MaskSense: Confronting the Visual Exploration Trap in Masked Image Generation

Yawen Shao[1,2*] **Jie Xiao**[2] **Kai Zhu**[2†] **Yu Liu**[2] **Hongchen Luo**[3]
Xueyang Fu[1] **Yang Cao**[1] **Wei Zhai**[1†] **Zheng-Jun Zha**[1]
[1] University of Science and Technology of China
[2] Tongyi Lab, Alibaba Group [3] Northeastern University

Abstract

Reinforcement learning (RL) has shown strong potential for aligning generative models with human intent, but adapting RL to masked generative models (MGMs) remains largely underexplored. Existing approaches that directly adapt Group Relative Policy Optimization (GRPO) to MGMs leave a foundational conflict unaddressed: MGMs' exploitative decoding nature is inherently at odds with the exploratory diversity that GRPO presupposes. In this work, we reveal a structural decoding bias termed the **visual exploration trap**: confidence-based sampling trades the exploration of diverse subject realizations for the greedy resolution of low-uncertainty background regions, causing premature collapse of the generative exploration space and starving GRPO of the rollout diversity it depends on. To this end, we propose **MaskSense**, a novel RL framework for MGMs that confronts the visual exploration trap at both the sampling and optimization stages. Specifically, we introduce a semantic anchored routing sampling mechanism that leverages semantic priors to preserve high-entropy subject token exploration while amplifying intra-group reward variance through exploratory-exploitative routing to yield more discriminative advantage estimates. Furthermore, we design a global entropy transition anchoring strategy to identify the most consequential decoding steps and concentrate gradient updates on them. Extensive experiments on multiple text-to-image benchmarks demonstrate that MaskSense substantially improves upon the base model, with GenEval accuracy improving from 54% to 86% and HPS from 28.89 to 36.39, achieving state-of-the-art performance.

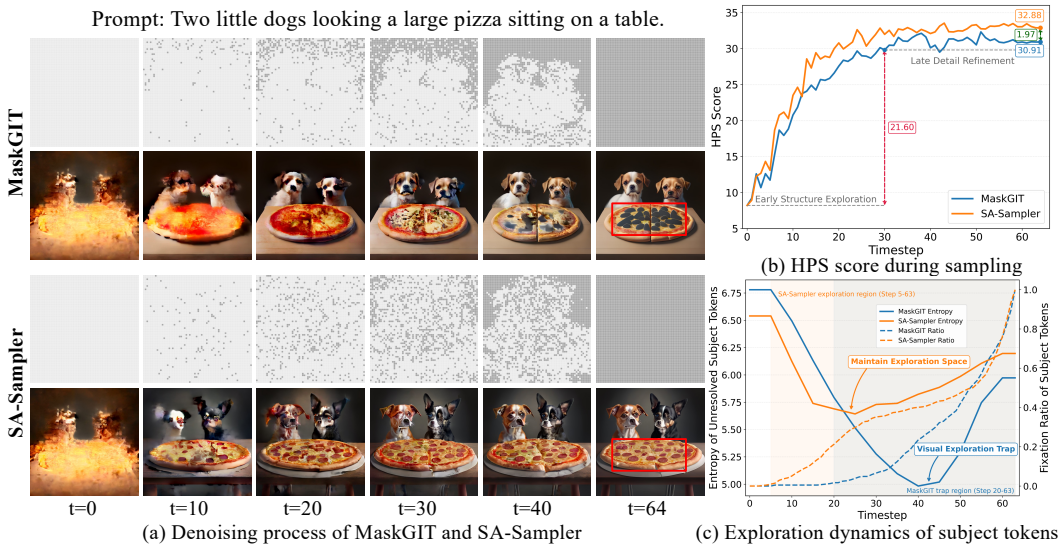
1 Introduction

Masked generative models (MGMs) [1, 5, 64, 15] have recently emerged as a compelling paradigm for visual synthesis, challenging the long-standing dominance of autoregressive models (AR) [58, 25, 49] and diffusion models (DMs) [47, 16, 24] by reformulating image generation as iterative masked token prediction. Central to this paradigm is the confidence-based sampling strategy of MaskGIT [6], which predicts all masked tokens in parallel at each step and retains only the most confident subset while remasking the remainder for subsequent iterations, making the decoding process exploitative.

Reinforcement learning (RL) [46, 41, 37] has proven highly effective in enhancing the reasoning capabilities of large language models (LLMs) [13, 17, 55, 42], while its application to MGMs remains comparatively underexplored relative to diffusion models [26, 43, 66] and AR models [18, 71]. Recent works [70, 29] have begun addressing this gap by applying Group Relative Policy Optimization (GRPO) to MGMs, yet retain the confidence-based sampler as a fixed component with minimal adaptation to the masked generation setting. This overlooks a fundamental prerequisite of

* Work done at Tongyi Lab.
† Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Denoising process of MaskGIT and SA-Sampler (b) HPS score during sampling (c) Exploration dynamics of subject tokens

Figure 1: **Motivation.** (a) MaskGIT keeps subject tokens masked while committing background regions early, whereas SA-Sampler resolves subject regions earlier via semantic anchoring. (b) SA-Sampler consistently discovers higher-reward trajectories through broader generative exploration. (c) SA-Sampler maintains higher subject-token fixation ratios and elevated entropy, effectively escaping the visual exploration trap.

GRPO: its effectiveness derives from group-relative advantage estimation, which presupposes rollouts sufficiently diverse to cover meaningful regions of the solution space, a requirement inherently at odds with the exploitative nature of confidence-based sampling.

In this paper, we identify and characterize precisely how this conflict manifests as a structural bias that we term the **visual exploration trap**. Specifically, tokens with lower uncertainty predominantly correspond to simple background textures in natural images [31], causing background regions to be committed well before semantically critical subject areas are ever explored (Fig. 1 (a)). Subject tokens, by contrast, retain high predictive entropy throughout the early generation steps, while their fixation ratio remains near zero (Fig. 1 (c)), indicating that substantial generative diversity is available yet remains entirely unexploited. By the time subject tokens begin to be retained after step 20, the surrounding background context has been largely committed, severely narrowing the space of potential generative branches [33, 55] and limiting the diversity and quality of the resulting subject regions [31]. The originally high entropy of subject tokens is thus suppressed before any meaningful subject decision is made, as the model effectively sacrifices diverse subject realizations to greedily resolve low-uncertainty background regions, directly undermining the rollout diversity that GRPO requires. This trap necessitates a rethink of RL for MGMs.

To tackle this challenge, we propose **MaskSense** (Fig. 2), a novel RL framework designed for MGMs that directly confronts the visual exploration trap at both the sampling and optimization stages. At the sampling stage, we propose the semantic anchored routing sampling (SARS) mechanism, built upon the Semantic Anchored Sampler (SA-Sampler), which employs semantic priors to anchor subject tokens during the early generation steps (Fig. 1 (a)). This design preserves the high-entropy exploration space that confidence-based sampling would otherwise suppress (Fig. 1 (c)), enabling the discovery of higher-reward trajectories through broader generative exploration (Fig. 1 (b)). SARS further routes rollouts stochastically between SA-Sampler and MaskGIT, generating a mixture of exploratory and exploitative trajectories within each group to deliberately amplify intra-group reward variance and sharpen discriminative advantage estimates. At the optimization stage, we further introduce global entropy transition anchoring (GETA) to concentrate gradient updates on the most structurally decisive decoding steps, avoiding diffuse optimization across steps of unequal importance. Extensive experiments on multiple text-to-image benchmarks validate the effectiveness of MaskSense.

The contributions of this paper are summarized as follows:

- We identify the visual exploration trap in MGMs, revealing that confidence-based sampling suppresses subject-token entropy before meaningful subject decisions are made, fundamentally depriving GRPO of the exploratory diversity essential to its optimization.

- We propose MaskSense, a novel RL framework for MGMs that promotes diverse rollout exploration through semantic anchored routing sampling, and concentrates gradient updates on the most consequential decoding timesteps via global entropy transition anchoring.
- Extensive experiments on multiple text-to-image benchmarks demonstrate that MaskSense achieves state-of-the-art performance, confirming the effectiveness of confronting the visual exploration trap for RL optimization in MGMs.

2 Related Work

Visual Generation Models. The rapid advancement of visual generation [53, 23, 1, 58, 63, 45, 73, 48, 62, 60, 54] has been propelled by three dominant architectures: DMs, AR, and MGMs. DMs [16, 47, 39, 35, 24, 27] reframe generation as an iterative denoising process, which suffers from slow inference. AR [50, 7, 34, 11, 59, 9, 67] predict discrete tokens sequentially via causal transformers, but their raster-order nature incurs prohibitive computational overhead at high resolution [38]. To address the limitations of both paradigms, MGMs [5, 6, 64, 1] cast image synthesis as an iterative mask prediction problem, where all tokens are decoded in parallel at each step and progressively refined conditioned on previously unmasked tokens. MaskGIT [6] pioneered this direction by introducing bidirectional masked modeling with effective mask scheduling and parallel decoding, establishing the foundational framework for subsequent works. Meissonic [1] further advances this paradigm through hybrid transformer architectures and feature compression, successfully scaling masked generative modeling to high-resolution text-to-image generation. The efficiency and growing capability of MGMs motivate further exploration of their potential through RL-based optimization.

Reinforcement Learning for Text-to-Image Generation. With recent advances in large language and multimodal modeling [74, 75, 21, 19, 57, 28, 44, 51], reinforcement learning (RL) [46, 41, 37] has proven effective in enhancing the reasoning capabilities of these models [17, 72, 8, 20], spurring its rapid adoption in text-to-image generation. Prior works primarily target preference alignment in diffusion-based models [52, 4], while more recent efforts successfully extend GRPO-based optimization to flow matching [26, 66, 43] and AR models [18, 71]. Within MGMs, MaskGRPO [29] formulates the iterative unmasking process as a multi-step decision-making problem and proposes transition probability to enable GRPO optimization. MaskGRPO [30] introduces modality-aware importance estimation to improve the image generation quality of multimodal discrete diffusion models. MaskFocus [70] identifies sampling steps via inter-step embedding similarity and incorporates dynamic sampling. However, none of these methods escape the confidence-based sampling strategy, leaving the visual exploration trap and principled gradient allocation across decoding steps fundamentally unaddressed as core challenges in RL optimization for MGMs.

3 Preliminaries

3.1 Masked Generative Models

Masked generative models (MGMs) can be formulated as discrete diffusion processes, in which a clean token sequence x_0 is progressively corrupted into either a new token or an absorbed state x_t at step t via a Markov process, and a learned reverse process subsequently reconstructs the original data from these corrupted observations. Specifically, the corruption mechanism operates over discrete tokens through a transition matrix Q_t, constructed as a sequential product of per-step matrices $Q_t = \prod_{s=1}^{t} Q_s$. The forward process is formulated as:

$$x_t \sim q(x_t|x_0, t), \quad q(x_t|x_0, t) = \text{Cat}(x_t;\, x_0 Q_t) \tag{1}$$

where Q_t encodes structured transition patterns, such as Gaussian kernels, nearest-neighbor mappings, or absorbing states. Meissonic adopts the absorbing state formulation, in which each token is independently replaced with a `[MASK]` token and remains in the absorbed state once masked.

To recover the original sequence, the reverse process $p_\theta(x_{t-1}|x_t)$ is modeled by a network that learns to progressively denoise the corrupted sequence, starting from the fully masked state x_T and iteratively transitioning toward the clean data x_0. The reverse process is formulated as:

$$p_\theta(x_{t-1}|x_t) = \sum_{x_0} q(x_{t-1}|x_t, x_0)\, p_\theta(x_0|x_t) \tag{2}$$

During inference, generation begins from a fully [MASK] canvas and proceeds over T iterative decoding steps [1]. At each step, the model predicts all masked tokens in parallel, commits a high-confidence subset, and remasks the remaining positions for subsequent refinement.

3.2 Group Relative Policy Optimization

Group Relative Policy Optimization (GRPO) [46] is a reinforcement learning algorithm that leverages the average reward of multiple sampled outputs as a dynamic baseline for advantage estimation. This principle has been recently adapted for masked generative models in Mask-GRPO [29], which applies GRPO to improve the performance of state-of-the-art MGMs [1, 63]. Given a prompt c, the policy model p_θ samples a group of G images $\left\{x_0^i\right\}_{i=1}^G$ along with the corresponding reverse-time trajectories $\left\{x_T^i, x_{T-1}^i, \cdots, x_0^i\right\}_{i=1}^G$. The advantage of the i-th image is then calculated by normalizing the group-level rewards as follows:

$$\hat{A}_t^i = \frac{R\left(x_0^i, c\right) - \operatorname{mean}\left(\left\{R\left(x_0^i, c\right)\right\}_{i=1}^G\right)}{\operatorname{std}\left(\left\{R\left(x_0^i, c\right)\right\}_{i=1}^G\right)} \tag{3}$$

Mask-GRPO [29] optimizes the policy model by maximizing the following objective:

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{c\sim\mathcal{C},\{x^i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|c)} \\ \left[\frac{1}{G}\frac{1}{T}\sum_{i=1}^G\sum_{t=1}^T\left(\min\left(r_t^i(\theta)\hat{A}_t^i, \operatorname{clip}\left(r_t^i(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_t^i\right) - \beta D_{\text{KL}}\left(\pi_\theta \| \pi_{\text{ref}}\right)\right)\right]\end{aligned} \tag{4}$$

where the probability ratio $r_t^i(\theta)$ is defined as:

$$r_t^i(\theta) = \frac{p_\theta\left(x_{t-1}^i \mid x_t^i, c\right)}{p_{\theta_{\text{old}}}\left(x_{t-1}^i \mid x_t^i, c\right)} \tag{5}$$

4 Method

4.1 Overview

MaskSense addresses a fundamental mismatch in applying GRPO to masked generative models: GRPO requires diverse rollout groups, whereas confidence-based decoding in MaskGIT prematurely commits to low-uncertainty trajectories, suppressing exploration. Meanwhile, applying GRPO uniformly across all decoding steps ignores their temporally uneven contribution to generation quality. MaskSense resolves this tension via two core modules, as shown in Fig. 2. **Semantic Anchored Routing Sampling** (SARS) introduces a routing strategy that integrates a novel semantic anchored sampler (SA-Sampler) alongside MaskGIT to balance exploration and exploitation, while **Global Entropy Transition Anchoring** (GETA) identifies the decoding timesteps with the largest entropy transition as optimization anchors, focusing GRPO updates where token commitment is most decisive.

4.2 Semantic Anchored Routing Sampling

SARS targets the visual exploration trap, in which confidence-driven decoding preferentially resolves low-uncertainty background regions while deferring commitment on semantically decisive subject tokens. To counteract this bias, SARS derives a prompt-conditioned semantic prior, injects it into the early token retention process, and routes rollout groups through a dual-branch structure.

Semantic Prior Construction. We construct the semantic prior from image-to-text cross-attention cached at decoding step t, where $t \in \{0, 1, \ldots, T\}$ denotes the number of completed decoding steps in contrast to t in Sec. 3. Let $\mathbf{F} \in \mathbb{R}^{B\times N_{\text{img}}\times N_{\text{txt}}}$ denote the cross-attention map extracted from each transformer block, where B is the batch size, N_{img} and N_{txt} denote the numbers of image and text tokens, respectively. We average over image positions to compute a global response score Φ_j for each text token $j \in \{1, \ldots, N_{\text{txt}}\}$, and select the top-$k$ tokens with the highest scores as semantic anchors $\mathcal{T} \subset \{1, \ldots, N_{\text{txt}}\}$, $|\mathcal{T}| = k$. The response score and importance weight are formulated as:

$$\Phi_j = \frac{1}{N_{\text{img}}}\sum_{u=1}^{N_{\text{img}}}\mathbf{F}_{u,j}, \quad w_n = \frac{\Phi_n}{\sum_{n\in\mathcal{T}}\Phi_n} \tag{6}$$

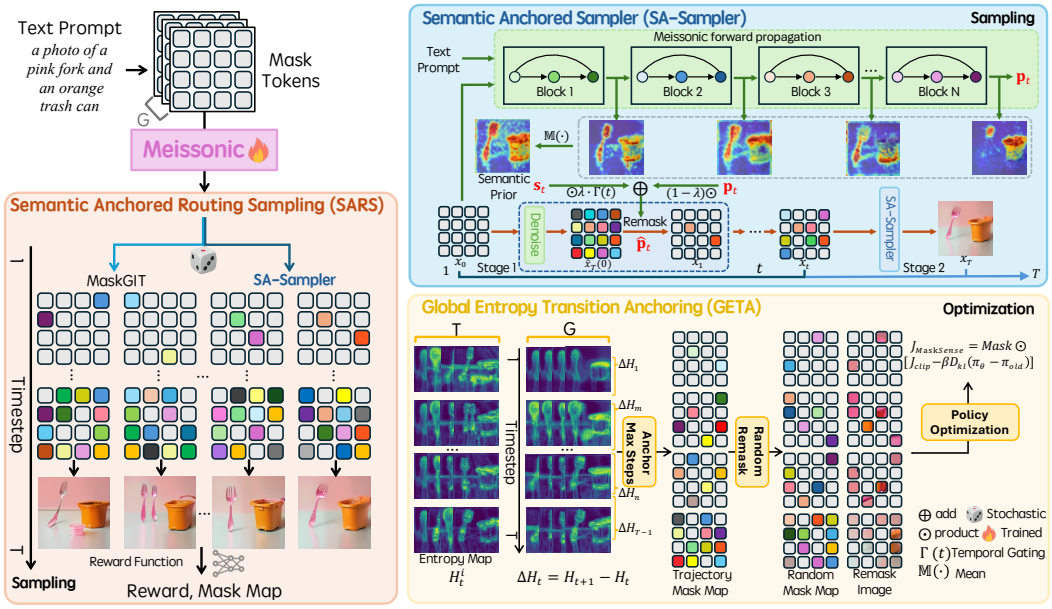


Figure 2: **Overview of MaskSense.** MaskSense comprises two core modules that jointly address the visual exploration trap. At the sampling stage, SARS promotes subject-token exploration via semantic anchored decoding and sharpens advantage estimation through dual-branch rollout routing. At the optimization stage, GETA concentrates GRPO updates on critical decoding steps anchored by entropy transitions, with random remasking applied to ensure optimization diversity.

where Φ_j quantifies the degree to which text token j is attended to across all image positions, and w_n denotes the normalized importance weight of anchor token n. The block-level semantic map $\mathbf{s}_b \in \mathbb{R}^{B\times N_{\text{img}}}$ and the final semantic prior $\mathbf{s}_t \in [0,1]^{B\times N_{\text{img}}}$ are subsequently computed as:

$$\mathbf{s}_b = \sum_{n\in\mathcal{T}} w_n \, \mathbf{F}_n, \quad \mathbf{s}_t = \Psi\left(\frac{1}{N_b}\sum_{b=1}^{N_b} \mathbf{s}_b\right) \tag{7}$$

where $\mathbf{F}_n \in \mathbb{R}^{B\times N_{\text{img}}}$ denotes the attention slice of text token n along the text token dimension of $\mathbf{F}$, N_b is the number of transformer blocks, and $\Psi(\cdot)$ denotes percentile normalization onto $[0,1]$.

Semantic Anchored Sampler (SA-Sampler). Since early steps determine the global structure of the generated image while late-stage injection risks disrupting fine-grained detail refinement (Fig. 1 (b)), we confine semantic prior guidance to the early decoding stage via a linearly decaying temporal gate:

$$\Gamma(t) = \max\left(0,\, 1 - \frac{r_t}{\tau}\right), \quad r_t = \frac{t}{T} \tag{8}$$

where r_t denotes the decoding progress ratio, T is the total number of decoding steps, and $\tau \in (0,1]$ defines the boundary of the early decoding stage within which semantic anchoring is active.

At each step t, the model predicts token probabilities $\mathbf{p}_t$ as confidence scores for all masked positions in parallel, and attains the projected terminal state $\hat{x}_T(t)$. Rather than modifying the predictor, we reweight this score with the semantic prior to obtain the semantic anchored retention score $\hat{\mathbf{p}}_t$:

$$\hat{\mathbf{p}}_t = (1-\lambda)\,\mathbf{p}_t + \lambda\,\Gamma(t)\,\mathbf{s}_t \tag{9}$$

where λ controls the strength of semantic guidance. The reweighted score $\hat{\mathbf{p}}_t$ is used exclusively to determine which tokens are retained as x_{t+1} and which are remasked.

Dual-Branch Routing. To balance exploration and exploitation within each GRPO rollout group, we stochastically partition the G sampled trajectories for each prompt into a semantic branch $\mathcal{G}_{\text{sa}}$ (SA-Sampler) and a standard branch $\mathcal{G}_{\text{mg}}$ (MaskGIT sampler), with $|\mathcal{G}_{\text{sa}}| = |\mathcal{G}_{\text{mg}}| = \frac{G}{2}$ and branch assignments fixed throughout the entire decoding process. By coupling exploratory and conservative trajectories within the same group, this design diversifies sampling outcomes, amplifies intra-group reward contrast, and yields more discriminative advantage estimates for GRPO.

4.3 Global Entropy Transition Anchoring

Even with improved rollout diversity, denoising steps contribute unevenly to policy learning: some involve only local token refinements, while others drive sharp collective uncertainty shifts that fundamentally reshape the generation trajectory, manifesting as large entropy transitions across denoising steps. GETA exploits this signal to concentrate GRPO updates on critical steps.

Entropy Transition Anchor Selection. GETA measures the magnitude of change in the model's global uncertainty across decoding steps. For each rollout, we record the token distribution at every timestep t and compute the group-averaged entropy, formulated as:

$$H_t = -\frac{1}{G}\sum_{i=1}^{G}\sum_{v\in\mathcal{V}} p_t^i(v)\log p_t^i(v) \tag{10}$$

where $\mathcal{V}$ is the codebook and $p_t^i(v)$ is the predicted probability of token v at step t in the i-th rollout.

The entropy transition magnitude between consecutive steps is then computed, and the top-k steps with the largest transitions are selected as anchor steps:

$$\Delta H_t = |H_{t+1} - H_t|, \quad \mathcal{K} = \mathrm{topk}\big(\{\Delta H_t\}_{t=1}^{T-1}, k\big) \tag{11}$$

where $\mathcal{K}$ denotes the set of selected anchor steps and k is the number of anchor steps to select.

Random Remasking for Optimization. Once the anchor set $\mathcal{K}$ is obtained, we construct optimization states via random remasking [70]. Let M_t^{traj} denote the binary mask at anchor step t in the original rollout. Rather than replaying this, we sample a random mask M_t^{rand} with the same mask budget:

$$\big|M_t^{\mathrm{rand}}\big| = \Big|M_t^{\mathrm{traj}}\Big|, \quad M_t^{\mathrm{rand}} \sim \mathcal{U}\Big(\{M\in\{0,1\}^{H\times W} : |M| = \Big|M_t^{\mathrm{traj}}\Big|\}\Big), \tag{12}$$

where $\mathcal{U}(\cdot)$ denotes the uniform distribution, $H \times W$ is the spatial resolution of the latent token map. This random mask is then applied to the final generated latent x_T to produce the remasked input:

$$\tilde{x}_t = M_t^{\mathrm{rand}} \odot [\texttt{MASK}] + \big(1 - M_t^{\mathrm{rand}}\big) \odot x_T, \tag{13}$$

where $\odot$ denotes element-wise product. By decoupling optimization from the exact trajectory mask pattern while preserving the number of masked positions, the policy is trained around each critical transition without overfitting to any single realized masking configuration.

The remasked input $\tilde{x}_t$ is subsequently used to compute the importance ratio $r_t^i(\theta)$ in Eq. 5, and the GRPO objective (Eq. 4) is optimized exclusively over the selected anchor steps $t \in \mathcal{K}$:

$$\mathcal{J}_{\text{MaskSense}}(\theta) = M_t^{\mathrm{rand}} \odot \mathcal{J}_{\text{GRPO}}(\theta)\big|_{t\in\mathcal{K}}, \tag{14}$$

where the mask M_t^{rand} restricts gradient updates to the masked positions of the remasked input $\tilde{x}_t$, ensuring that only the tokens requiring prediction contribute to the optimization objective. The complete MaskSense algorithm is detailed in Algorithm 1 in appendix.

5 Experiments

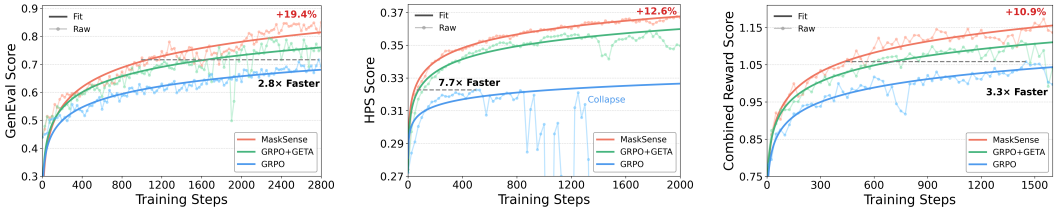
5.1 Experimental Setup

Training Settings. All experiments are conducted on Meissonic [1], a masked text-to-image model at 1024×1024 resolution, trained on 8 NVIDIA A100 GPUs. Following the standard masked-generation RL setup [70], we adopt a classifier-free guidance scale of 5 during both training and inference, perform 64 iterative decoding steps, and set the number of selected optimization anchors to 6. For the compositional image generation task [12], we use 50K GenEval-style prompts sourced from Flow-GRPO [26]. For preference alignment, we use 10K training samples from HPS v2 dataset [61]. More implementation details are provided in Appendix A.

Evaluation Details. We evaluate MaskSense from two complementary perspectives: (i) *task-specific accuracy* on in-distribution test sets, which directly measures the model's ability to handle the target task. (ii) *general image quality* on DrawBench [40], which assesses broader generalization beyond the training distribution. The latter is quantified using a comprehensive set of metrics, including image quality (DeQA [69]) and human preference scores (PickScore [22], ImageReward [65]).

Table 1: **Quantitative Comparison on the GenEval.** Best scores are **bolded**. Results for other models are from [70]. Δ indicates improvements over Meissonic. Obj.: Object; Attr.: Attribution.

Model	Overall↑	Single Obj.↑	Two Obj.↑	Counting↑	Color↑	Position↑	Color Attr.↑
Diffusion-based Models							
LDM [39]	0.37	0.92	0.29	0.23	0.70	0.02	0.05
SD-XL [35]	0.55	0.98	0.74	0.39	0.85	0.15	0.23
SD3.5-M [10]	0.63	0.98	0.78	0.50	0.81	0.24	0.52
DALLE-3 [3]	0.67	0.96	0.87	0.47	0.83	0.43	0.45
FLUX.1-dev [23]	0.66	0.98	0.81	0.74	0.79	0.22	0.45
AR-based Models							
LlamaGen [49]	0.32	0.71	0.34	0.21	0.58	0.07	0.04
JanusFlow [32]	0.63	0.97	0.59	0.45	0.83	0.53	0.42
Janus-Pro-1B [7]	0.73	0.98	0.82	0.51	**0.89**	0.65	0.56
Emu3(+Rewriter) [56]	0.66	0.99	0.81	0.42	0.80	0.49	0.45
Masked Generative Models							
Show-o [63]	0.68	0.98	0.80	0.66	0.84	0.31	0.50
Show-o+PARM [14]	0.69	0.97	0.75	0.60	0.83	0.54	0.53
Mask-GRPO [29]	0.73	0.99	0.90	0.69	0.85	0.35	0.59
Meissonic [1]	0.54	0.99	0.66	0.42	0.86	0.10	0.22
+ GRPO [46]	0.72	0.98	0.84	0.76	0.85	0.45	0.41
+ MaskGRPO [30]	0.73	**1.00**	0.87	0.83	0.87	0.39	0.48
+ MaskFocus [70]	0.76	**1.00**	0.91	0.85	0.87	0.42	0.54
+ **MaskSense (Ours)**	**0.86**	0.99	**0.93**	**0.86**	**0.89**	**0.86**	**0.65**
Δ Over Base Model	+0.32	+0.00	+0.27	+0.44	+0.03	+0.76	+0.43



(a) Compositional Image Gen. (b) Human Preference Alignment (c) Multi-Reward Optimization

Figure 3: **Comparison of Training Dynamics Curves.** Fig.(a) to (c) refer to the tasks of compositional image generation, human preference alignment, and multi-reward optimization, respectively.

5.2 Main Results

Quantitative Comparison. Tab. 1 and 2 demonstrate that MaskSense achieves superior performance across both tasks, improving GenEval by 0.32 and HPS by 7.50 over the Meissonic baseline. Furthermore, these gains generalize beyond the training reward domain, as MaskSense achieves the best scores across DrawBench metrics, confirming broader generalization. As shown in Fig. 3, MaskSense exhibits superior training stability and faster convergence to a higher final performance.

Qualitative Comparison. Fig. 4 provides qualitative evidence consistent with the quantitative results. MaskSense produces richer fine-grained details, complex textures, and higher visual fidelity for preference-oriented prompts, while other methods yield flatter appearance and incomplete stylistic rendering. For compositional prompts, MaskSense more reliably satisfies object-level constraints including count, spatial arrangement, and color attribution, demonstrating significant advantages of MaskSense in aligning with target preferences. See Appendix C for more qualitative comparisons.

5.3 Analysis

Ablation Study. Tab. 3 shows that removing either SARS or GETA consistently degrades performance across all metrics, confirming both components are indispensable to MaskSense. As further evidenced by Fig. 3, the full model achieves more stable training dynamics and faster convergence throughout optimization, ultimately reaching a stronger performance.

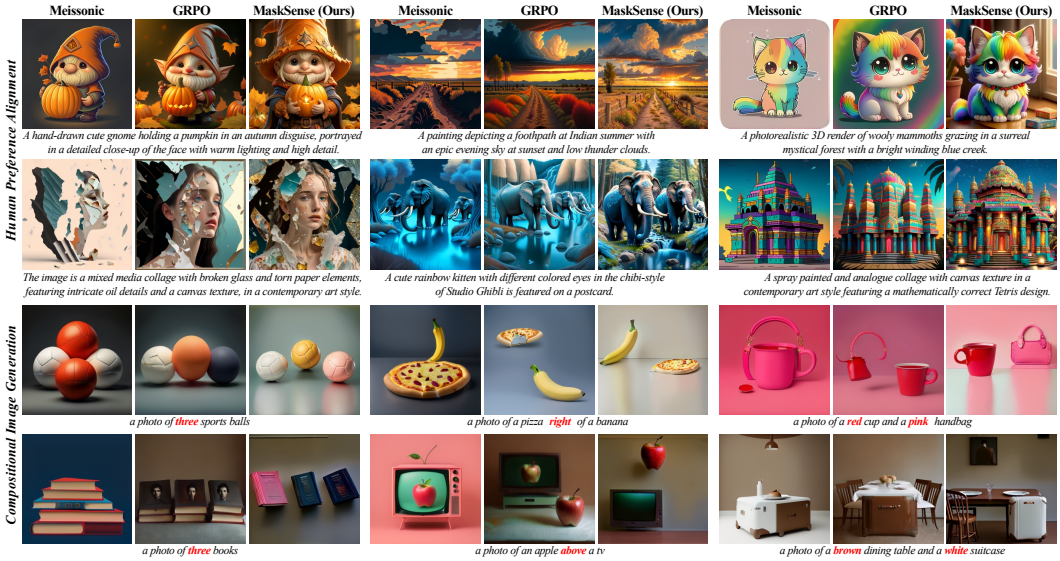


Figure 4: **Qualitative Comparison.** MaskSense achieves superior visual quality with richer fine-grained details and more faithful satisfaction of object-level constraints across both tasks.

Table 2: **Quantitative Comparison on the Human Preference Alignment Benchmark.**

Model	HPS↑	DEQA↑	PicS.↑	ImgR.↑
SD3.5-M [10]	30.17	4.24	22.50	0.98
FLUX.1-dev [23]	31.13	4.37	**22.97**	1.06
Meissonic [1]	28.89	4.00	21.63	0.39
+ GRPO [46]	32.31	4.15	21.96	0.93
+ MaskGRPO [30]	35.48	4.35	22.34	1.06
+ MaskFocus [70]	35.52	4.39	22.39	1.09
+ **MaskSense (Ours)**	**36.39**	**4.42**	22.54	**1.13**
Δ Over Base Model	+7.50	+0.42	+0.91	+0.74

Table 3: **Core Ablations.** PicS.: PickScore; ImgR.: ImageReward.

Model	GenEval↑	HPS↑	DEQA↑	PicS.↑	ImgR.↑
Meissonic [1]	0.54	28.89	4.00	21.63	0.39
+ GRPO [46]	0.72	32.31	4.15	21.96	0.93
w/o GETA	0.79	32.62	4.20	22.09	0.95
w/o SARS	0.81	35.64	4.35	22.19	1.06
MaskSense	**0.86**	**36.39**	**4.42**	**22.54**	**1.13**

Table 4: **Hyper-parameter Study**. We investigate the influence of the semantic strength λ, the temporal gate τ in the SARS, and the number of anchor steps k in the GETA, respectively.

(a) Semantic Strength λ

Metric	0	0.3	**0.5**	0.7	1
HPS	35.64	36.16	36.39	36.20	36.00
DEQA	4.35	4.41	4.42	4.42	4.37
PickScore	22.19	22.42	22.54	22.38	22.35
ImageReward	1.06	1.12	1.13	1.04	1.06

(b) Temporal Gate τ

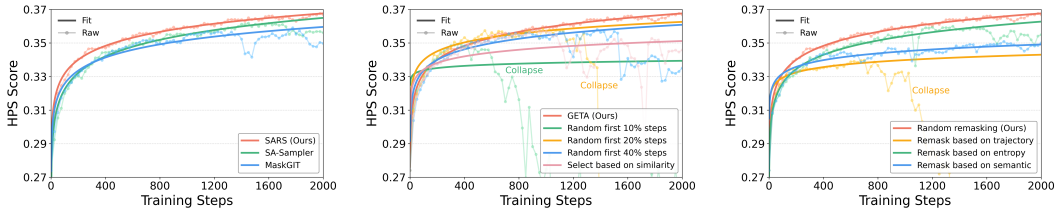
0	0.3	**0.5**	0.7	1
35.64	35.88	36.39	36.11	35.78
4.35	4.39	4.42	4.37	4.37
22.19	22.38	22.54	22.34	22.20
1.06	1.11	1.13	1.07	1.04

(c) Anchor Number k

4	**6**	8
35.79	36.39	36.18
4.36	4.42	4.39
22.37	22.54	22.50
1.07	1.13	1.11

Effect of Semantic Anchored Routing Sampling. We compare SARS against MaskGIT and SA-Sampler alone in Fig. 5 (a). SA-Sampler alone already improves over MaskGIT, confirming that anchoring subject tokens during early decoding stage effectively alleviates the visual exploration trap. However, the full SARS design brings a further and more stable gain, as semantic bias alone is insufficient without the dual-branch routing to mix exploratory and conservative trajectories, which amplifies reward contrast and sharpens advantage estimates. Moreover, Tab. 4 (a) and (b) show that both semantic guidance strength λ and temporal gate τ achieve optimal performance at 0.5.

Effect of Cross-Branch Reward Contrast in Dual-Branch Routing. Jointly normalizing MaskGIT and SA-Sampler trajectories can introduce a branch-dependent component, motivating us to distinguish cross-branch contrast from trajectory-level quality. Let $b_i \in \{\mathrm{MG}, \mathrm{SA}\}$ denote the branch of rollout i, and let μ_b and μ be the branch-wise and pooled reward means. The centered reward decomposes as $R_i - \mu = (R_i - \mu_{b_i}) + (\mu_{b_i} - \mu)$, where the two terms capture within-branch trajectory quality and cross-branch reward contrast, respectively. We remove the latter mean-offset term using branch-conditioned normalization within each $G = 8$ rollout group, which enforces



(a) Effect of Sampling Strategy (b) Effect of Anchoring Strategy (c) Effect of Remasking Strategy

Figure 5: **Ablation Analysis on Key Strategy Choices.** (a) SARS effectively balances exploration and exploitation, outperforming single-branch sampling strategies. (b) GETA anchors critical decision steps, enabling more principled policy optimization than heuristic alternatives. (c) Random remasking diversifies anchor states, producing more stable optimization.

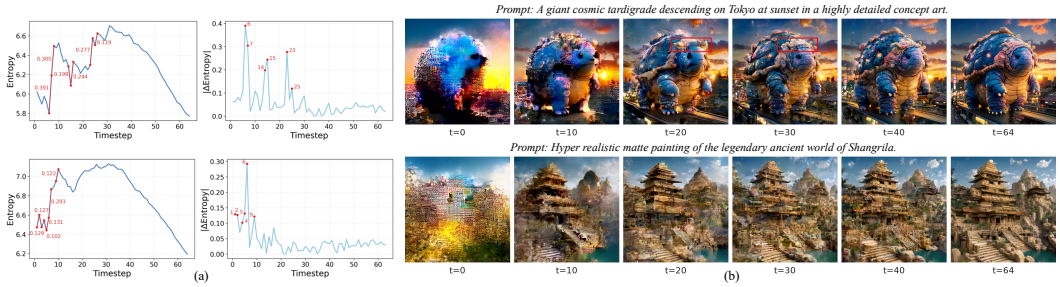


Figure 6: **Anchored Timesteps of Diverse Prompts.** (a) Global entropy and absolute entropy difference between consecutive timesteps. (b) Visualization of generation process across timesteps.

$\mathbb{E}\left[\hat{A}_i \mid b_i = \mathrm{SA}\right] = \mathbb{E}\left[\hat{A}_i \mid b_i = \mathrm{MG}\right] = 0$. Dual-branch training still outperforms MaskGIT-only + GETA (36.08 vs. 35.64 HPS), while pooled normalization reaches 36.39, showing that cross-branch contrast contributes to, but does not account for, the overall gain. A 300-prompt overlap analysis further supports this conclusion: the branch reward ranges overlap for 90.7% of prompts, with 26.9% of SA-Sampler rollouts below the pooled mean and 25.7% of MaskGIT rollouts above it. Thus, sampler identity does not determine the advantage sign. Since all variants use MaskGIT for evaluation, the observed gain stems from the learned policy itself.

Effect of Global Entropy Transition Anchoring. We compare GETA against random early-step selection and similarity-based selection [70] in Fig. 5 (b). Although restricting anchors to the first 10% or 20% of steps brings faster initial gains, it ultimately leads to training collapse, possibly due to highly masked states inducing unstable optimization targets. Similarity-based selection appears more stable initially but exhibits increasing instability in later stages, as embedding similarity captures visual representational change but fails to align with the model's most decisive policy decision points. As shown in Fig. 6, the top row exhibits dominant transitions spread across the first 30 steps where each committed decision critically shapes the overall image structure, while the bottom row concentrates transitions within the first 10 steps with later steps contributing only minor detail refinement. This prompt-adaptive distribution of critical steps validates the rationality of GETA's entropy-guided anchor selection. Besides, Tab. 4 (c) demonstrates that $k = 6$ achieves the best trade-off between optimization focus and trajectory coverage under the same GPU hours.

Effect of Random Remasking. We conduct a systematic study of remasking strategies for anchor-state reconstruction in Fig. 5 (c), as masked positions directly affect anchor-state probability estimation. Random remasking achieves the best convergence and stability, while trajectory-based remasking leads to early collapse. Entropy-guided and semantics-guided remasking remain trainable yet consistently underperform, indicating that randomizing masked positions while preserving the mask budget produces greater anchor-state diversity and more robust optimization.

Discussion of Reward Hacking. Beyond task-specific metrics, we assess reward hacking via the DrawBench [40] metrics reported in Tab. 2. MaskSense demonstrates outstanding alignment capability, indicating that the model generalizes well beyond the training reward without overfitting to optimized metrics. Furthermore, we present the joint multi-reward training curves of GenEval and

Table 5: **Transfer to Show-o on HPS.** Evaluation HPS scores and cumulative training GPU hours.

Step	0	200	400	600	800	1K	1.2K	1.4K	1.6K	1.8K	2K
Mask-GRPO HPS	26.54	27.28	28.09	28.54	29.05	29.44	29.56	29.72	29.74	29.72	29.81
MaskSense HPS	26.54	27.60	28.22	28.92	29.30	29.66	29.82	29.98	29.97	30.05	30.09
Mask-GRPO GPU Hours	0.00	11.10	22.17	33.21	44.30	55.37	66.44	77.51	88.61	99.66	110.72
MaskSense GPU Hours	0.00	11.45	22.89	34.31	45.73	57.15	68.56	79.96	91.37	102.77	114.18

Table 6: **VQA Evaluation on GenEval.** Results cover all prompts with four images per prompt.

Model	**Overall↑**	**Single Obj.↑**	**Two Obj.↑**	**Counting↑**	**Color↑**	**Position↑**	**Color Attr.↑**
Meissonic	0.7006	0.9750	0.8030	0.5687	0.9149	0.6400	0.5925
+ GRPO	0.8616	**0.9875**	0.9343	0.9109	0.9335	0.8233	0.7700
+ **MaskSense**	**0.9189**	0.9844	**0.9432**	**0.9234**	**0.9362**	**0.9567**	**0.8662**
Δ Over Base Model	+0.2183	+0.0094	+0.1402	+0.3547	+0.0213	+0.3167	+0.2737

Table 7: **Human Preference Evaluation on HPS.** Image order is randomized in each comparison.

Comparison	**Dimension**	**Opponent Win**	**MaskSense Win**	**Tie**
Meissonic vs. MaskSense	Aesthetic	28 (7%)	**348 (87%)**	24 (6%)
Meissonic vs. MaskSense	Semantic	24 (6%)	**92 (23%)**	284 (71%)
GRPO vs. MaskSense	Aesthetic	84 (21%)	**216 (54%)**	100 (25%)
GRPO vs. MaskSense	Semantic	32 (8%)	**60 (15%)**	308 (77%)

CLIP [36] rewards in Fig. 3 (c), which remain consistently stable with superior performance throughout training. These advancements demonstrate the strong robustness of the proposed MaskSense.

Model Transferability and Independent Evaluations. *Transfer to Show-o.* We integrate SARS and GETA into the official Mask-GRPO implementation on Show-o [63] while retaining its training configuration. As shown in Tab. 5, MaskSense consistently outperforms Mask-GRPO and reaches 30.09 HPS at 2K steps, versus 29.81 for Mask-GRPO and 26.54 for the pretrained model, with only 3.1% additional GPU hours. This demonstrates transfer beyond Meissonic [1] with modest overhead.

Independent Evaluations. Using Qwen-VL-Plus [2] as an independent VQA evaluator, we assess object existence, counting, color, and spatial relations on the GenEval test set. Tab. 6 shows that MaskSense achieves 91.89% overall accuracy, outperforming GRPO (86.16%) and Meissonic (70.06%). We further conduct human pairwise comparisons on HPS prompts. As shown in Tab. 7, MaskSense is aesthetically preferred over Meissonic and GRPO in 87% and 54% of cases, respectively, while also leading in semantic alignment despite frequent ties. These reward-independent evaluations confirm improvements in both compositional faithfulness and perceptual quality.

6 Conclusion

We present MaskSense to address the visual exploration trap in masked generative models, where confidence-based sampling prematurely suppresses subject-token diversity and undermines the effectiveness of GRPO optimization. By anchoring subject tokens during early decoding via semantic priors, MaskSense preserves the high-entropy exploration space that confidence-based sampling would otherwise collapse, enabling sharper intra-group advantage estimation through dual-branch rollout routing. Building on the enriched rollout diversity, MaskSense further identifies timesteps of maximal entropy transition as optimization anchors to address the uneven contribution of individual steps to policy learning, concentrating gradient updates where token commitment is most decisive. Extensive experiments demonstrate the gains achieved by MaskSense and validate the critical role of confronting the visual exploration trap for RL optimization in masked generative models.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants Grants 62225207, 62436008, 62422609, 62306295, and 62576328, in part by the CAS Project for Young Scientists in Basic Research under Grant YSBR-122, and in part by the Fundamental Research Funds for the Central Universities under Grant WK2102026001.

References

[1] 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*, 2024.

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

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[4] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

[5] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

[6] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11315–11325, 2022.

[7] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

[8] Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

[9] Ziyun Dai, Xiaoqiang Li, Shaohua Zhang, Yuanchen Wu, and Jide Li. See different, think better: Visual variations mitigating hallucinations in lvlms. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 3310–3319. ACM, October 2025.

[10] 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.

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, June 2021.

[12] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.

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

[14] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Rui Huang, Haoquan Zhang, Manyuan Zhang, Jiaming Liu, Shanghang Zhang, Peng Gao, et al. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025.

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[17] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[18] Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot, 2025.

[19] Zhehan Kan, Xinghua Jiang, Yanlin Liu, Xiaochen Yang, Zhixiang Wei, Shifeng Liu, Yubo Zhu, Qingmin Liao, Wenming Yang, Xin Li, et al. Uvu: Improving multimodal understanding via vision-language unified autoregressive paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26230–26239, 2026.

[20] Zhehan Kan, Yanlin Liu, Kun Yin, Xinghua Jiang, Xin Li, Haoyu Cao, Yinsong Liu, Deqiang Jiang, Xing Sun, Qingmin Liao, et al. Taco: Think-answer consistency for optimized long-chain reasoning and efficient data learning via reinforcement learning in lvlms. *arXiv preprint arXiv:2505.20777*, 2025.

[21] Zhehan Kan, Yubo Zhu, Xinghua Jiang, Zhixiang Wei, Shifeng Liu, Wei Tong, Sheng Zhong, Qingmin Liao, Wenming Yang, Xin Li, et al. Vivas: Vitalizing visual perception in vlm pre-training via vision-language unified autoregressive supervision. In *European Conference on Computer Vision*, pages 321–338. Springer, 2026.

[22] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

[23] Black Forest Labs. https://github.com/black-forest-labs/flux, 2024.

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

[25] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yi Xin, Xinyue Li, Qi Qin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024.

[26] 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. *arXiv preprint arXiv:2505.05470*, 2025.

[27] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22(4):730–751, 2025.

[28] Fan Lu, Wei Wu, Kecheng Zheng, Shuailei Ma, Biao Gong, Jiawei Liu, Wei Zhai, Yang Cao, Yujun Shen, and Zheng-Jun Zha. Benchmarking large vision-language models via directed scene graph for comprehensive image captioning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19618–19627. IEEE, 2025.

[29] Yifu Luo, Xinhao Hu, Keyu Fan, Haoyuan Sun, Zeyu Chen, Bo Xia, Tiantian Zhang, Yongzhe Chang, and Xueqian Wang. Reinforcement learning meets masked generative models: Mask-grpo for text-to-image generation. *arXiv preprint arXiv:2510.13418*, 2025.

[30] Tianren Ma, Mu Zhang, Yibing Wang, and Qixiang Ye. Consolidating reinforcement learning for multimodal discrete diffusion models. *arXiv preprint arXiv:2510.02880*, 2025.

[31] Xiaoxiao Ma, Feng Zhao, Pengyang Ling, Haibo Qiu, Zhixiang Wei, Hu Yu, Jie Huang, Zhixiong Zeng, and Lin Ma. Towards better & faster autoregressive image generation: From the perspective of entropy. *arXiv preprint arXiv:2510.09012*, 2025.

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

[33] Zanlin Ni, Shenzhi Wang, Yang Yue, Tianyu Yu, Weilin Zhao, Yeguo Hua, Tianyi Chen, Jun Song, Cheng Yu, Bo Zheng, et al. The flexibility trap: Rethinking the value of arbitrary order in diffusion language models. *arXiv preprint arXiv:2601.15165*, 2026.

[34] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4055–4064. PMLR, 10–15 Jul 2018.

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

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[37] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[38] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[41] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[42] Yawen Shao, Jie Xiao, Kai Zhu, Yu Liu, Hongchen Luo, Xueyang Fu, Yang Cao, Wei Zhai, and Zheng-Jun Zha. Back on track: Aligning rewards and states for reasoning in diffusion large language models. In Hung-Yi Lee, Sunipa Dev, Jessy Li, and Desmond Elliott, editors, *Findings of the Association for Computational Linguistics: EMNLP 2026*, pages 6874–6890, Budapest, Hungary, October 2026. Association for Computational Linguistics.

[43] Yawen Shao, Jie Xiao, Kai Zhu, Yu Liu, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Anchoring values in temporal and group dimensions for flow matching model alignment. *arXiv preprint arXiv:2512.12387*, 2025.

[44] Yawen Shao, Wei Zhai, Yuhang Yang, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Great: Geometry-intention collaborative inference for open-vocabulary 3d object affordance grounding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17326–17336. IEEE, 2025.

[45] Yawen Shao, Yubo Zhu, Ziyun Dai, Zixun Fang, Kai Zhu, Zeyinzi Jiang, Yufeng Ai, Siyang Sun, Haolan Xue, Yu Shang, et al. Towards unified evaluation of prompt enhancers for video generation. *arXiv preprint arXiv:2610.11736*, 2026.

[46] 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.

[47] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[48] Xinyang Song, Xi Yong, Qi Li, Jianxin Sun, Yunfan Liu, and Zhenan Sun. Stylevideo: One-shot image-guided video style transfer. *Machine Intelligence Research*, 23(5):1216–1228, 2026.

[49] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

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

[51] Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Xiangyang Liu, Hang Yan, Yunfan Shao, Qiong Tang, Shiduo Zhang, et al. Moss: An open conversational large language model. *Machine Intelligence Research*, 21(5):888–905, 2024.

[52] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

[53] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[54] Wan Team. Wan AI: Leading ai video generation model. `https://wan.video`, 2026.

[55] Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025.

[56] 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.

[57] Zhixiang Wei, Yi Li, Zhehan Kan, Xinghua Jiang, Zuwei Long, Shifeng Liu, Hongze Shen, Wei Liu, Xiaoyu Tan, Haojia Lin, et al. Youtu-vl: Unleashing visual potential via unified vision-language supervision. *arXiv preprint arXiv:2601.19798*, 2026.

[58] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12966–12977, 2025.

[59] Pingyu Wu, Kai Zhu, Yu Liu, Longxiang Tang, Jian Yang, Yansong Peng, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Towards sequence modeling alignment between tokenizer and autoregressive model. In *International Conference on Learning Representations*, volume 2026, pages 132194–132217, 2026.

[60] Pingyu Wu, Kai Zhu, Yu Liu, Liming Zhao, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Improved video vae for latent video diffusion model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18124–18133. IEEE, 2025.

[61] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

[62] Junyuan Xiao, Dingkang Liang, Xin Zhou, Yixuan Ye, Tongtong Su, Guangmo Yi, Bin Xia, Qiang Lyu, Shurui Shi, Jun Huang, et al. Divide and conquer: Decoupled representation alignment for multimodal world models. In *European Conference on Computer Vision*, pages 433–451. Springer, 2026.

[63] 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.

[64] Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025.

[65] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

[66] 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.

[67] Jian Yang, Dacheng Yin, Yizhou Zhou, Fengyun Rao, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Mmar: Towards lossless multi-modal auto-regressive probabilistic modeling. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7974–7985. IEEE, 2025.

[68] Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025.

[69] Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14483–14494, 2025.

[70] Guohui Zhang, Hu Yu, Xiaoxiao Ma, Yaning Pan, Hang Xu, Jie Huang, and Feng Zhao. Maskfocus: Focusing policy optimization on critical steps for masked image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5956–5966, June 2026.

[71] Guohui Zhang, Hu Yu, Xiaoxiao Ma, JingHao Zhang, Yaning Pan, Mingde Yao, Jie Xiao, Linjiang Huang, Jie Huang, and Feng Zhao. Group critical-token policy optimization for autoregressive image generation. In *International Conference on Learning Representations*, volume 2026, pages 51697–51719, 2026.

[72] Siyan Zhao, Devaansh Gupta, Qinqing Zheng, and Aditya Grover. d1: Scaling reasoning in diffusion large language models via reinforcement learning, 2025.

[73] Yubo Zhu, Zhehan Kan, Jingyi Yang, Miaolin Chen, Jinbo Xing, Kai Zhu, Zijian Wang, Sheng Zhong, and Wei Tong. When does visual generation help visual understanding in unified multimodal models? In Hung-Yi Lee, Sunipa Dev, Jessy Li, and Desmond Elliott, editors, *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, pages 8749–8788, Budapest, Hungary, October 2026. Association for Computational Linguistics.

[74] Yubo Zhu, Dongrui Liu, Zecheng Lin, Wei Tong, Sheng Zhong, and Jing Shao. The llm already knows: Estimating llm-perceived question difficulty via hidden representations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1160–1176, 2025.

[75] Yubo Zhu, Yawen Shao, Ziyun Dai, Zixun Fang, Kai Zhu, Siyang Sun, Haolan Xue, Chuxin Wang, Tingyu Weng, Jingming Luo, et al. Wanpe: Towards cinematic prompt enhancement for modern text-to-video generation. *arXiv preprint arXiv:2609.30221*, 2026.

A Details on the Experimental Setup

A.1 Dataset and Reward

Dataset. For the compositional image generation task, 50K prompts are constructed using the official GenEval [12] scripts, following Flow-GRPO [26], covering six sub-tasks with the prompt sampling ratio set to Position : Counting : Attribute Binding : Colors : Two Objects : Single Object = 7 : 5 : 3 : 1 : 1 : 0. The same dataset is also used for multi-reward joint optimization. For human preference alignment, we use 10K training samples from the HPS v2 dataset [61], a large-scale benchmark collecting pairwise human judgments over generated images across diverse prompt styles. We further evaluate out-of-distribution generalization on DrawBench [40], a benchmark comprising 200 prompts across 11 diverse categories that probes broad generative capabilities beyond the training distribution.

Reward. For compositional image generation, we adopt the GenEval reward, which is computed based on rule-based detection of object presence, count, spatial position, and color attribution. For human preference alignment, we employ the HPS [61] reward model. For multi-reward joint optimization, we simultaneously optimize GenEval and CLIP [36] rewards.

A.2 Implementation Details

Training. All experiments are conducted on Meissonic [1], a masked text-to-image model at 1024×1024 resolution, trained on 8 NVIDIA A100 GPUs. We use the AdamW optimizer with a learning rate of 2×10^{-6}. The maximum prompt length and completion length are set to 512 and 1024, respectively, with a sampling temperature of 1.0. For GRPO optimization, we set the group size to $G = 8$, with a per-device batch size of 1 and gradient accumulation steps of 1. Following the standard masked-generation RL setup [70], we adopt a classifier-free guidance scale of 5 during both training and inference, and perform 64 iterative decoding steps. The number of selected optimization anchors is set to $k = 6$, and the KL penalty coefficient is set to $\beta = 0.01$. For SARS, the semantic guidance strength is set to $\lambda = 0.5$ and the temporal gate threshold to $\tau = 0.5$.

Evaluation. For compositional image generation, we evaluate on the GenEval benchmark [12] using its official evaluation pipeline, which detects object bounding boxes and colors and infers their spatial relations. The evaluation set comprises 553 prompts, for each of which we generate 4 images to ensure stable and reliable score estimation. For human preference alignment, we evaluate on a held-out test set of 400 prompts from HPS v2 [61]. For out-of-distribution generalization, we evaluate on DrawBench [40], a benchmark of 200 prompts spanning 11 diverse categories, generating 5 images per prompt to ensure broader coverage and stable generalization assessment. To ensure fairness and stability during evaluation, all models are assessed using the original MaskGIT sampler as the decoding strategy at evaluation time, regardless of the sampling strategy employed during training. The proposed SA-Sampler is exclusively utilized during the RL training phase to promote exploratory diversity in rollout generation and is not applied at evaluation time. All quantitative results reported in the main paper are obtained under the above evaluation protocol. To ensure statistical reliability, all metrics are averaged over multiple generated images per prompt, which reduces variance and yields more stable estimates across different random seeds, ensuring that the reported numbers reflect the model's generative capability rather than individual sampling artifacts.

A.3 Quality Metrics

The details of task accuracy metrics and quality metrics are as follows:

- GenEval [12]: An object-focused evaluation framework that benchmarks T2I models on six compositional image generation tasks. The official pipeline detects object bounding boxes and colors and infers spatial relations between them. Rewards follow rule-based criteria: (i) Counting: $r = 1 - |N_{gen} - N_{ref}|/N_{ref}$; (ii) Position / Color: a partial reward is granted when the object count is correct, with the remainder awarded upon correct prediction of position or color.
- HPS [61]: A scoring model obtained by fine-tuning CLIP [36] on the Human Preference Dataset v2 (HPD v2), a large-scale collection of 798K pairwise human preference annotations over images generated from diverse prompts and models. HPS v2 predicts the

Algorithm 1 MaskSense Training Process

Input: Reward model RM; Prompt dataset $\mathcal{C}$; Decoding steps T; Anchored steps $\mathcal{K}$; Group size G.
Output: Optimized policy model parameters θ

Initialize policy model π_θ
while not converged **do**
  Sample batch of prompts $\mathcal{C}_b \sim \mathcal{C}$
  Update old policy model: $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
  ▷ *// Sampling Stage*
  for each prompt $c \in \mathcal{C}_b$ **do**
    for generating i-th image from $i = 1$ **to** G **do**
      for decoding timestep $t = 0$ **to** $T - 1$ **do**
        Compute semantic prior $\mathbf{s}_t$ in Eq. 7 and retention score $\hat{\mathbf{p}}_t$ in Eq. 9
        Retain tokens to obtain x_{t+1}^i via assigned branch sampler
      end for
      Compute group-averaged entropy H_t in Eq. 10 and advantage $\hat{A}_t^i$ in Eq. 3
    end for
    ▷ *// Optimization Stage*
    Compute entropy transition magnitudes ΔH_t and select anchor set $\mathcal{K}$ in Eq. 11
    for each anchor step $t \in \mathcal{K}$ **do**
      Sample random remasking M_t^{rand} with same budget as M_t^{traj} in Eq. 12
      Construct remasked input $\tilde{x}_t$ from final latent x_T in Eq. 13
      Compute policy optimization objective $\mathcal{J}_{\text{MaskSense}}(\theta)$ in Eq. 14
    end for
  end for
end while

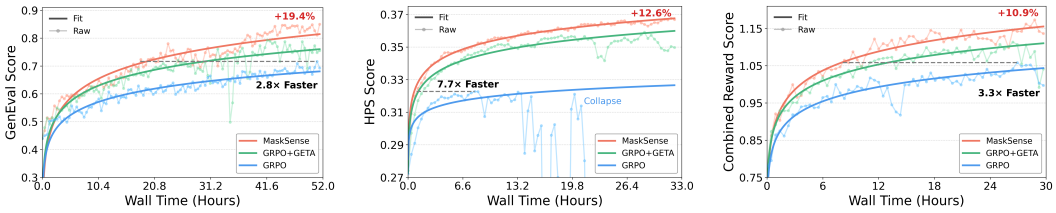
likelihood that a generated image will be preferred by humans, serving as a reliable proxy for aesthetic quality and text-image alignment.

- DeQA [69]: A multimodal large language model-based image quality assessment metric that estimates overall perceptual quality by accounting for distortions, texture degradation, and AI-specific artifacts, trained via a fidelity loss over soft score distributions.
- PickScore [22]: A preference scoring model fine-tuned from CLIP [36] on Pick-a-Pic, a large-scale dataset of real user preference choices between pairs of images generated from identical prompts, providing judgments strongly aligned with human evaluation across text-image alignment and visual aesthetics.
- ImageReward [65]: The first general-purpose reward model for text-to-image generation, trained to capture human preference dimensions encompassing text-image alignment, perceptual fidelity, and harmlessness.

B Limitations and Societal Impacts

Limitations. While our method demonstrates strong performance across multiple benchmarks, several limitations remain. First, the proposed reinforcement learning framework is grounded in the general MaskGIT paradigm, which is adopted by the vast majority of existing masked generative models as their core sampling strategy. It is therefore in principle compatible with a broad family of such models. Validating and extending its application to a wider range of base architectures [68] will be one of the primary focuses of our future work. Second, the current work focuses solely on image generation, leaving the exploration of reinforcement learning for multi-modal discrete diffusion models to future work, including the analysis of sampling and optimization across different modalities toward a unified RL framework for multi-modal masked generative models [64].

Societal Impacts. This work aims to advance text-to-image generation by aligning masked generative models with human preferences and compositional understanding. The proposed framework has the potential to benefit a wide range of creative, educational, and industrial applications. However, improved generative capabilities may also be exploited to produce misleading or harmful visual content, raising concerns about misinformation and malicious use. Additionally, as the model is



(a) Compositional Image Gen. (b) Human Preference Alignment (c) Multi-Reward Optimization

Figure 7: **Comparison of Compute Overhead Dynamics.** Fig.(a) to (c) refer to the tasks of compositional image generation, human preference alignment, and multi-reward optimization, respectively.

trained on large-scale data, it may inadvertently inherit and amplify societal biases present in the training corpus. We encourage responsible deployment of the proposed method and hope to inspire future research on safety, fairness, and interpretability in generative modeling.

C Additional Experiments

Observation Discussion. Fig. 8 and 9 illustrate the sampling trajectories of the MaskGIT sampler and SA-Sampler across diverse prompt categories, including abstract art, natural landscapes, character portraits, architectural scenes, and still-life compositions. A consistent pattern emerges across all categories: the MaskGIT sampler systematically resolves background regions in the early decoding steps while deferring subject tokens to later stages, confirming that the visual exploration trap is a pervasive structural bias rather than a prompt-specific artifact. In contrast, the SA-Sampler, without any modification to the model parameters, redirects early token commitment toward semantically critical subject regions, enabling the discovery of higher-quality generative trajectories through broader subject-space exploration. Fig. 10 further reveals how RL training reshapes the decoding behavior of the model. While the pre-training MaskGIT sampler follows a rigid background-to-subject commitment order, the MaskSense trained model exhibits a markedly more flexible and diverse sampling process, with subject tokens being resolved earlier and more consistently across different prompts. This behavioral shift demonstrates that MaskSense not only improves final image quality but fundamentally restructures the model's generative strategy toward a more diverse and flexible sampling process.

The corresponding prompts in Fig. 8, 9, 10 are: (1) A spray painted and analogue collage with canvas texture in a contemporary art style featuring a mathematically correct Tetris design. (2) A town of pod homes integrated in a forest area with water and trees, depicted in a detailed watercolor by Lurid. (3) An oil painting portrait of a beautiful dryad wearing an ombre velvet gown, with long hair and a tiara, adorned with dozens of jeweled necklaces, and illuminated with dramatic cinematic lighting. (4) A digital painting of a small universe with intricate details and ornate features, featuring the styles of Claude Monet and Vincent van Gogh. (5) A lot of building on each side of the road, with a very curvy road in the middle. (6) A red cobweb is seen inside a marble with an hourglass, lightning and intricate details, creating a sense of awe with swirling mist.

Computational Efficiency. Although MaskSense introduces additional operations for semantic prior extraction in the SA-Sampler and global entropy computation in GETA, the resulting computational overhead is negligible in practice. As shown in Fig. 7, the GPU hours curve of MaskSense remains fully consistent with the training dynamics curves in Fig. 3 of the main paper at the same convergence speed. Taking the human preference alignment task as an example, MaskSense incurs an additional overhead of only 0.064 GPU hours per 100 training steps compared to the baseline. This marginal cost is negligible relative to the substantial performance gains achieved by MaskSense, demonstrating that our method improves generation quality with virtually no additional computational burden. The trend curves are obtained by robust logarithmic fitting $y = a + b\log(1 + x)$ with outlier rejection.

More Visualized Results. To further validate the effectiveness of MaskSense, we provide additional qualitative comparisons on both the human preference alignment task (Fig. 11) and the compositional image generation task (Fig. 12), supplementing the visualizations presented in the main paper.

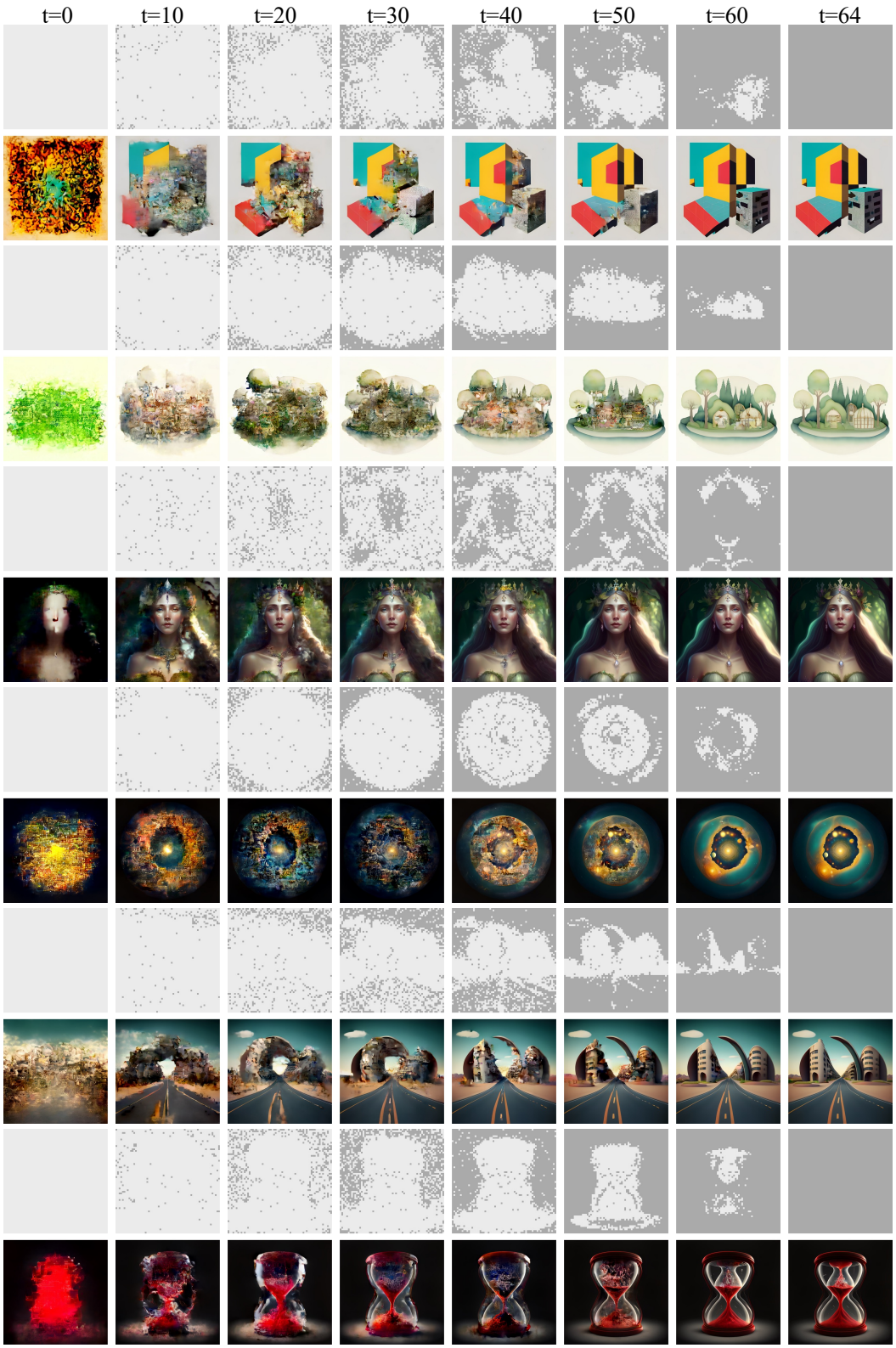


Figure 8: **Sampling Process Visualization of MaskGIT Sampler on Meissonic baseline.**

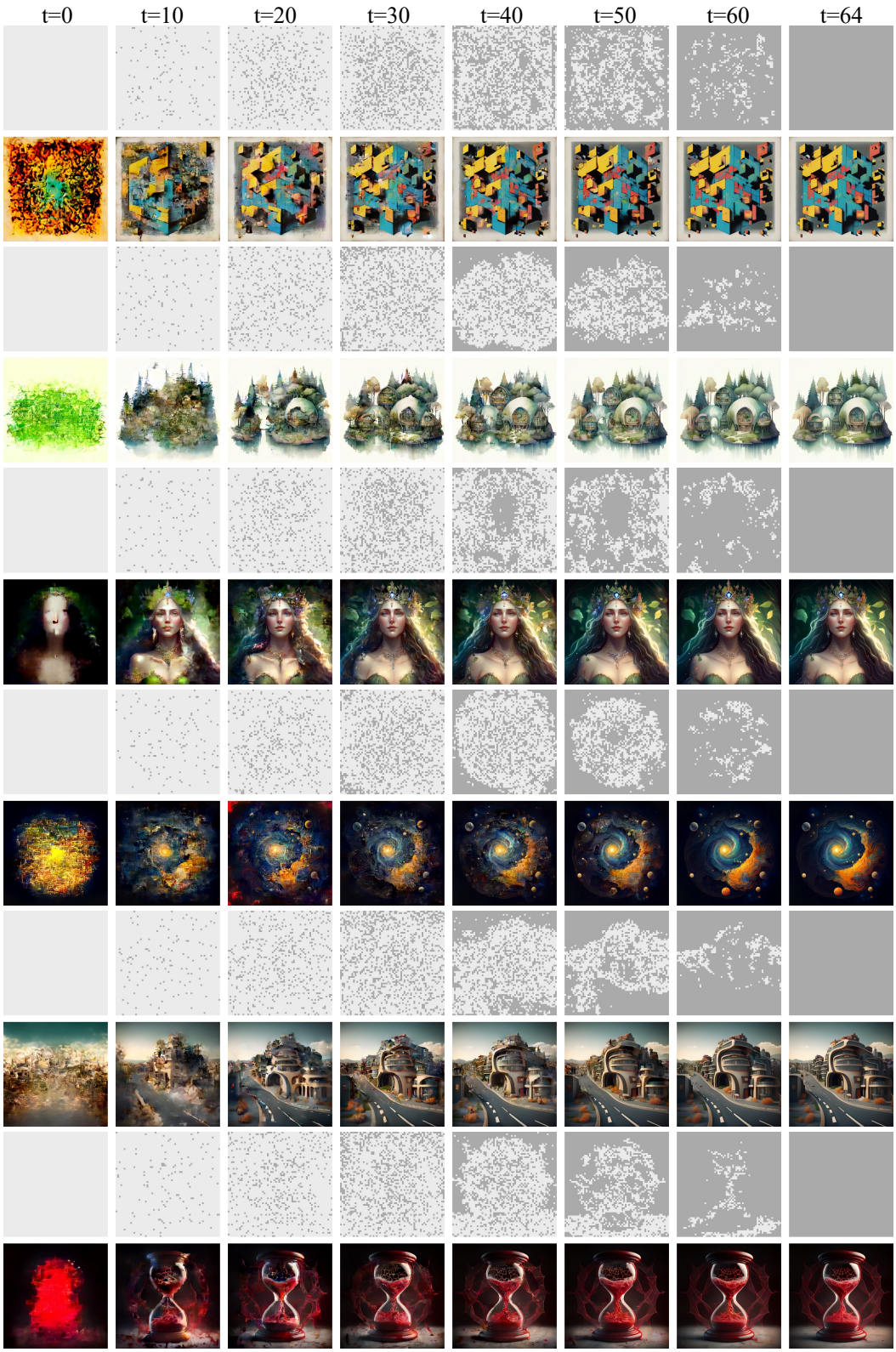


Figure 9: **Sampling Process Visualization of SA-Sampler on Meissonic baseline.**

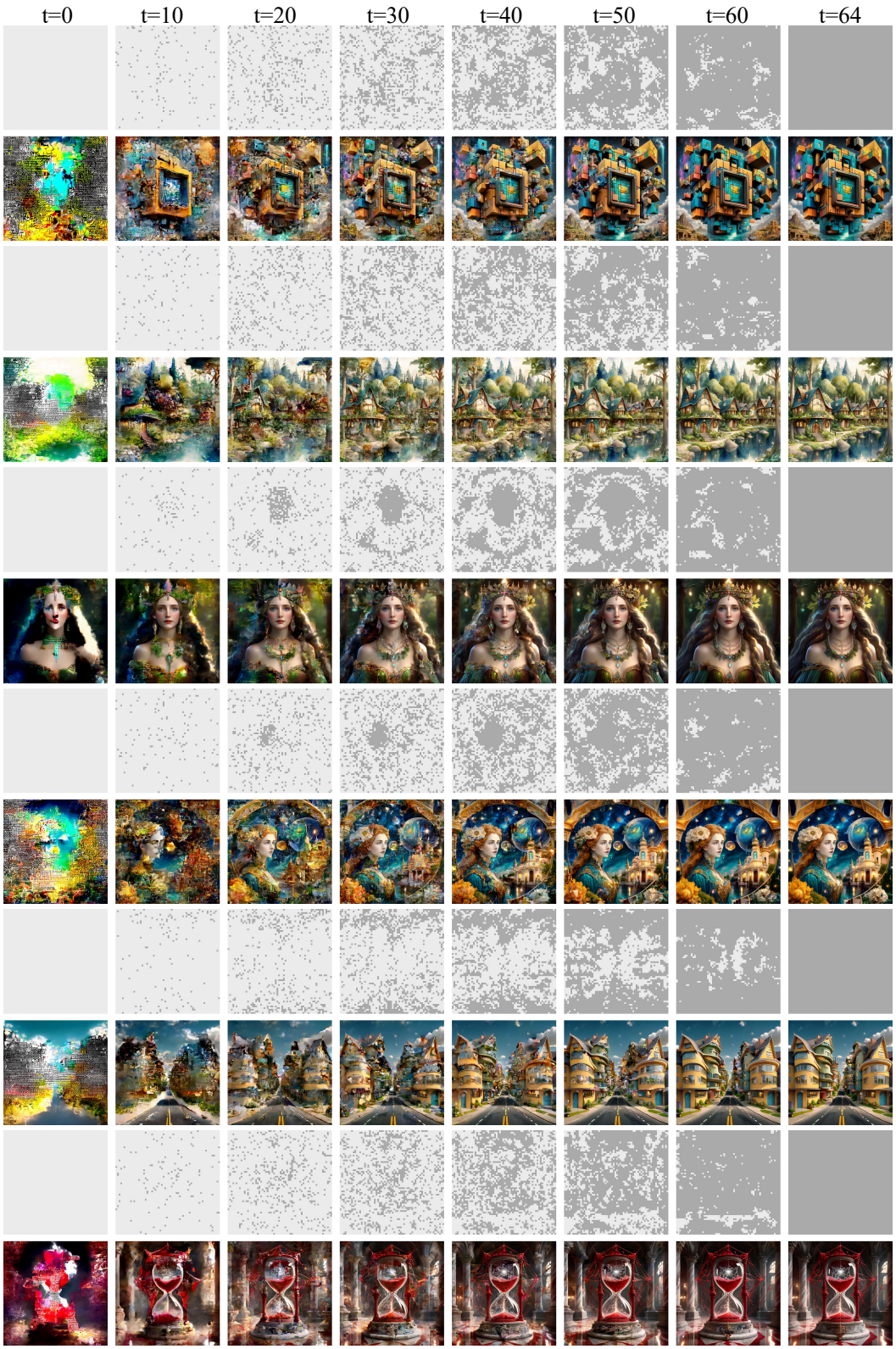


Figure 10: **Sampling Process Visualization of MaskGIT Sampler on MaskSense after RL Training.**

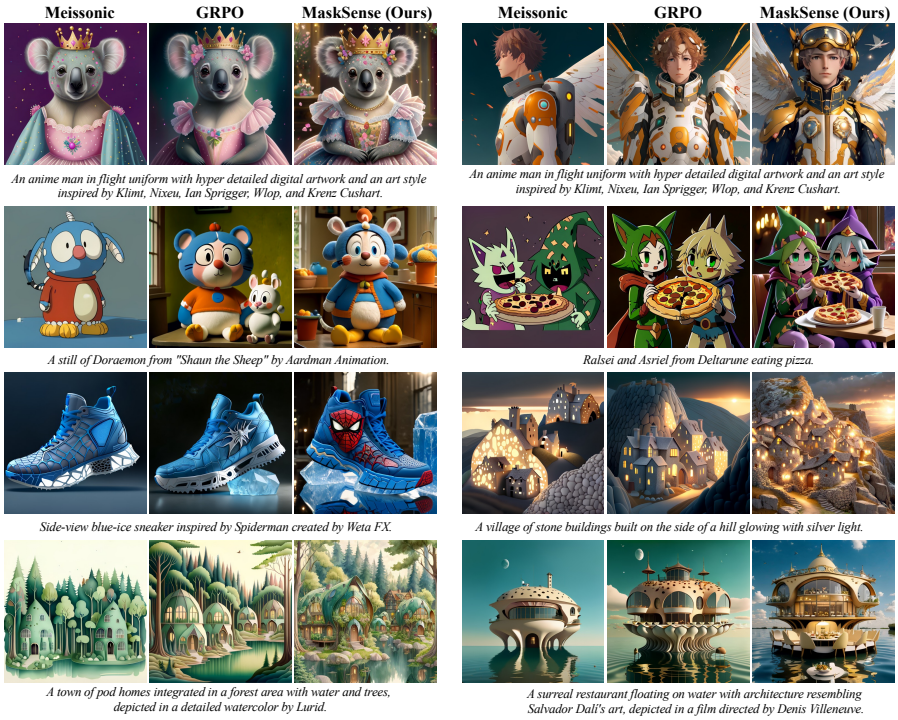


Figure 11: **Qualitative Comparisons on Human Preference Alignment.** MaskSense generates images with richer stylistic details, superior texture fidelity, and more faithful rendering of complex artistic styles compared to Meissonic and GRPO baselines.

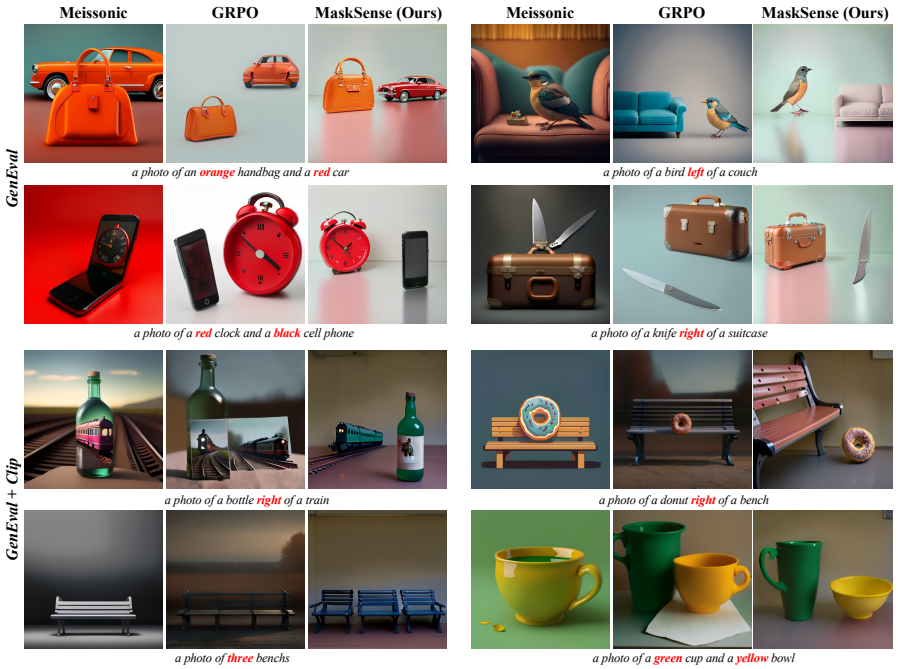


Figure 12: **Qualitative Comparisons on Compositional Image Generation.** MaskSense more reliably satisfies compositional constraints including object count, spatial arrangement, and color attribution across GenEval and GenEval+CLIP settings.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims we made.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in the Sec. B.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide details for our theory.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Our results are reproducible, and we will release the codes and models.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes].

Justification: We will open-source our codes after acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specify all the training and evaluation details in Sec. A.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: This paper ensures statistical reliability by averaging all metrics over multiple generated images per prompt, as detailed in Sec. A.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sufficient information on the computer resources is provided in Sec. A.2 & C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We recognize the importance of ethical standards in research and take this matter seriously.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential societal impacts of our work in Sec. B.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We commit to continuously monitoring the usage of our released models and codes and will take action to restrict access or provide additional guidance if we identify concerning patterns of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the assets we used have been properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: All models will be released upon acceptance.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.