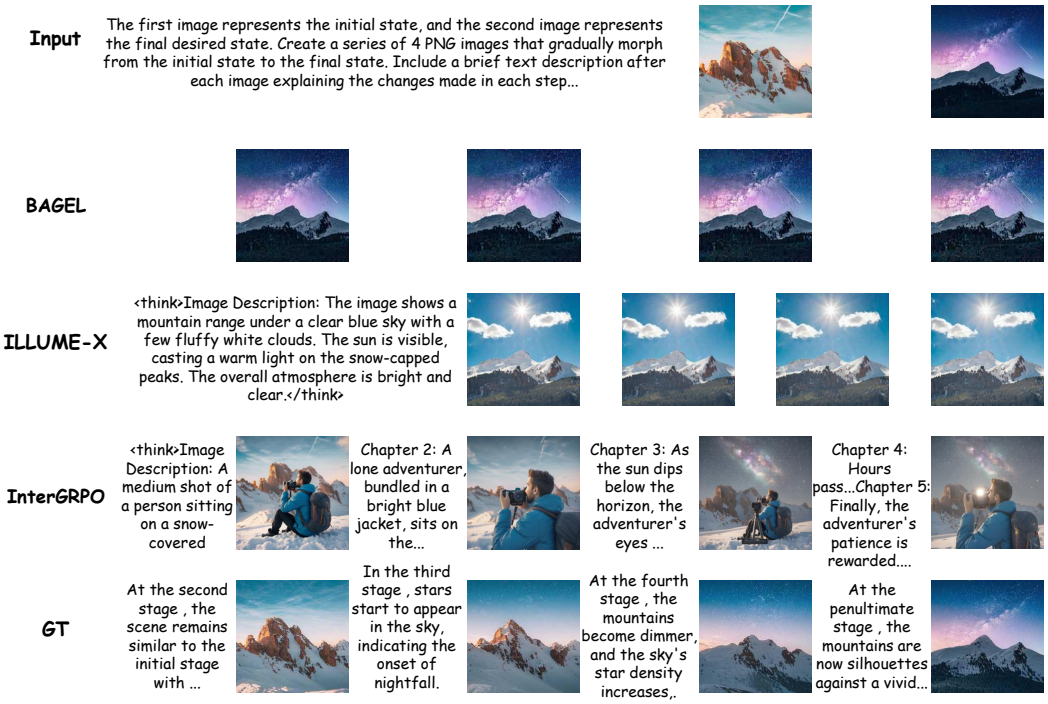


Figure 12: **Results of Progressive on ISG-Bench**

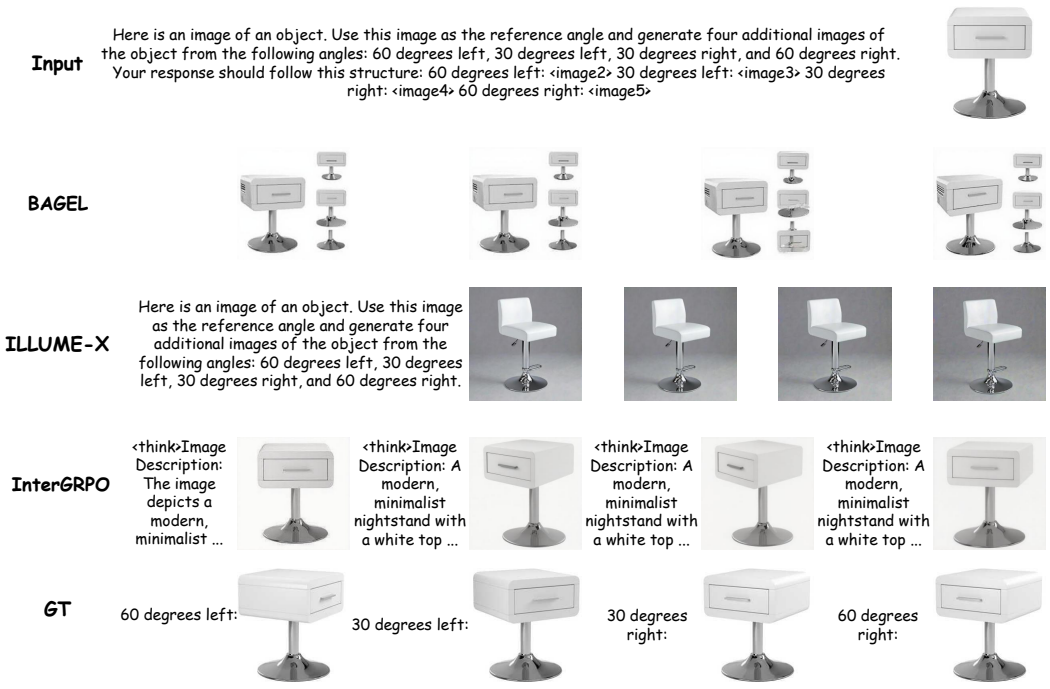


Figure 13: **Results of 3D Scene on ISG-Bench**



Figure 14: **Results of Image Decomposition on ISG-Bench**



Figure 15: **Results of Image-Text Complementation on ISG-Bench**

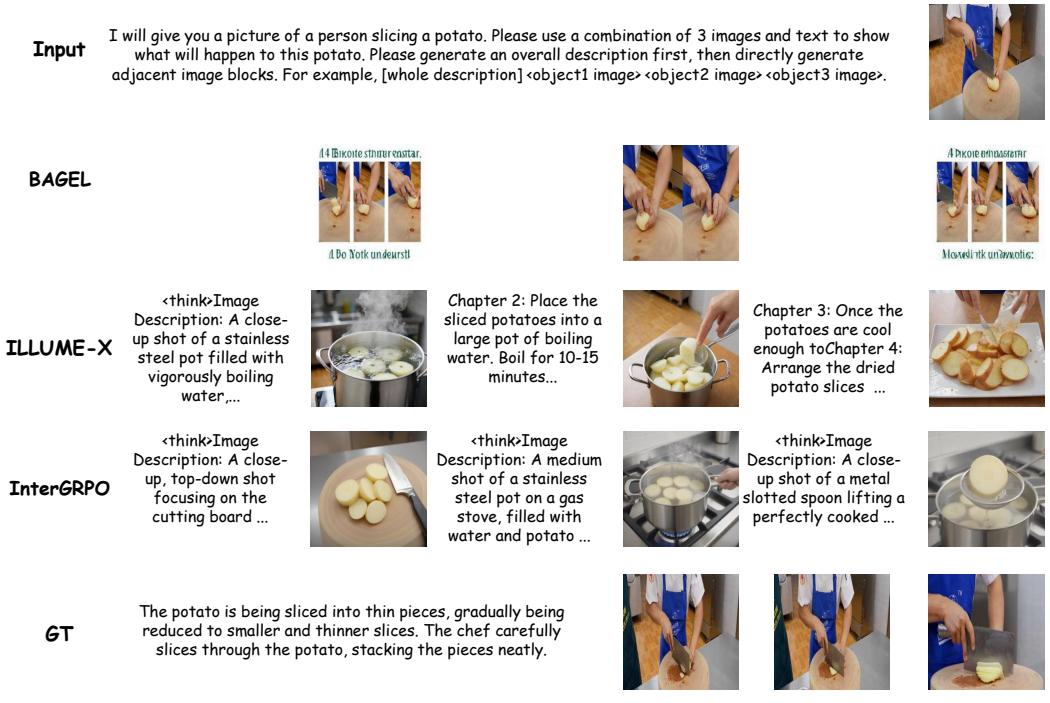


Figure 16: **Results of Temporal Prediction on ISG-Bench**

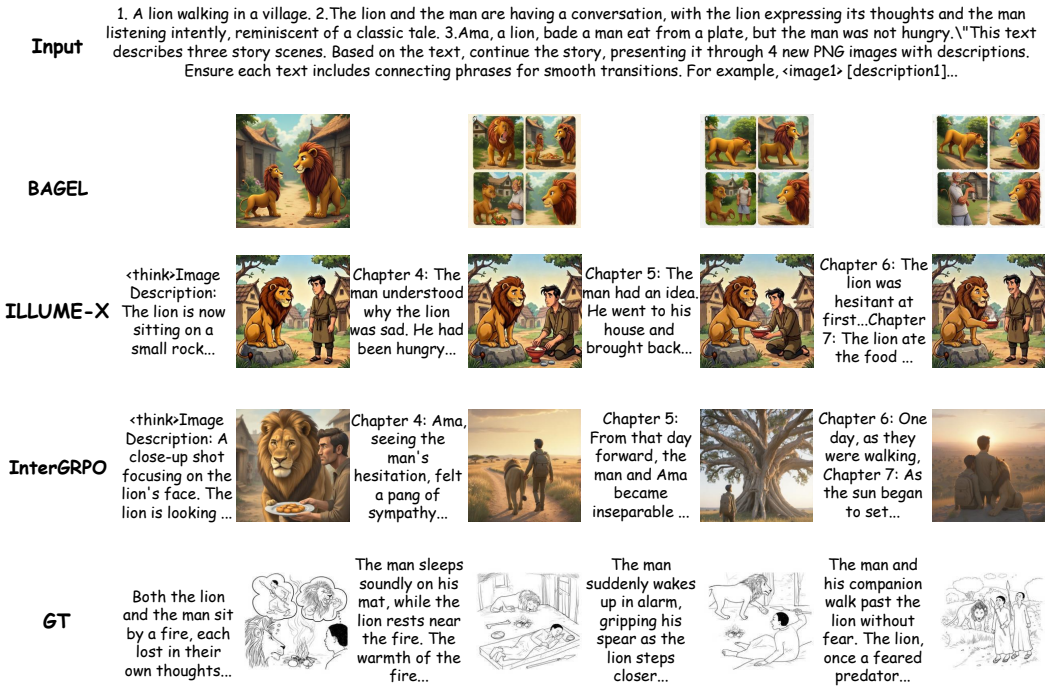


Figure 17: **Results of Visual Story Telling on ISG-Bench**

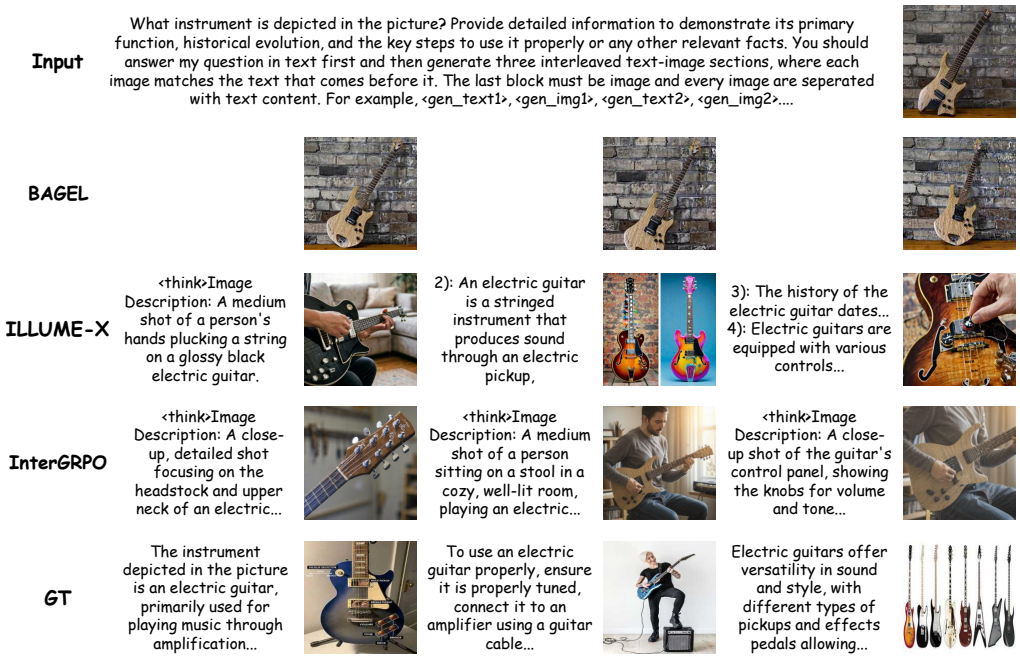


Figure 18: **Results of VQA on ISG-Bench**

```
SYSTEM_PROMPT = """You are an expert tutorial writer and visual director.
Create a practical, accurate tutorial with exactly {step_num} sequential steps.
Keep the instructions safe, clear, easy to follow, and appropriate for minors.
Exclude violent, sexual, or otherwise age-inappropriate content."""

USER_PROMPT = """Create an illustrated how-to tutorial from these topic tags:
{topic_tags}

First infer one specific, natural tutorial topic from all the tags. Then plan
exactly {step_num} sequential steps. Each step must describe one meaningful
action and must be suitable for a corresponding illustration.

Safety requirements:
- All tutorial text and visual descriptions must be appropriate for minors.
  Do not include violence, weapons, sexual or suggestive content, nudity,
  graphic injury, substance abuse, or other age-inappropriate material.
- Do not provide dangerous, illegal, invasive-medical, or otherwise harmful
  instructions. For risky topics, give only cautious, non-invasive guidance
  and recommend a qualified professional where appropriate.
- Do not invent facts, tools, or materials.
- Do not put written labels in an illustration unless essential.
- Describe each image as a plausible real-world scene suitable for a
  photorealistic image. Use an illustrated style only rarely (about 30% of
  tutorials), when it is especially helpful for clarity or the topic.

Respond in English with only valid JSON in this exact shape:
{{
  "global-cot": "A concise tutorial plan, required materials, and safety notes.",
  "image-cot": [
    {{"image-1": "Detailed visual description for step 1"}}
  ],
  "texts": [
    {{"text-1": "Step 1 instruction"}}
  ],
  "title": "How to ..."
}}

The image-cot and texts arrays must each contain exactly {step_num} entries,
numbered consecutively from 1 to {step_num}."""
```

Figure 19: **Prompt for Generating How-to Samples for SIL-Bench**

You are good at telling stories including interleaved images and texts. Now you need to tell a story about the topic of “TOBEREPLACE”. Ensure the story is well-organized into appropriate chapters, with each chapter containing a reasonable amount of plot and avoiding excessive events. The plot is simple and healthy, without negative guidance, suitable for children aged 3-12.

Please generate your chain-of-thought (CoT) to finish the above storytelling task and the story texts following the rules:

Content

global chain-of-thought (**global-cot**)

1. Think about and describe the overall idea and specific steps for writing this story.
2. Think about and describe the main characters, plot, and typical environment that need to appear in this story.Use demonstrative pronoun or the description like "the boy/the princess/the old man" instead of naming the characters
3. Think about and describe the issues that need attention regarding the overall text description and illustrations for this story.

image-level chain-of-thought (**image-cot**)

1. Think and describe the content of each illustration as much detail as possible, for example: the style of the image and its visual elements. The image style is mainly cartoon.
2. Think and describe the differences between the current illustration and the previous one, for example: new characters introduced, new scenes transitioned to, changed overall atmosphere, plot developments.
3. Think and describe the impact of the current illustration on the visual content of subsequent illustrations, for example: changes in the surrounding environment caused by actions, character consistency.

story texts

1. Generate the texts in sequence to create the story.
2. Keep the texts logically coherent and smooth.
3. Ensure each text is corresponding to an story image.

Title (**Title**)

Think and describe the title of this story.

Output Format

1. Use English to reponse
2. The story should contain **[ST_NUM]** image-text pairs
3. Do not render text in image if unnecessary
4. Your response should be structure in the next JSON format:

```
{
   "global-cot": Your Output,
   "image-cot": [
      {
         "image-1": Your Output
      },
      {
         "image-2": Your Output
      }, ...
   ]
   "texts": [
      {
         "text-1": Your Output
      },
      {
         "text-2": Your Output
      }, ...
   ],
   "title": Your Output
}
```

Figure 20: **Prompt for Generating Visual Storytelling Samples for SIL-Bench**

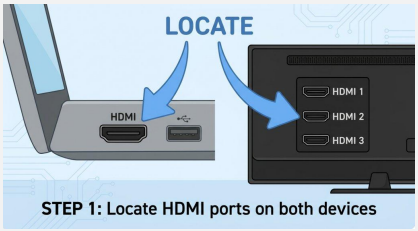
I will give you an image. Based on it, please tell me next 3 steps on How to Wash and Re-waterproof Your Ski Pants. Generate an image for each step, and write a brief description after each image.For example, <image1> [description1], <image2> [description2], <image3> [description3].



1): First, prepare your ski pants by emptying all pockets and brushing off any loose dirt. Then, close all zippers, Velcro straps, and snaps to prevent them from snagging or damaging the fabric during the wash.

I will give you an image. Based on it, please tell me next 3 steps on How to Connect a Computer to a TV via HDMI. Generate an image for each step, and write a brief description after each image.For example, <image1> [description1], <image2> [description2], <image3> [description3].

1): First, locate an available HDMI port on both your computer and television. These are usually labeled \"HDMI\" and have a distinct trapezoidal shape.



Please tell me 3 steps on How to Select a Site and Set Up a Camping Tarp. Generate an image for each step, and write a brief description after each image.For example, <image1> [description1], <image2> [description2], <image3> [description3].

I will give you an image. Based on it, please tell me next 3 steps on How to Knit the Waffle Stitch. Generate an image for each step, and write a brief description after each image.For example, <image1> [description1], <image2> [description2], <image3> [description3].

1): Begin by casting on a multiple of 3 stitches plus 2 extra stitches (e.g., 17 stitches) onto your knitting needles.



I will give you an image. Based on it, please tell me next 3 steps on How to Apply Spot-On Parasite Treatment to a Cat. Generate an image for each step, and write a brief description after each image.For example, <image1> [description1], <image2> [description2], <image3> [description3].

1): Preparation: Choose a vet-recommended spot-on treatment specifically designed for cats and your cat's weight. Ensure your cat is calm.

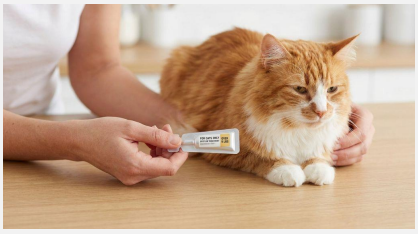


Figure 21: **How-to cases on SIL-Bench**

Based on these 1 images, continue the story, presenting it through 3 new PNG images with descriptions. Ensure each text includes connecting phrases for smooth transitions. For example, [description1] <image1>, [description2] <image2>.



Chapter 1: Welcome to the Sunshine Supply Hub! It is a very big place where thousands of useful things are kept safe and tidy. The wise owl manager keeps everything organized, and the busy squirrel team makes sure every box is in its perfect spot, waiting to be needed.



Continue the story, presenting it through 3 new PNG images with descriptions. Ensure each text includes connecting phrases for smooth transitions. For example, [description1] <image1>, [description2] <image2>.

Chapter 1: Once there was a girl who was tired of wearing the same old clothes every day. She wanted an adventure. While looking around her playroom, she spotted a dusty treasure chest in the corner that she hadn't opened in a long time.
This text describes 1 story scenes. Based on the text, continue the story, presenting it through 3 new PNG images with descriptions. Ensure each text includes connecting phrases for smooth transitions. For example, [description1] <image1>, [description2] <image2>.

Brave Explorer and the Glowing Cave Adventure'. Based on the image and the provided text prompts, continue the story, presenting it through 3 new PNG images with descriptions. Ensure each text includes connecting phrases for smooth transitions. For example, [description1] <image1>, [description2] <image2>.

Chapter 1: Once, there was a brave little explorer who loved adventures. One sunny morning, equipped with a helmet and a map, they stood before a mysterious cave entrance, ready to find the legendary hidden light. A tiny, friendly firefly buzzed around, eager to join the journey.



Chapter 1: The little blue truck finished his busy morning deliveries. 'Whew! Good work,' he beeped happily. Now it was time to find a place to rest.



Continue the story, presenting it through 3 new PNG images with descriptions. Ensure each text includes connecting phrases for smooth transitions. For example, [description1] <image1>, [description2] <image2>.

Figure 22: **Visual Storytelling cases on SIL-Bench**