EFFICIENT REINFORCEMENT FOR VISUAL-TEXTUAL THINKING WITH DISCRETE DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite promising multimodal understanding results, discrete diffusion models reason predominantly in text, using images as inputs but not as output for visual thinking that supports grounding and spatial reasoning. We introduce GRPO-based post-training that enables a unified discrete diffusion model to first generate an intermediate reasoning image and then use it to guide textual reasoning, while addressing two key challenges. First, to reduce visual rollout costs during RL, we leverage bidirectional attention for *localized visual editing*, denoising only a small subset of tokens while preserving all the other tokens. Second, we introduce *modality-factorized credit assignment* with segment-specific rewards, ensuring that image updates exclude future text while text updates use the completed image. Across five multimodal benchmarks, **LocFac-RL** improves average accuracy over standard GRPO by 7.9% on LaViDa-O and 11.9% on MMaDA-Parallel. Compared with global editing on the same backbones, it reduces training time by 12.3% and 17.1% while improving average accuracy by 4.1% and 38.3%, respectively.

1 INTRODUCTION

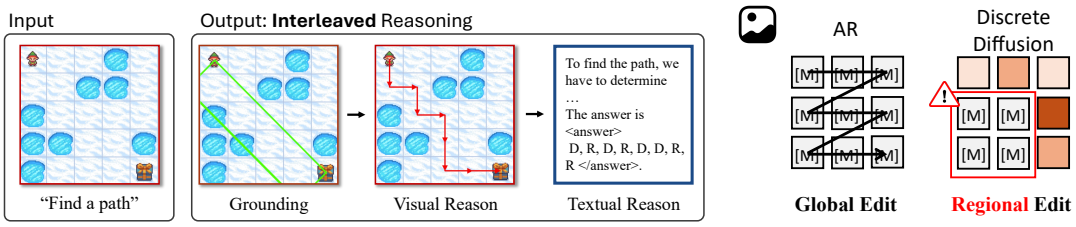
Multimodal discrete diffusion models have emerged as alternatives to fully autoregressive architectures (Swerdlow et al., 2025; Li et al., 2025e). Their iterative denoising process supports parallel token updates and bidirectional conditioning, enabling controllable refinement and flexible quality–efficiency trade-offs. Recent unified models extend this formulation to multimodal understanding and generation within a single framework, supporting visual interpretation, generation, and editing (AI et al., 2026; Li et al., 2025f; Swerdlow et al., 2025; Yang et al., 2026a; Wang et al., 2026; Shi et al., 2026; Li et al., 2026).

Prior work has demonstrated benefits from interleaved visual–textual reasoning, which creates intermediate visual states and produces textual reasoning traces as shown in Fig. 1a. Interleaved visual-textual reasoning outperforms text-only reasoning across diverse multimodal tasks, such as visual grounding and spatial reasoning tasks (Gu et al., 2026; Zhang et al., 2026a; Cheng et al., 2025). However, existing unified discrete diffusion models remain predominantly text-centric, neglecting the critical role of visual reasoning.

To address this limitation, we study a single transition from visual to textual reasoning as a first step toward interleaved visual–textual reasoning, training models to generate an intermediate visual edit before producing a textual response. While supervised fine-tuning provides direct supervision for intermediate visual states, it does not explicitly optimize these edits for downstream reasoning performance. We therefore investigate post-training with Group Relative Policy Optimization (GRPO; Shao et al., 2024) to refine visual–textual generation using task-level rewards. In this setting, we identify two key challenges in adapting GRPO to unified discrete diffusion models for visual–textual reasoning.

First, generating intermediate visual states incurs substantial rollout cost, which GRPO amplifies by sampling multiple responses per prompt. As illustrated in Fig. 1b, causal decoding in AR models constrains local editing, whereas bidirectional attention in discrete diffusion models enables an edited region to condition on the surrounding image. We exploit this capability through *localized visual editing*, masking and denoising only tokens within a predicted editing region while preserving the remaining image tokens. Unlike full-canvas AR and discrete diffusion baselines, our method avoids regenerating unchanged regions, reducing visual rollout cost.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Our goal: textual-visual interleaved reasoning.

(b) AR vs. Discrete Diffusion.

Figure 1: (a) We train a unified discrete diffusion model to localize and edit an image region, then use the edited image to guide textual reasoning. (b) Unlike AR models that regenerate the full image token sequence, discrete diffusion models support regional editing for efficient GRPO rollouts.

Second, DDMs create a conditioning mismatch between image rollouts and actor updates. During an actor update, bidirectional attention allows an earlier image segment to access subsequent response text, even though that text is unavailable when the image is generated during rollout. We address this mismatch with *modality-factorized credit assignment*, which evaluates earlier visual reasoning independently of subsequent text.

Our method **LocFac-RL**, **Loc**alized and **Fac**torized RL with discrete diffusion models, enhances visual-language reasoning with the following contributions:

- **Systematization of Multi-modal RL with Discrete Diffusion.** We develop an RL framework for multi-modal discrete diffusion models with bidirectional context modeling and parallel decoding.
- **Grounded Localized Visual Rollouts.** We combine edit-grounding SFT with localized visual editing during RL to reduce rollout costs while preserving editing quality and reasoning performance.
- **Modality-Factorized Credit Assignment.** We address cross-modal credit misassignment under parallel decoding with a factorized reward scheme that decouples image and text updates.

2 Related Works

RL for Visual-Textual Thinking. Works that directly leverage image generation into the reasoning process (Li et al., 2025c; Chern et al., 2025; Cheng et al., 2026; Xu et al., 2026), often rely on large-scale synthetic SFT data, sometimes combined with GRPO, to encourage coherent multi-step reasoning through generated visual artifacts. These works are primarily built on auto-regressive (AR) or hybrid architectures (Chern et al., 2024; Team, 2024), which require token-level sequential decoding. This leads to higher latency and makes localized visual updates inefficient, as modifications require recomputing downstream tokens due to causal dependencies.

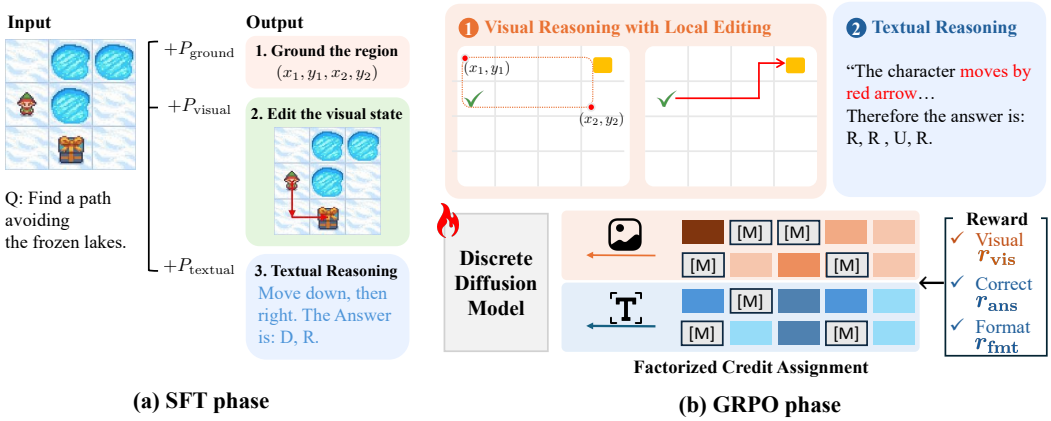
Unified Discrete Diffusion Models. Unified DDMs generate both images and text through iterative denoising in a discrete token space. Unlike AR models, DDMs allow parallel updates across tokens and naturally support masking-based conditioning. Representative multimodal diffusion models include Li et al. (2025f); Swerdlow et al. (2025); Yang et al. (2026a); Wang et al. (2026); Shi et al. (2026); Li et al. (2026); Tian et al. (2025). Among these models, LaViDa-O (Li et al., 2026) and MMaDA-Parallel (Tian et al., 2025) are models with image editing ability and textual reasoning, which are explicitly trained on large-scale image editing datasets, while others are focused on text-based image generation. Extended related works are introduced in Appendix A.

3 Preliminaries: GRPO with Discrete Diffusion Models

Let π_θ denote the current policy, and $\pi_{\theta_{\text{old}}}$ the old policy used to generate samples, while π_{ref} is a fixed reference policy. For each prompt $q \sim P(\mathcal{Q})$, the old policy samples a group of G responses, referred to as rollouts, $\mathcal{O} = \{\mathbf{o}_1, \ldots, \mathbf{o}_G\}$, each assigned a scalar reward r_i.

The GRPO objective maximizes a clipped policy gradient objective with KL regularization, applied at the token level for causal AR models:

$$\mathcal{L}^{\text{AR}}(\theta) = \mathbb{E}_{\substack{q\sim P(\mathcal{Q}) \\ \{\mathbf{o}_i\}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}} \left[\frac{1}{G}\sum_{i=1}^{G} \left(\frac{1}{|\mathbf{o}_i|}\sum_{t=1}^{|\mathbf{o}_i|} \left(\min\left(\rho_{i,t}^{\text{AR}}(\theta)A_i,\ \text{clip}_\epsilon(\rho_{i,t}^{\text{AR}}(\theta))A_i\right) \right) - \beta\, D_{\text{KL}}(\pi_\theta(\mathbf{o}_i \mid q) \,\|\, \pi_{\text{ref}}(\mathbf{o}_i \mid q)) \right) \right], \quad (1)$$



(a) SFT phase **(b) GRPO phase**

Figure 2: **Overview of the proposed training framework.** The framework consists of SFT and GRPO phases. (a) During SFT, the model is trained to generate grounding, visual reasoning, and textual reasoning in an interleaved input–output format with task-specific prompts P_{ground}, P_{visual}, and P_{textual}. (b) During GRPO, the model performs interleaved reasoning by sequentially generating: (i) grounding regions that require editing, (ii) localized visual edits on the predicted grounding, and (iii) textual reasoning based on the edited visual content. The rollout sequence is used to update the policy model in a *modality-factorized* manner.

where the importance sampling ratio at each token $o_{i,t}$ is $\rho_{i,t}^{\text{AR}}(\theta) = \frac{\pi_\theta(o_{i,t}|q,\mathbf{o}_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,\mathbf{o}_{i,<t})}$, and ϵ controls the clipping range between $1-\epsilon$ and $1+\epsilon$. Unlike standard PPO, GRPO assigns a group-normalized advantage shared across all tokens of a response as $A_i = \frac{r_i - \mu_{\mathcal{G}}}{\sigma_{\mathcal{G}} + \delta}$, where $\mu_{\mathcal{G}}$ and $\sigma_{\mathcal{G}}$ are the mean and standard deviation of rewards within the group $\mathcal{G} = \{r_1, \ldots, r_G\}$, and δ is a small constant for numerical stability.

Discrete diffusion models, unlike AR variants, do not have a sequential order. Therefore, following the previous work (Tian et al., 2025), the GRPO objective function for discrete diffusion model is defined as

$$\mathcal{L}^{\text{DDM}}(\theta) = \mathbb{E}_{\substack{q\sim P(\mathcal{Q}) \\ \{\mathbf{o}_i\}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}} \left[\frac{1}{G}\sum_{i=1}^{G} \left(\frac{1}{|\mathbf{o}_{i,M}|} \sum_{k=1}^{|\mathbf{o}_{i,M}|} \left(\min\left(\rho_{i,k}^{\text{DDM}}(\theta)A_i,\ \text{clip}_\epsilon(\rho_{i,k}^{\text{DDM}}(\theta))A_i\right) - \beta\, D_{\text{KL}}\left(\pi_\theta(\mathbf{o}_i \mid q) \,\|\, \pi_{\text{ref}}(\mathbf{o}_i \mid q)\right) \right) \right] . \quad (2)$$

Here, the importance sampling ratio for discrete diffusion models, $\rho_{i,k}^{\text{DDM}}(\theta)$, is defined over a masked token $o_{i,k} \in \mathbf{o}_{i,M}$, conditioned on the prompt q and the unmasked sequence $\mathbf{o}_{i,\overline{M}}$. More specifically, M is a binary mask over the sequence $\mathbf{o}_i$, while $\overline{M}$ denotes the complement.

Formally, importance sampling ratio for the masked token $o_{i,k} \in \mathbf{o}_{i,M}$ is defined as $\rho_{i,k}^{\text{DDM}}(\theta) = \frac{\pi_\theta(o_{i,k}|q,\mathbf{o}_{i,\overline{M}})}{\pi_{\theta_{\text{old}}}(o_{i,k}|q,\mathbf{o}_{i,\overline{M}})}$.

4 Methodology

We introduce **LocFac-RL**, a two-phase framework for training unified discrete diffusion models. In the first phase, supervised fine-tuning (SFT) equips the base model with grounding, visual editing, and textual reasoning capabilities. In the second phase, GRPO-based reinforcement learning (RL) further optimizes visual editing and textual reasoning using task-level rewards, encouraging intermediate visual states that support downstream reasoning rather than merely imitating supervised targets. We describe the overall training pipeline in Section 4.1.

The GRPO phase incorporates two mechanisms designed specifically for unified discrete diffusion models: (i) *localized visual editing* (Section 4.2), which accelerates visual rollouts by restricting denoising to task-relevant regions, and (ii) *modality-factorized credit assignment* (Section 4.3), which prevents cross-modal interference in the GRPO objective.

4.1 Overall Framework: Cold-Start SFT and GRPO

Problem Formulation. Given an input image I_q and a text query T_q, the model π_θ produces an interleaved reasoning trace consisting of a generated image I_{gen} followed by textual reasoning T_{gen}. The image is synthesized *before* the text so that the textual reasoning can condition on the edited visual state. To enable localized editing, the model additionally predicts a bounding box B in pixel coordinate format that specifies the region of I_q to be modified. Each evaluated response contains one intermediate edited image followed by a textual response. The present experiments focus on a single transition from visual to textual reasoning, rather than multiple interleaved visual and textual reasoning steps.

SFT phase. The SFT phase equips the base model with two prerequisite abilities: predicting the grounding region that specifies where to edit, and generating visual and textual reasoning traces.

We construct a supervised dataset for grounding, localized image editing, and textual reasoning, as illustrated in Fig. 2a. For each image–text question, we use task-specific prompts P_{ground}, P_{visual}, and P_{textual}, paired with their respective target outputs: a bounding box for grounding, a locally edited image, and a textual reasoning trace. Full prompt templates are provided in Appendix B.

GRPO phase. Starting from the SFT-initialized policy π_θ, we sample visual-textual rollouts by generating a localized visual edit followed by a textual reasoning trace, from which we extract the final answer as shown in Fig. 2b. We then apply GRPO using rewards computed from the edited image I_{gen} and the textual reasoning trace T_{gen}:

- **Visual reward** $r_{\text{vis}} \in [-1, 1]$: CLIP similarity between I_{gen} and the ground-truth image I_{gt}.
- **Answer reward** $r_{\text{ans}} \in \{0, 1\}$: 1 if the answer extracted from T_{gen} matches the ground-truth answer, and 0 otherwise.
- **Format reward** $r_{\text{fmt}} \in \{0, 1\}$: 1 if the answer is enclosed in the required `<answer>` and `</answer>` tags, and 0 otherwise.

For each sampled response, we compute the visual reward $r_{i,\text{vis}}$ and the textual reward $r_{i,\text{ans}} + \frac{1}{2} r_{i,\text{fmt}}$. The advantage calculation is explained in Section 4.3. Also, as visual quality indirectly reflects grounding accuracy, we do not assign a separate reward to the grounding bounding box.

4.2 Localized Editing in Visual Rollout for Accelerated GRPO.

A key bottleneck in the RL stage is the cost of visual rollouts: naive GRPO training with discrete diffusion models regenerates the entire image at every rollout, incurring substantial denoising overhead. We alleviate this cost by exploiting the bidirectional attention of DDMs to perform *localized editing*, in which only the image tokens inside a predicted grounding region are updated while the surrounding tokens are kept fixed. This is fundamentally infeasible for AR models: their causal structure means that any change to an earlier token propagates to all subsequent tokens, forcing global regeneration even for spatially local edits.

Localized editing procedure. Given an input image $I_q \in \mathbb{R}^{H \times W}$ and text query T_q, the localized editing rollout proceeds in three steps: (1) **Grounding prediction.** The model first predicts a bounding box $B = (x_1, y_1, x_2, y_2)$ in pixel space, specifying the region to be edited. (2) **Coordinate downsampling.** The pixel-space box B is downsampled onto the VQ-latent grid of shape $H_{\text{VQ}} \times W_{\text{VQ}}$, yielding a corresponding region B_{VQ} over image tokens. (3) **Selective masking and denoising.** Image tokens inside B_{VQ} are replaced with the [MASK] token, while tokens outside B_{VQ} are copied directly from the input image and held fixed. The denoising process then reconstructs only the masked tokens, conditioned on the preserved context.

Computational cost. Let K denote the number of denoising steps for full-image editing, $N = H_{\text{VQ}} W_{\text{VQ}}$ the total number of image tokens, and n_{local} the number of tokens inside the predicted editing region. We scale the local denoising budget in proportion to the region size, $K_{\text{local}} = \max\left(1, \left\lceil K \cdot \frac{n_{\text{local}}}{N} \right\rceil\right)$. This allocation reduces the number of denoising forward passes for localized edits while preserving image tokens outside the region.

4.3 MODALITY-FACTORIZED CREDIT ASSIGNMENT

Our image-first rollout generates an intermediate image before producing a textual response. Joint masking of image and text tokens during actor updates can violate this order. Image-token predictions condition on response text that is unavailable during image generation. We introduce modality-factorized credit assignment to alleviate this issue.

Modality-Factorized Conditioning and Rewards. Let $X = [X_I, X_T]$ denote a multimodal output consisting of an image segment X_I followed by a textual reasoning segment X_T. We decompose the reward as

$$R(X) = R_1(X_I) + R_2(X_I, X_T),$$

where $R_1(X_I) = r_{\text{vis}}$ depends only on the generated image, and $R_2(X_I, X_T) = r_{\text{ans}} + \frac{1}{2} r_{\text{fmt}}$ evaluates the format and correctness of the final answer conditioned on that image.

We construct two actor-update inputs from each response. For the image update, we mask only image tokens, exclude the response text X_T from the conditioning context, and assign reward R_1. For the text update, we hold the completed image tokens fixed, mask only text tokens, and assign reward $R_1 + R_2$. Thus, each update uses the preceding context available in the image-first rollout, together with unmasked tokens from its own segment.

KL-regularized interpretation. Following the idealized policy improvement formulation of Liu et al. (2019), a policy update that maximizes expected reward while penalizing KL divergence from π_k has the form

$$\pi_{k+1}(X \mid q) \propto \pi_k(X \mid q) \exp\left(\beta^{-1} R(X)\right),$$

where $\beta > 0$ controls the strength of the KL penalty,

For image segment X_I and text segment X_T, let M denote the tokens being updated and $\overline{M}$ denote the complementary tokens fixed. Applying this formulation to interleaved output $X = [X_I, X_T]$, the idealized image and text updates, denoted by $\widetilde{\pi}^I_{k+1}$ and $\widetilde{\pi}^T_{k+1}$, are proportional to

$$\pi_k\left(X_{I,M} \mid q, X_{I,\overline{M}}\right) \exp\left(\beta^{-1} R_1(X_I)\right) \quad \text{and,} \tag{3}$$

$$\pi_k\left(X_{T,M} \mid q, X_I, X_{T,\overline{M}}\right) \exp\left(\beta^{-1} \left[R_1(X_I) + R_2(X_I, X_T)\right]\right), \quad \text{respectively.} \tag{4}$$

GRPO implementation. Motivated by these idealized conditional updates[1], we train a shared policy using separate masked-token GRPO objectives for the image and text segments, with the corresponding segment-specific rewards. For a group of G sampled responses, the corresponding advantages are

$$A^I_i = \frac{R_1(X_{I,i}) - \mu_I}{\sigma_I + \delta}, \qquad A^T_i = \frac{R_1(X_{I,i}) + R_2(X_{I,i}, X_{T,i}) - \mu_T}{\sigma_T + \delta},$$

where (μ_I, σ_I) and (μ_T, σ_T) are the group statistics of the visual and combined rewards, respectively. Each objective is evaluated only on masked tokens in its own segment, using the conditioning contexts specified above. This prevents direct access to future response text during image updates while allowing text updates to use the completed visual state.

5 EXPERIMENTS

In this section, we provide the details of training and evaluation configuration in Section 5.1. We demonstrate the performance and the efficiency gain of our method through experiments in Section 5.2. Finally, we demonstrate component-wise ablations on factorized credit assignment, local region editing, and vision-language modalities in Section 5.3.

[1]For a fixed image, $R_1(X_I)$ cancels upon normalization of the ideal text conditional in Eq. (4), but it can affect GRPO text advantages when the group contains responses with different images.

Table 1: Performance comparison on multimodal benchmarks. Scores are reported as accuracy (%). Δ (%) denotes the relative percentage improvement over the corresponding SFT baseline. GRPO uses global editing with unified credit assignment, whereas **LocFac-RL** uses local editing with factorized credit assignment.

Arch	Model	MM-Vet	MMMU (val)	VStarBench	ChartQA	BLINK Jigsaw	Average Accuracy
AR	**ANOLE**	7.8	2.6	7.3	**7.3**	28.0	10.6
	+ SFT	7.8	20.2	18.3	**5.8**	36.7	17.76
	+ GRPO	**9.6**	**21.0**	**20.9**	5.7	**38.7**	**19.18** (△+8.00%)
D-Diff	**LaViDa-O**	9.36	43.33	19.90	38.68	44.77	31.21
	+ SFT	11.93	43.11	31.94	19.64	43.61	**30.05**
	+ GRPO	18.81	42.56	33.84	62.40	45.25	**40.57** (△+35.01%)
	+ **LocFac-RL**	**23.94**	**43.67**	**36.13**	**67.12**	**48.00**	**43.77** (△+45.66%)
	MMaDA-Parallel	5.05	0.00	0.00	0.00	0.68	1.15
	+ SFT	**14.68**	7.67	13.61	3.64	52.67	18.45
	+ GRPO	7.37	17.67	27.75	3.96	**54.00**	**22.15** (△+20.05%)
	+ **LocFac-RL**	12.84	**24.89**	**28.27**	**4.6**	53.33	**24.79** (△+34.36%)

5.1 Experimental Setup

Models. We evaluate our method on two unified discrete diffusion models: LaViDa-O and MMaDA-Parallel. Among open-source multimodal discrete diffusion models, we use LaViDa-O and MMaDA-Parallel since they are pre-trained for image editing, while MMaDA (Yang et al., 2026a) and Muddit (Shi et al., 2026) are limited to generation without editing capabilities. Following prior work (Cheng et al., 2026), we adopt ANOLE as an AR baseline.

Dataset. We partition the 19K *ThinkMorph*[2] examples (Gu et al., 2026) into three disjoint subsets: 15K for SFT, 3.8K for GRPO, and 200 for held-out evaluation. For SFT, we combine the 15K ThinkMorph examples with 180K examples from *ZebraCoT* (Li et al., 2025a), yielding 195K examples in total. Because these datasets lack explicit spatial supervision for intermediate edits, we automatically identify modified regions by comparing each reasoning image with its source image. Details of automatic grounding region identification are provided in Appendix C.

For GRPO, we use the 3.8K ThinkMorph examples together with 1K randomly sampled examples from *ArxivQA* (Li et al., 2024), following prior works (Cheng et al., 2026; Zhao et al., 2026a). The 200 held-out ThinkMorph examples are excluded from both training stages.

Evaluation Setting. Across all models, generation follows an interleaved reasoning format: image editing is performed first, followed by textual reasoning conditioned on the generated visual and textual context. We use greedy decoding (temperature 0.0), following the baseline model's hyperparameter settings, with task-specific maximum generation lengths reported in Table 7 in the appendix. Since decoding is deterministic, we do not repeat evaluations across random seeds.

We evaluate our method on MM-Vet, MMMU-val, V*Bench, ChartQA, and BLINK-Jigsaw using the LMMs-Eval package. For benchmarks requiring 1024 tokens, MM-Vet and BLINK, we utilize Chain-of-Thought (CoT) prompting with task-specific default templates. Performance is measured via exact-match accuracy. For MM-Vet, we measure the accuracy using the LLM judge GPT-4-0613. Other training and evaluation hyper-parameters are specified in Appendices D and E.

Baselines. We compare our method against two axes:

- **GRPO Baseline.** The two components of **LocFac-RL** are *local* editing and *factorized* credit assignment. Regular GRPO, which uses *global* editing and *unified* credit assignment, serves as the baseline for our proposed method and is denoted as *GRPO* in the tables. For the ablations, we introduce the following variants by ablating either 'local' editing ($\leftrightarrow$ global editing) or 'factorized' credit assignment ($\leftrightarrow$ unified credit assignment) of **LocFac-RL**.

[2]We exclude ThinkMorph Jigsaw Assembly from grounding-prompt supervision because it requires global canvas modifications and therefore has no localized edit target.

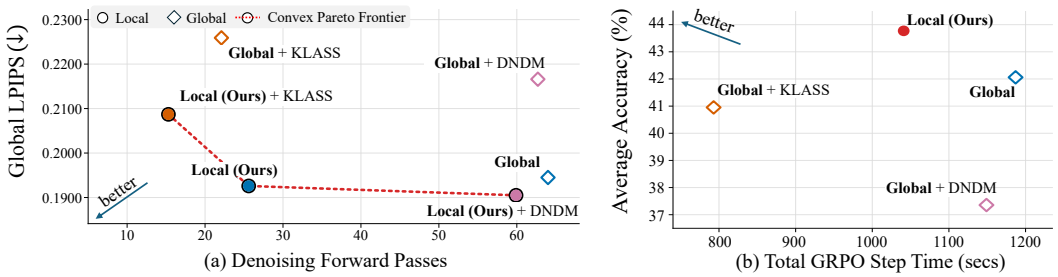


Figure 3: **Visual-rollout efficiency of LocFac-RL.** (a) LPIPS (lower is better) versus denoising forward passes on ThinkMorph. Local editing achieves comparable or lower global LPIPS with fewer denoising forward passes in the evaluated configurations. (b) Mean accuracy across five benchmarks versus total GRPO step time, including rollout and actor updates. *Global* and *Local* are full-canvas and target-region baselines without fast sampling; *KLASS* and *DNDM-k-Absorb* use fast sampling. All runs use factorized credit assignment with LaViDa-O.

- **Training-free Fast Sampling Baseline.** We select two representative methods for training-free fast unmasking strategies: KLASS (Kim et al., 2025) and DNDM-k-Absorb (Chen et al., 2024). KLASS accelerates the sampling speed by selecting tokens to unmask with high predictive confidence and low KL divergence between consecutive steps. DNDM-k-Absorb decides the number of unmasking tokens in advance from non-Markov absorbing diffusion process, and reveal the tokens based on the confidence order.

5.2 Efficiency and Performance Analysis

Performance Analysis. Table 1 shows the post-training effects on the AR model, Anole, and two discrete diffusion models, LaViDa-O and MMaDA-Parallel. For Anole, SFT provides most of the improvement, while RL yields a further gain. On both diffusion backbones, **LocFac-RL** outperforms standard GRPO and exceeds the average accuracy of the evaluated Anole GRPO baseline under the common image-first evaluation protocol.

We omit results for RL without SFT phase because the model fails to reliably follow the structured output formats required for reward extraction and lacks a stable starting policy for the complex reasoning and image generation tasks necessary for effective reinforcement learning. Note that the base MMaDA-Parallel model achieves a score of only 1.15. While trained for image editing rather than complex reasoning, it also fails to follow the specified answer format, further limiting its measured performance.

Efficiency Analysis. To evaluate the efficiency of **LocFac-RL**, we compare it with the training-free fast sampling methods KLASS and DNDM-k-Absorb, each applied to both global and local editing, in Figure 3a. **LocFac-RL**, labeled **Local (Ours)**, achieves high image editing quality (LPIPS ↓) compared to all global editing variants combined with fast sampling methods. It also requires fewer denoising forward passes, except for Global + KLASS, which is slightly faster but produces lower-quality edits. Combining local editing with fast sampling provides additional speed–quality trade-offs. KLASS reduces denoising passes at higher LPIPS, whereas DNDM-k-Absorb lowers LPIPS at the cost of more passes. Together, these configurations form the observed Pareto frontier.

Figure 3b compares downstream accuracy and GRPO step time. Local editing without fast sampling achieves a higher mean benchmark score than all global editing variants with fast sampling. Global + KLASS is faster but scores 2.82%p lower than **LocFac-RL**. Global + DNDM-k-Absorb provides neither a time nor an accuracy advantage over **LocFac-RL**. Overall, **LocFac-RL** achieves the highest mean benchmark score while offering a favorable trade-off between rollout cost and visual fidelity. Extensive results over five benchmarks are reported in Appendix F.

5.3 Ablation Studies

Beyond the combined effect of local editing and factorized credit assignment shown in Table 1, we evaluate credit assignment while holding local editing fixed, and editing strategy while holding

Table 2: **Ablation on two components** (a) factorized credit assignment and (b) local editing. Scores are reported in accuracy (%).

(a) Ablation on factorized credit assignment. Experiments are conducted with local editing.

		MM-Vet	MMMU (val)	VStarBench	ChartQA	BLINK Jigsaw	**Average Accuracy**
LaViDa-O	Unified	23.58	42.56	27.75	**67.40**	47.33	41.72
	Factorized	**23.94**	**43.67**	**36.13**	67.12	**48.00**	**43.77** (Δ+4.9%)
MMaDA-Parallel	Unified	10.09	**25.89**	**30.37**	4.44	47.67	23.69
	Factorized	**12.84**	24.89	28.27	**4.60**	**53.33**	**24.79** (Δ+4.6%)

(b) Ablation on local editing. Experiments for LaViDa-O and MMaDA-Parallel are conducted with factorized credit assignment. ANOLE with standard GRPO is included as an AR reference.

		MM-Vet	MMMU (val)	VStarBench	ChartQA	BLINK Jigsaw	**Average Accuracy**	**Train Time** (sec/step)
Anole (AR)	Global Edit	9.6	21.0	20.9	5.7	38.7	19.18	2489
LaViDa-O	Global Edit	19.71	**43.78**	**37.17**	66.16	43.48	42.06	1187
	Local Edit	**23.94**	43.67	36.13	**67.12**	**48.00**	**43.77** (Δ+4.1%)	1041 (Δ-12.3%)
MMaDA-Parallel	Global Edit	**12.84**	9.89	11.00	4.56	51.33	17.92	1221
	Local Edit	**12.84**	**24.89**	**28.27**	**4.6**	**53.33**	**24.79**(Δ+38.3%)	1012 (Δ-17.1%)

Table 3: Ablation on Modality during Reinforcement Learning Training. Scores reported in accuracy (%).

Model	**RL Variant**	MM-Vet	MMMU (val)	VStarBench	ChartQA	BLINK Jigsaw	**Average Accuracy**
LaViDa-O	Image Only RL	18.35	42.78	3.60	60.20	45.75	34.14
	Text Only RL	17.43	42.60	**36.13**	57.24	45.75	39.83
	Both (**LocFac-RL**)	**23.94**	**43.67**	**36.13**	**67.12**	**48.00**	**43.77**
MMaDA-Parallel	Image Only RL	9.63	14.78	26.18	3.60	48.00	20.44
	Text Only RL	11.93	15.33	26.18	3.84	**55.33**	22.52
	Both (**LocFac-RL**)	**12.84**	**24.89**	**28.27**	**4.60**	53.33	**24.79**

factorized credit assignment fixed. We also compare single-modality and interleaved text–image reasoning under GRPO.

We additionally compare GRPO results under single-modality reasoning (text-only or image-only) and interleaved text–image reasoning to assess the benefit of combining both reasoning modalities.

Factorized vs. Unified Credit Assignment. With local editing, factorized credit assignment improves average accuracy on both models (Table 2a). Under global editing, however, it benefits LaViDa-O but degrades MMaDA-Parallel. This interaction suggests that the effectiveness of factorized credit assignment depends on the editing strategy.

Global vs. Local Editing. With factorized credit assignment, local editing improves average accuracy while reducing GRPO step time on both models (Table 2b), with a particularly large accuracy gain on MMaDA-Parallel. These results support local editing as an effective and efficient complement to factorized updates, although improvements vary across benchmarks.

Single-Modality vs. Interleaved Reasoning. We compare joint visual–textual reasoning with single-modality rollouts that use either visual or textual reasoning (Table 3). Interleaved reasoning achieves the highest average accuracy on both models and outperforms either baseline on most benchmarks, suggesting that visual and textual reasoning provide complementary benefits during RL.

5.4 Grounding and Visual Reasoning Analysis

Qualitative analysis. Fig. 4 compares intermediate visual reasoning outputs across training stages. In the selected examples, **LocFac-RL** improves the alignment of visual edits with the question. Also, the completeness of image reconstruction is greatly enhanced relative to the base and SFT models. Across the illustrated examples, **LocFac-RL** produces more task-relevant visual edits, including more accurate localization and complete reconstruction of the missing image region. We provide extensive qualitative samples in Section G.1.

Quantitative Analysis. Following the Supervised Fine-Tuning (SFT) phase, we observe a marked improvement in the alignment between predicted and ground-truth bounding boxes when identifying

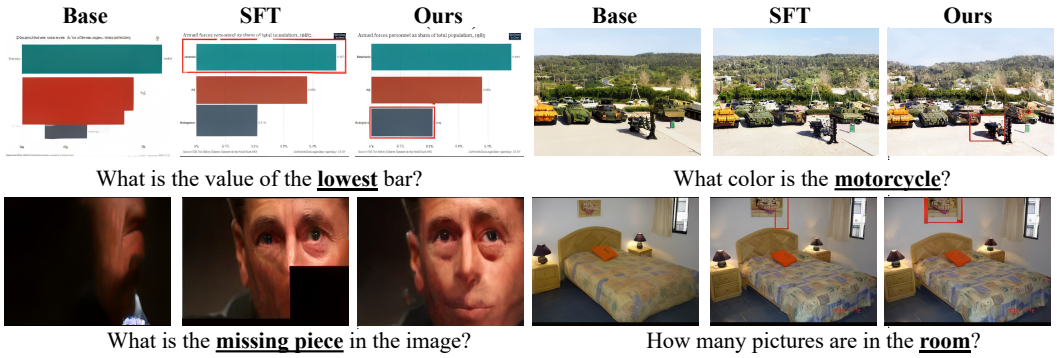


Figure 4: Qualitative examples of intermediate visual reasoning produced by the base model (LaViDa-O), the SFT model, and **LocFac-RL**.

localized regions for editing. This spatial precision is critical for the success of region-specific denoising. To quantify this progress, we evaluate the Intersection over Union (IoU) scores for LaViDa-O.

Table 4 shows the improvement in grounding and visual reasoning ability, tested with held-out evaluation set from ThinkMorph. The mean grounding IoU increases from 0.384 before task-specific post-training to 0.716 after SFT and 0.826 after **LocFac-RL**.

We also evaluate reference-image similarity using CLIP similarity (ViT-L/14). The mean similarity increases from 0.637 to 0.825 after SFT and to 0.831 after **LocFac-RL**. These results indicate improved similarity to the reference images.

Table 4: Grounding accuracy and visual reasoning fidelity Analysis.

	Grounding (IoU)	Visual Reasoning (CLIP sim)
Base model	0.384	0.637
+ SFT	0.716	0.825
+ **LocFac-RL**	**0.826**	**0.831**

Because global similarity may remain high even when a small edit is incorrect, this metric does not by itself establish the correctness or usefulness of the visual reasoning step. We provide the failure cases and analysis for such failure case in appendix Section G.2.

6 Limitations

Our framework has three limitations. First, its discrete modality space does not support the high-fidelity synthesis offered by continuous-latent diffusion models, therefore using the discrete model to control diffusion decoders is a potential extension. Second, RL depends on SFT initialization and is sensitive to reward signals, motivating robust cold-start SFT data and reliable image and text rewards to stabilize exploration. Third, local editing requires the predicted region to cover all necessary changes. Grounding errors can exclude relevant content, while global transformations, such as zooming or novel-view synthesis, may exceed a small region. Enlarging the region accommodates broader changes but reduces computational savings.

7 Conclusion

We have introduced **LocFac-RL**, a unified framework for interleaved visual-language reasoning that enables efficient learning across text and visual domains. Our approach aims to equip models with the ability to generate, refine, and ground intertwined textual and visual outputs in open-ended environments. Through supervised fine-tuning for interleaved reasoning, a reinforcement learning framework for general applicability, accelerated rollout via localized editing, and a masked discrete diffusion language modeling strategy for credit assignment across modalities, our method significantly improves both the scalability and effectiveness of multimodal learning, addressing key limitations of auto-regressive baselines and costly rollout generation.

AI USE STATEMENT

In this work, we used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, assist in the writing of proofs, implement methods, and design or provide feedback on research methodology or experiments. We have not used generative AI tools for providing critical ingredients for proving mathematical claims, supporting qualitative and thematic data analysis, assisting with translation, and interpreting results. Generating synthetic data, cleaning and reformatting dataset are not applicable to this work.

Additionally, we used generative AI tools to summarize or analyze existing literature, sourcing/searching for information, identify relevant literature, and propose a title or keywords for a research paper. We have reviewed all AI-assisted work for potential plagiarism or hallucination by manually checking the sources. Also, the LLM-generated code was verified and tested for correctness by one author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

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.

Jinbin Bai, Tian Ye, Wei Chow, Enxin Song, Qing-Guo Chen, Xiangtai Li, Zhen Dong, Lei Zhu, and Shuicheng YAN. Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=GJsuYHhAga.

Zixiang Chen, Huizhuo Yuan, Yongqian Li, Yiwen Kou, Junkai Zhang, and Quanquan Gu. Fast sampling via discrete non-markov diffusion models with predetermined transition time. *Advances in Neural Information Processing Systems*, 37:106870–106905, 2024.

Dongjie Cheng, Yongqi Li, Zhixin Ma, Hongru Cai, Yupeng Hu, Wenjie Wang, Liqiang Nie, and Wenjie Li. Omni-r1: Towards the unified generative paradigm for multimodal reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 38907–38924, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1937. URL https://aclanthology.org/2026.findings-acl.1937/.

Zihui Cheng, Qiguang Chen, Xiao Xu, Jiaqi Wang, Weiyun Wang, Hao Fei, Yidong Wang, Alex Jinpeng Wang, Zhi Chen, Wanxiang Che, et al. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. *arXiv preprint arXiv:2505.15510*, 2025.

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.

Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. Thinking with generated images. *arXiv preprint arXiv:2505.22525*, 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.

Shengyuan Ding, Xinyu Fang, Ziyu Liu, Yuhang Zang, Yuhang Cao, Xiangyu Zhao, Haodong Duan, Xiaoyi Dong, Jianze Liang, Bin Wang, et al. Arm-thinker: Reinforcing multimodal generative reward models with agentic tool use and visual reasoning. *arXiv preprint arXiv:2512.05111*, 2025.

Lijie Fan, Luming Tang, Siyang Qin, Tianhong Li, Xuan Yang, Siyuan Qiao, Andreas Steiner, Chen Sun, Yuanzhen Li, Tao Zhu, et al. Unified autoregressive visual generation and understanding with continuous tokens. *arXiv preprint arXiv:2503.13436*, 2025.

Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=mB3vxfrQZM.

Zemin Huang, Zhiyang Chen, Zijun Wang, Tiancheng Li, and Guo-Jun Qi. Reinforcing the diffusion chain of lateral thought with diffusion language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=aDTcN3yZGE.

Seo Hyun Kim, Sunwoo Hong, Hojung Jung, Youngrok Park, and Se-Young Yun. Klass: Kl-guided fast inference in masked diffusion models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 92267–92301. Curran Associates, Inc., 2025. doi: 10.52202/085713-3087. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/8514c1f12dcdd061ac8521feb78f28ca-Paper-Conference.pdf.

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

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*, 2025b.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. In *International Conference on Machine Learning*, pp. 36340–36364. PMLR, 2025c.

Han Li, Xinyu Peng, Yaoming Wang, Zelin Peng, Xin Chen, Rongxiang Weng, Jingang Wang, Xunliang Cai, Wenrui Dai, and Hongkai Xiong. Onecat: Decoder-only auto-regressive model for unified understanding and generation. *arXiv preprint arXiv:2509.03498*, 2025d.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14369–14387, 2024.

Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. Lavida: A large diffusion language model for multimodal understanding. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 105101–105134. Curran Associates, Inc., 2025e. doi: 10.52202/085713-3511. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/975affbe7b5f3b55fba62247b6877b1c-Paper-Conference.pdf.

Shufan Li, Jiuxiang Gu, Kangning Liu, Zhe Lin, Zijun Wei, Aditya Grover, and Jason Kuen. Lavida-o: Elastic large masked diffusion models for unified multimodal understanding and generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=zV5FeNiWDO.

Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2779–2790, 2025f.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural trust region/proximal policy optimization attains globally optimal policy. *Advances in neural information processing systems*, 32, 2019.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, JUN ZHOU, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=KnqiC0znVF.

Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pp. 2564–2571, USA, 2011. IEEE Computer Society. ISBN 9781457711015. doi: 10.1109/ICCV.2011.6126544. URL https://doi.org/10.1109/ICCV.2011.6126544.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Qingyu Shi, Jinbin Bai, Zhuoran Zhao, Wenhao Chai, Kaidong Yu, Jianzong Wu, Shuangyong Song, Yunhai Tong, Xiangtai Li, Xuelong Li, and Shuicheng YAN. Beyond text-to-image: Liberating generation with a unified discrete diffusion model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=pG0WTde3pR.

Alexander Swerdlow, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. Unified multimodal discrete diffusion. *arXiv preprint arXiv:2503.20853*, 2025. doi: 10.48550/arXiv.2503.20853.

Hongxuan Tang, Hao Liu, and Xinyan Xiao. Ugen: Unified autoregressive multimodal model with progressive vocabulary learning. *arXiv preprint arXiv:2503.21193*, 2025.

Xiaohang Tang, Rares Dolga, Sangwoong Yoon, and Ilija Bogunovic. wd1: Weighted policy optimization for reasoning in diffusion language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=L2rfd2Czbj.

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

Ye Tian, Ling Yang, Jiongfan Yang, Anran Wang, Yu Tian, Jiani Zheng, Haochen Wang, Zhiyang Teng, Zhuochen Wang, Yinjie Wang, et al. Mmada-parallel: Multimodal large diffusion language models for thinking-aware editing and generation. *arXiv preprint arXiv:2511.09611*, 2025.

Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025.

Jin Wang, Yao Lai, Aoxue Li, Shifeng Zhang, Jiacheng Sun, Ning Kang, Chengyue Wu, Zhenguo Li, and Ping Luo. Fudoki: Discrete flow-based unified understanding and generation via kinetic-optimal velocities. *Advances in Neural Information Processing Systems*, 38:81402–81440, 2026.

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.

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

Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=wsnse46kRO.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. MMaDA: Multimodal large diffusion language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026a. URL https://openreview.net/forum?id=wczmXLuLGd.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33510–33520, 2026b.

Lijun Yu, Jose Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A Ross, and Lu Jiang. Language model beats diffusion - tokenizer is key to visual generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gzqrANCF4g.

Lei Zhang, Junjiao Tian, Zhipeng Fan, Kunpeng Li, Jialiang Wang, Weifeng Chen, Markos Georgopoulos, Felix Juefei-Xu, Yuxiang Bao, Julian McAuley, et al. Think in strokes, not pixels: Process-driven image generation via interleaved reasoning. *arXiv preprint arXiv:2604.04746*, 2026a.

YiFan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think beyond images. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=gCWLkqK45O.

Kesen Zhao, Beier Zhu, Junbao Zhou, Xingyu Zhu, Zhongqi Yue, and Hanwang Zhang. Thinking with images as continuous actions: Numerical visual chain-of-thought. *arXiv preprint arXiv:2602.23959*, 2026a.

Siyan Zhao, Devaansh Gupta, Qinqing Zheng, and Aditya Grover. d1: Scaling reasoning in diffusion large language models via reinforcement learning. *Advances in Neural Information Processing Systems*, 38:56729–56762, 2026b.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and XingYu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xUyMXkI958.

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 *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=SI2hI0frk6.

Zetong Zhou, Dongping Chen, Zixian Ma, Zhihan Hu, Mingyang Fu, Sinan Wang, Yao Wan, Zhou Zhao, and Ranjay Krishna. Reinforced visual perception with tools. *arXiv preprint arXiv:2509.01656*, 2025b.

A Extended Related Works

A.1 Unified Multimodal Generative Models

Auto-regressive models. AR models treat multi-modal data as a single sequence and generate tokens causally. Early approaches rely on discrete visual tokens via vector quantization (e.g., Chameleon (Team, 2024), Anole (Chern et al., 2024), Janus (Wu et al., 2025), Emu3 (Wang et al., 2024)). More recent works explore different visual representations: discrete-token models such as UGen (Tang et al., 2025) and OneCAT (Li et al., 2025d) unify text and image tokens in a shared vocabulary, while continuous-token models such as UniFluid (Fan et al., 2025) remove the quantization bottleneck by directly modeling continuous latent embeddings.

A key limitation of AR models lies in their causal structure. Image generation and editing must follow a predefined token order, making localized editing difficult. Practical solutions rely on masking or token reordering, but cannot fully recover bidirectional conditioning.

Hybrid models. Hybrid approaches combine AR-based text generation with continuous diffusion or flow-based image synthesis (e.g., BAGEL (Deng et al., 2025), Transfusion (Zhou et al., 2025a)). These models decouple semantic reasoning from visual generation, but require trajectory-level modeling and multi-step optimization, making them fundamentally different from both AR and discrete diffusion formulations.

A.2 Multi-Modal Discrete Diffusion Models

The generative process with discrete diffusion iteratively predicts the clean state $\mathbf{x}_0$ over masked positions given $\mathbf{x}_t$ as an input at time $t \in (0, 1]$. Starting from the fully masked sequence $\mathbf{x}_T$, a shared Transformer backbone predicts the categorical distribution over clean tokens, $p_\theta(\mathbf{x}_0 \mid \mathbf{x}_t)$, for all masked positions. The total denoising step K defines the sampling schedule with time t, and the noised sequence $\mathbf{x}_t$ is iteratively remasked using confidence score of the model until it reaches the fully denoised state at $\mathbf{x}_0$.

Multi-modal discrete diffusion models receive both image and text as inputs, and are capable of generating the image and text tokens in discrete space. More specifically, the input images are reconstructed into a sequence of visual tokens with image tokenizer, where MMaDA-Parallel uses MagVIT-v2 (Yu et al., 2024) and LaViDa-O uses the encoder of Meissonic (Bai et al., 2025), while text is processed by the discrete diffusion backbones tokenizer like LLaDA-8B (Nie et al., 2026). Both image and text tokens are jointly modeled within a unified probabilistic framework.

A.3 Visual Thinking and Intermediate Visual Reasoning

External tool-based methods. Zhang et al. (2026b); Zhou et al. (2025b); Ding et al. (2025); Zheng et al. (2026); Wang et al. (2025) augment models with explicit visual operations (e.g., zooming, cropping, region selection). These approaches operate directly in pixel space and iteratively refine visual evidence. They typically rely on reinforcement learning (e.g., GRPO) to encourage tool usage, as supervised fine-tuning alone is insufficient.

Internal latent reasoning. Some works (Yang et al., 2026b; Li et al., 2025b) perform reasoning over internal visual representations without explicit image manipulation. Training typically combines SFT with RL to stabilize multi-modal reasoning trajectories.

Intermediate image generation. Other approaches explicitly generate intermediate images as part of reasoning. For example, Chern et al. (2025) produces visual subgoals and self-refined hypotheses, while Xu et al. (2026) formulates reasoning as sequences of generated visual states. These methods often rely on synthetic SFT data combined with RL to encourage coherent multi-step reasoning.

Visual thinking methods emphasize iterative reasoning, bidirectional context, and explicit spatial grounding. However, they typically rely on external tools or multi-step interaction and do not optimize a unified generative likelihood. In contrast, unified AR and DDM models operate within a single generative framework. Notably, the need for bidirectional conditioning and flexible editing

highlighted by these works aligns more naturally with diffusion-based approaches than with auto-regressive models.

A.4 GRPO WITH DISCRETE DIFFUSION MODELS

In AR language models, GRPO relies on an exact token-level likelihood factorization, enabling unbiased importance sampling ratios and a sum of clipped policy gradients over tokens. In contrast, discrete diffusion language models generate text via iterative denoising without tractable sequence likelihoods, making GRPO's importance sampling ratio term approximate (Zhao et al., 2026b; Tang et al., 2026). Consequently, diffusion variants of GRPO fundamentally shift from token-level, causally factorized optimization to trajectory-level objectives over denoising processes.

For language-only diffusion, d1 (Zhao et al., 2026b) extends GRPO to masked diffusion by introducing a surrogate likelihood-based objective over denoised sequences; wd1 (Tang et al., 2026) replaces unstable importance ratios with a ratio-free weighted log-likelihood formulation to reduce variance; and lateral-thought-style method (Huang et al., 2025) shifts optimization to the diffusion chain itself, assigning credit across intermediate denoising states rather than tokens.

B PROMPT TEMPLATES

The prompt template for supervised finetuning and GRPO phase is shared to induce the desired behavior from the model at rollout. These are the list of templates used for training:

- **Grounding (P_{ground}):** `Your job is to identify the region where auxiliary line, box, or editing could help solve the following problem. Give bounding boxes in LOC format.`
- **Visual Reasoning (P_{visual}):** `Edit the region where auxiliary line, box, or drawing could help solve the following problem.`
- **Textual Reasoning (P_{textual}):** `Let's think step-by-step to solve the question. Put your final answer in <answer> </answer> tags.`

C SFT DATASET: GROUNDEDITREASON

To localize such edits, we first align each intermediate reasoning image with the original image by estimating a homography transformation using Oriented FAST and rotated BRIEF (ORB) features (Rublee et al., 2011), establishing pixel-wise spatial correspondence under geometric variations. We then compute a difference map between the aligned images. To improve robustness, we apply Gaussian smoothing to suppress noise and minor pixel-level discrepancies while preserving structured edits. The resulting map is thresholded and post-processed to extract a tight bounding box covering the most salient modified region. This procedure yields region-level supervision aligned with reasoning steps, enabling the model to associate intermediate reasoning with precise spatial edits and improving its readiness for grounded editing in the subsequent RL stage.

The dataset utilizes a normalized spatial representation where raw ground truth coordinates, $\mathbf{B}_{\text{raw}} = [x_{\min}, y_{\min}, x_{\max}, y_{\max}]$, are transformed from the original image dimensions to a standardized target resolution $H \times W$. For integration into the LaViDa-O Supervised Fine-Tuning (SFT) framework, these coordinates are serialized into a discrete `LOC` string format: the sequence is enclosed by structural tokens $\langle$LOC_BEGIN$\rangle$ and $\langle$LOC_END$\rangle$, with each integer coordinate v_i represented as $\langle$LOC_$v_i\rangle$.

In Fig. 5, we show samples from the augmented dataset, **GroundEditReason**. We construct the dataset by placing bounding boxes over the regions that differ between the ground-truth reasoning image and the original problem image.

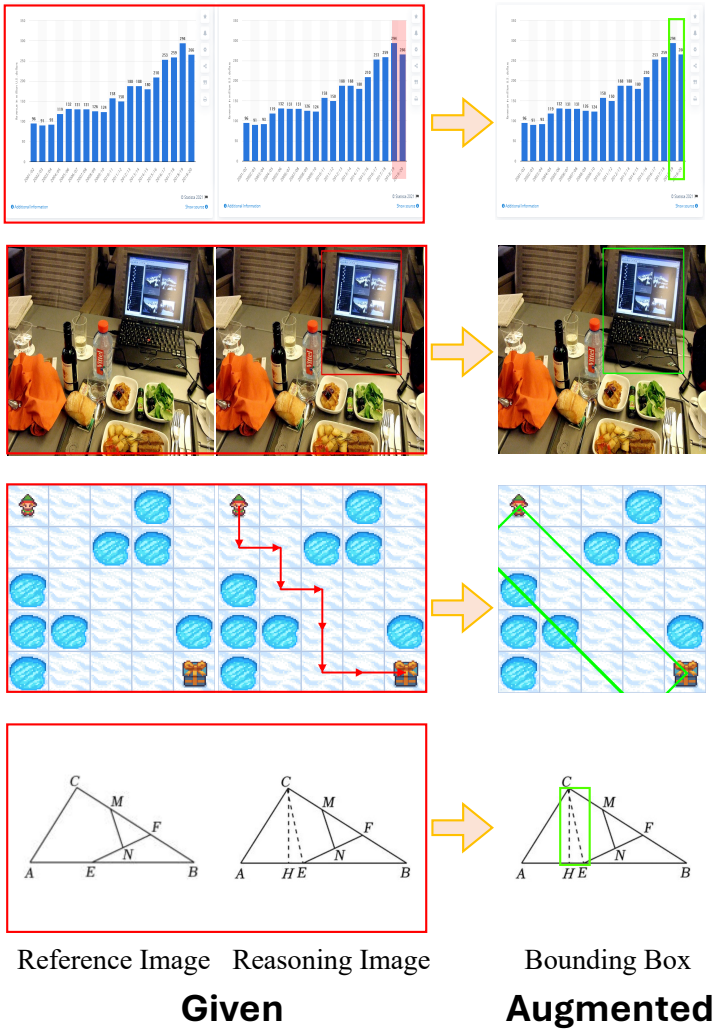


Figure 5: **GroundEditReason Sample.**

D TRAINING DETAILS

D.1 SUPERVISED FINE-TUNING DETAILS

SFT Dataset. The supervised finetuning dataset comprises 15K ThinkMorph examples from Spatial Navigation, Chart Refocus, and Visual Search, and 180K examples from ZebraCoT (Li et al., 2025a).

SFT Hyperparameters. We summarize the SFT hyperparameters in Table 5. The model is trained for 10 epochs with bf16 precision, cosine learning-rate decay with warmup, and a maximum sequence length of 8192 tokens.

Hyperparameter	Setting	Description
Training Setup		
Number of GPUs	8 (NVIDIA H100)	distributed training setup
Precision	bf16	mixed-precision training
DeepSpeed Stage	ZeRO-2	optimizer/state sharding
Number of Epochs	10	total training epochs
Optimization		
Learning Rate	5×10^{-6}	peak learning rate
Vision Tower Learning Rate	2×10^{-6}	vision encoder learning rate
Weight Decay	0.0	L2 regularization
Warmup Ratio	0.03	warmup proportion
LR Scheduler	cosine	decay schedule

Table 5: SFT hyperparameters used in our experiments. Settings are taken from the training configuration.

D.2 GRPO Details

GRPO Dataset. We use 3.8K ThinkMorph examples disjoint from the SFT and test splits, together with 1K randomly sampled examples from *ArxivQA*.

GRPO Hyperparameters. The hyperparameter settings for GRPO-based reinforcement learning experiments are reported in Table 6.

Hyperparameter	Setting	Description
Objective		
KL Loss Coefficient	0.04	KL regularization weight
PPO Clip Range (Low)	0.20	lower clipping bound (ϵ)
PPO Clip Range (High)	0.28	upper clipping bound
Batching		
Train Batch Size	64	prompts per update
Generations per Prompt	8	sampled responses per prompt
Effective Batch Size	512	64×8 rollouts
Optimization		
Epoch	1	Update epoch for GRPO
Learning Rate	1×10^{-5}	peak LR
Warmup Ratio	1×10^{-4}	LR warmup proportion
Adam β_1, β_2	(0.9, 0.99)	optimizer momentum terms
Weight Decay	0.10	L2 regularization
Gradient Clipping	1.0	global-norm clip
LR Scheduler	constant+warmup	schedule type
Rollout		
Temperature	0.6	sampling temperature
Max Prompt Length	2048	input token limit
Max Completion Length	512	output token limit
Diffusion Steps	256	steps for image generation
Image Edit Steps	64	iterative edit refinement
Image Generation Size	1024 (LaViDa-O) / 512 (MMaDA-Parallel & ANOLE)	generated image size
Denoise Tokens per Step (Text)	2	Mask sampling per generation

Table 6: GRPO training hyperparameters used in our experiments.

E Evaluation Details

Table 7: Evaluation Datasets and Generation Token Lengths in text. The number of image tokens is fixed to 1024 for both LaViDa-O and MMaDA-Parallel.

Dataset Name	**Text Token Length**	**Hugging Face Source (URL)**
MM-Vet	1024	huggingface.co/datasets/whyu/mm-vet
ChartQA	16	huggingface.co/datasets/lmms-lab/chartqa
V-STaR Bench	16	huggingface.co/datasets/craigwu/vstar_bench
BLINK	1024	huggingface.co/datasets/BLINK-Benchmark/BLINK
MMMU (val)	16	huggingface.co/datasets/MMMU/MMMU

Chain-of-Thought (CoT) Prompts. For datasets requiring reasoning, we prepend CoT instruction:‘`First please perform reasoning, and think step by step to provide best answer to the following question.`’

F Additional Experimental Results

Table 8 reports the full per-benchmark accuracy and timing results underlying the main-text visualization. **LocFac-RL** achieves the highest average accuracy and matches or exceeds both fast-sampling baselines on every benchmark. Global editing with KLASS is faster but yields lower accuracy, whereas global editing with DNDM-k-Absorb is both slower and less accurate than **LocFac-RL**. These results highlight the accuracy-efficiency trade-off with KLASS and show that fast sampling alone does not necessarily provide a training-time advantage over localized editing.

Table 8: Comparison of accuracy and training efficiency.

Method	MM-Vet	MMMU-Val	VStarBench	ChartQA	BLINK Jigsaw	Average	Visual Rollout Time ↓	Total GRPO Time ↓
Global + KLASS	20.18	43.22	34.03	60.00	47.33	40.95	**146**	**793**
Global + DNDM-k-Absorb	22.80	41.44	**36.13**	39.08	47.33	37.36	537	1149
Ours (LocFac-RL)	**23.94**	**43.67**	**36.13**	**67.12**	**48.00**	**43.77**	404	1,041

G QUALITATIVE RESULTS

G.1 VISUAL REASONING GENERATIONS.



Figure 6: Qualitative visualization of the interleaved reasoning process using localized editing.

To evaluate the visual coherence of the proposed method, we provide a qualitative assessment of the intermediate reasoning images generated by the unified models. Figures 6 and 7 show selected intermediate images produced before and after post-training.



Figure 7: Qualitative visualization of the interleaved reasoning process using localized editing.

G.2 FAILURE ANALYSIS

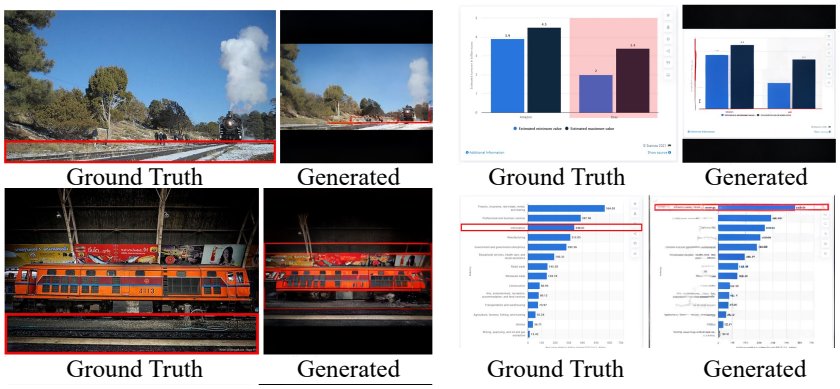


Figure 8: Failure samples with low IoU and CLIP similarity.

We also demonstrate the worst-4 cases where the IoU score and CLIP similarity are both the lowest in Fig. 8. The failure in identifying the edit region directly leads to compromised image generation.

H ASSETS AND EXTERNAL RESOURCES

We use several publicly available codebases, pretrained models, and evaluation frameworks in this work. All assets are used in accordance with their respective licenses and intended research usage policies.

- **TRL (Transformer Reinforcement Learning)** from Hugging Face is used for reinforcement learning training and optimization. License: Apache License 2.0. https://github.com/huggingface/trl
- **d1** is used as a reference implementation for diffusion-based reasoning and GRPO training. License: Apache License 2.0. https://github.com/dllm-reasoning/d1
- **Anole-7B** is used as a pretrained multimodal foundation model. License: Chameleon License (research-oriented license inherited from Meta Chameleon). https://huggingface.co/GAIR/Anole-7b
- **MMaDA-Parallel-M** is used as a pretrained multimodal diffusion model. License: MIT License. https://huggingface.co/tyfeld/MMaDA-Parallel-M
- **LaViDa-O-v1.0** is used as a pretrained multimodal diffusion model. License: Adobe Research License. https://huggingface.co/jacklishufan/LaViDa-O-v1.0
- **MMaDA-Parallel** inference and implementation codebase. License: MIT License. https://github.com/tyfeld/MMaDA-Parallel/tree/main
- **LaViDa-O** inference and implementation codebase. License: Adobe Research License. https://github.com/adobe-research/LaVida-O/tree/main
- **LMMS-Eval** is used as the evaluation framework for multimodal benchmarking. License: Apache License 2.0. https://github.com/evolvinglmms-lab/lmms-eval

We acknowledge and thank the original authors and maintainers of these resources for making their work publicly available. We follow the licenses, usage restrictions, and attribution requirements specified by each respective project and model release.