STAR: SPATIOTEMPORAL ADAPTIVE REWARD ALLOCATION FOR TEXT-TO-IMAGE RL POST-TRAINING

Anonymous authors
Paper under double-blind review

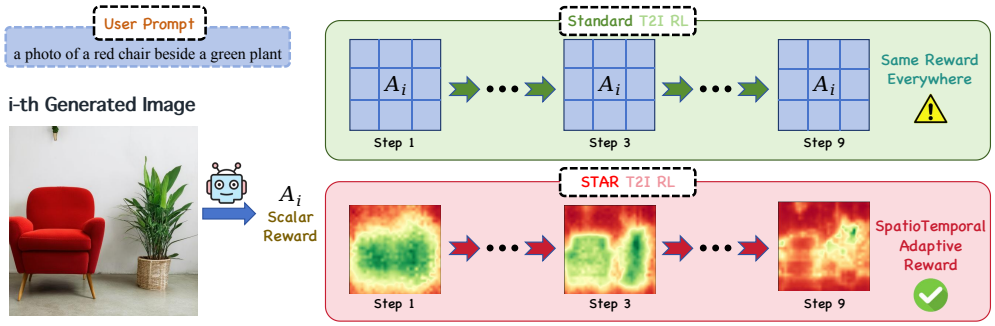


Figure 1: Overview of STAR. Text–image attention converts an image-level group-relative advantage into timestep-specific spatial rewards, focusing policy updates on prompt-relevant latent regions.

ABSTRACT

Existing RL post-training methods for text-to-image generation usually convert the final-image reward into a single scalar advantage and apply it with the same strength to the entire generative trajectory. However, text-to-image generation naturally has temporal and spatial structure: different denoising steps are responsible for different generation stages, and the content that truly determines text alignment often appears only in part of the image. This granularity mismatch makes it difficult for policy updates to focus on the generative components that actually affect the reward. To address this issue, we propose **SpatioTemporal Adaptive Reward (STAR) Allocation** for RL post-training of text-to-image diffusion and flow models. STAR uses text-image attention inside the generative model and starts from the core content that the user truly cares about in the prompt. It constructs spatial allocation maps that dynamically vary across denoising steps and rollouts, and allocates the same group-relative advantage to more relevant latent regions with almost no additional computational overhead. STAR then applies stronger policy updates to these regions through a spatially resolved policy objective. We use Stable Diffusion 3.5 Medium as the base model and evaluate on three tasks: GenEval, OCR text rendering, and PickScore. Experimental results show that STAR improves compositional semantic alignment, text rendering, and preference optimization without changing the external reward source, achieving $\mathbf{0.9759}$, $\mathbf{0.9757}$, and $\mathbf{23.60}$ on GenEval, OCR, and PickScore, respectively.

1 INTRODUCTION

Diffusion models have become mainstream frameworks for high-fidelity text-to-image generation (Ho et al., 2020; Dhariwal & Nichol, 2021; Song et al., 2021b;a; Rombach et al., 2022). Recent text-to-image systems further improve resolution and language understanding through latent diffusion, rectified-flow transformers, and large-scale prompt-image training (Podell et al., 2024; Saharia et al., 2022; Esser et al., 2024). Although these models have made substantial progress in visual quality, compositional generalization under complex prompts remains a major weakness: models often make errors in object counting, attribute binding, and spatial relations (Ghosh et al., 2023; Huang et al.,

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

2023). Since such errors can usually only be judged on the completed image by automatic evaluators, preference models (Xu et al., 2023; Kirstain et al., 2023), or human feedback, RL-style post-training provides an effective path for directly optimizing image-level objectives.

A key limitation of existing RL post-training for text-to-image generation lies in the granularity of credit assignment. The final reward is usually an image-level scalar, whereas diffusion and flow trajectories have explicit temporal and spatial structure: sampling spans multiple denoising or flow steps, each containing many latent spatial tokens. Existing methods, including Flow-GRPO (Liu et al., 2025), typically apply this scalar reward uniformly across timesteps and spatial locations, assigning the same reward weight to different denoising steps and image regions. However, these components of the generative process differ in their contributions to text alignment. For example, given the prompt "a bicycle," text alignment depends primarily on whether the bicycle region is generated correctly, while background regions such as the sky or the road have a smaller influence. When prompts additionally specify colors, counts, and spatial relations, the reward depends more strongly on whether the relevant objects, attributes, and spatial layout satisfy the textual requirements. Uniform reward weights do not reflect these differences in regional contributions, making it difficult for policy updates to focus on the regions most relevant to the prompt.

These differences in contribution also arise along the temporal dimension. In diffusion and flow models, different denoising stages place different emphasis on global layout, object formation, and detail refinement (Lipman et al., 2023; Liu et al., 2023), and the contribution of the same latent region to text alignment can change as generation progresses. Reward allocation should therefore distinguish spatial regions and adapt across denoising steps. As shown in Figure 1, spatiotemporal adaptive allocation uses spatial reward distributions that vary across timesteps, allowing regional reward weights to evolve with the generation process. We study spatiotemporal credit assignment in this structured process, allocating the same image-level reward across denoising steps and latent regions so that policy updates focus on the components most relevant to the prompt at each stage.

We propose **SpatioTemporal Adaptive Reward (STAR) Allocation** for RL post-training of attention-based text-to-image (T2I) diffusion and flow models. STAR uses text-image attention within the generative model to construct spatial reward allocation maps that vary across denoising steps. Given the same scalar image-level reward, STAR uses these maps to adjust reward weights across timesteps and latent regions, and modulates policy update strength through a spatially resolved likelihood-ratio objective. Regions corresponding to key objects, attributes, and relations in the prompt receive stronger updates, while background and weakly related regions retain weaker updates, balancing text alignment with overall generation quality.

We instantiate STAR in compositional T2I RL post-training and compare it with scalar-allocation baselines under the same external reward source. Experimental results show that STAR achieves 0.9759, 0.9757, and 23.60 on GenEval, OCR, and PickScore, respectively.

Our contributions are as follows:

- We propose **SpatioTemporal Adaptive Reward (STAR) Allocation**, a method for T2I diffusion/flow RL post-training. STAR converts image-level rewards into training signals over denoising time and latent space, allowing updates to focus on the generative regions corresponding to key objects, attributes, and relations in the prompt.
- We use text-image interactions inside the model to construct timestep-specific spatial allocation maps, extend existing RL post-training objectives into a spatially resolved policy objective, and provide a unified reward-routing interface for attention-based T2I diffusion/flow models.
- We validate STAR on GenEval, OCR, and PickScore, where it achieves 0.9759, 0.9757, and 23.60, respectively, showing that SpatioTemporal Adaptive Reward (STAR) Allocation can simultaneously improve compositional semantic alignment, text rendering, and preference optimization.

2 Related Work

RL post-training for T2I. In recent years, directly optimizing downstream rewards has become an important direction for improving diffusion and flow-based T2I models, following the broader policy-

gradient view of reinforcement learning (Sutton & Barto, 1998; Williams, 1992). DDPO (Black et al., 2024) formulates the denoising process as a multi-step decision-making problem and optimizes diffusion models with policy-gradient methods. DPOK (Fan et al., 2023) further introduces online RL and KL regularization into T2I diffusion fine-tuning. Diffusion-DPO (Wallace et al., 2024) builds on direct preference optimization (Rafailov et al., 2023), while AlignProp (Prabhudesai et al., 2023) optimizes human preferences through differentiable reward backpropagation. More recently, Flow-GRPO (Liu et al., 2025) extends online policy-gradient training to flow matching models and enables efficient post-training through stochastic sampling and denoising reduction. These methods demonstrate the effectiveness of RL-style post-training for T2I models, but they usually treat the final reward as a scalar signal applied to the whole image or the entire trajectory. STAR studies how the same reward signal should be allocated within a structured denoising trajectory.

Fine-grained rewards and credit assignment. Existing work has begun to recognize that image-level scalar rewards provide overly coarse feedback for diffusion alignment. In A Dense Reward View, Yang et al. (2024) revisit preference alignment from a sequential decision-making perspective and derive a dense objective that emphasizes early denoising steps. B^2-DiffuRL (Hu et al., 2025) addresses sparse terminal rewards by estimating the contributions of intermediate denoising actions through backward progressive training and branch-based sampling. Another line of work improves the reward source itself. ImageReward (Xu et al., 2023) learns a general T2I human preference reward model. ImageDoctor (Guo et al., 2025) evaluates generated images across four dimensions: plausibility, semantic alignment, aesthetics, and overall quality. It also produces pixel-level heatmaps that localize flaws, providing fine-grained rewards for preference optimization. SpatialReward (Zhou et al., 2026) decomposes prompts into entities, attributes, and spatial relations, and combines localization results from specialized detectors with vision-language model reasoning to assess how well generated images satisfy spatial constraints. It then uses these scores for RL training to improve the spatial consistency of generated images. These methods provide more fine-grained feedback by improving the reward model. We further investigate how the generative model's own text-image interactions can allocate an existing image-level reward across denoising steps and latent regions.

3 METHOD

STAR reallocates an existing image-level reward to the structured generative process of diffusion/flow models. Given a prompt c, the model generates a latent trajectory

$$\tau = (z_T, z_{T-1}, \ldots, z_0), \qquad z_t \in \mathbb{R}^{C \times h_t \times w_t}, \tag{1}$$

where z_0 corresponds to the final image. Let $\Omega_t = \{1, \ldots, h_t\} \times \{1, \ldots, w_t\}$ denote the set of spatial locations in the latent feature map at step t. The external reward source provides scalar feedback only on the final sample, denoted as

$$R_i = R(z_0^i, c), \qquad A_i = \mathcal{B}(R_i, c), \tag{2}$$

where i denotes the i-th rollout under the same prompt, and A_i is the normalized reward obtained by group-relative normalization or baseline subtraction. A standard RL-style objective uses A_i as a trajectory-level weight, as in group-relative policy optimization (Shao et al., 2024); STAR routes A_i to denoising steps and latent regions.

3.1 ATTENTION-GUIDED ALLOCATION

STAR represents each prompt as a set of reward-relevant text units:

$$\mathcal{K}(c) = \{q_1, \ldots, q_K\}, \tag{3}$$

where q_k denotes a core component in the prompt. These components are the conditions that the final image needs to satisfy, including subjects, styles, attributes, counts, and spatial relations. For example, for the prompt "generate a futuristic-style building," $\mathcal{K}(c)$ contains the subject "building" and the visual style requirement "futuristic style." Dataset-specific extraction rules and prompt templates for text-unit extraction are provided in Appendix A.

Suppose the full prompt c is converted by the text encoder tokenizer into L_c tokens. Each text unit q_k corresponds to several text tokens, whose token indices are denoted as

$$\mathcal{T}_k \subseteq \{1, \ldots, L_c\}. \tag{4}$$

STAR reads routing signals from a selected set of text-image attention layers $\mathcal{L}$. When generating rollout i, attention-based T2I backbones establish interactions between image latent tokens and text tokens at each denoising step t. For any $\ell \in \mathcal{L}$, this layer provides an attention map $\alpha_t^{i,\ell} \in \mathbb{R}^{h_t \times w_t \times L_c}$; the attention weight from latent position u to text token j is denoted as

$$\alpha_t^{i,\ell}(u, j). \tag{5}$$

Here $u \in \Omega_t$ is a latent spatial position and j is a text token index. In joint-attention backbones such as Stable Diffusion 3.5 and Qwen-Image (Esser et al., 2024; Wu et al., 2025), this value corresponds to the image-to-text block of the full attention matrix; in cross-attention backbones such as Stable Diffusion XL (Podell et al., 2024), it corresponds to the cross-attention weights from image queries to text keys. STAR takes the maximum image-to-text attention weight over the selected layers:

$$\alpha_t^i(u, j) = \max_{\ell \in \mathcal{L}} \alpha_t^{i,\ell}(u, j), \tag{6}$$

which gives the timestep-specific text-image routing map $\alpha_t^i \in \mathbb{R}^{h_t \times w_t \times L_c}$ for rollout i at step t.

Next, STAR extracts where each text unit appears in the current trajectory. Given the token set $\mathcal{T}_k$ of text unit q_k, STAR constructs its spatial map $M_t^{i,k} \in \mathbb{R}^{h_t \times w_t}$, where the routing strength at latent position u is

$$M_t^{i,k}(u) = \frac{1}{|\mathcal{T}_k|} \sum_{j \in \mathcal{T}_k} \alpha_t^i(u, j). \tag{7}$$

STAR aggregates these key components into a trajectory-timestep specific heatmap $H_t^i \in \mathbb{R}^{h_t \times w_t}$:

$$H_t^i(u) = \frac{1}{K} \sum_{k=1}^{K} M_t^{i,k}(u). \tag{8}$$

This heatmap captures the spatial distribution of attention to the selected text units. STAR maps it to bounded reward-allocation coefficients $C_t^i \in \mathbb{R}^{h_t \times w_t}$:

$$C_t^i(u) = 1 + (\lambda - 1) \frac{H_t^i(u) - \min_{v \in \Omega_t} H_t^i(v)}{\max_{v \in \Omega_t} H_t^i(v) - \min_{v \in \Omega_t} H_t^i(v) + \epsilon}. \tag{9}$$

Here $\lambda \geq 1$ controls the allocation strength, and ϵ ensures numerical stability. Larger $C_t^i(u)$ increases the coefficient of the corresponding local log-density contribution, while all locations retain positive allocation coefficients.

3.2 Spatially Resolved Policy Objective

STAR uses the allocation map to convert the normalized reward into spatial training weights. For the i-th rollout and denoising step t, we define

$$A_t^i(u) = A_i \, C_t^i(u), \qquad A_t^i \in \mathbb{R}^{h_t \times w_t}. \tag{10}$$

This mapping preserves the sign of A_i and assigns larger reward weights to locations with higher aggregated text–image attention.

The policy update acts on spatial log-probability contributions. Let

$$\ell_{\theta,t}^i \in \mathbb{R}^{h_t \times w_t}, \qquad \bar{\ell}_{\theta,t}^i = \frac{1}{|\Omega_t|} \sum_{u \in \Omega_t} \ell_{\theta,t}^i(u), \tag{11}$$

where $\ell_{\theta,t}^i(u)$ is the local log-density quantity at position u, and $\bar{\ell}_{\theta,t}^i$ is its spatial average. The scalar comparator uses the same channel and spatial reductions. STAR defines the spatial ratio as

$$\rho_{\theta,t}^i(u) = \mathrm{sg}\big[\exp\big(\bar{\ell}_{\theta,t}^i - \bar{\ell}_{\mathrm{old},t}^i\big)\big] \, \frac{\exp(\ell_{\theta,t}^i(u))}{\mathrm{sg}\Big[\exp(\ell_{\theta,t}^i(u))\Big]}, \tag{12}$$

where $\bar{\ell}_{\mathrm{old},t}^i$ is the corresponding quantity stored during sampling, and $\mathrm{sg}[\cdot]$ denotes stop-gradient. The ratio has the same forward value at every location; its specified derivative routes gradients to local log-density contributions.

Finally, STAR optimizes the clipped spatial objective, following the clipped-ratio policy update used in PPO-style training (Schulman et al., 2017):

$$\mathcal{J}_{\mathrm{STAR}}(\theta) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{1}{|\Omega_t|}\sum_{u\in\Omega_t}\min\left(\rho^i_{\theta,t}(u)A^i_t(u), \mathrm{clip}(\rho^i_{\theta,t}(u), 1-\varepsilon, 1+\varepsilon)A^i_t(u)\right)\right] - \beta\mathcal{D}_{\mathrm{KL}}. \tag{13}$$

Here $\mathcal{D}_{\mathrm{KL}}$ regularizes denoising transitions by penalizing deviations from the frozen reference policy.

4 Theoretical Analysis

Overall. Theorem 5 shows that, under the stated assumptions, STAR uses the reward signal more effectively for policy optimization: for sufficiently small common step sizes, STAR yields a positive local optimization gain that exceeds that of spatially uniform weighting. To establish this result, we follow the design in Section 3: Theorem 1 motivates timestep-specific heatmaps, Theorem 2 links the spatial ratio to local gradient reweighting, and Theorems 3–4 supply the required coherence and alignment. Complete proofs are in Appendix B.

Theorem 1 (Spatiotemporal reward sensitivity)**.** Assume $p_\theta(z_{t-1} \mid z_t, c) = \mathcal{N}(\mu_{\theta,t}(z_t, c), \sigma_t^2 I)$ has a differentiable mean and strictly positive, parameter-independent variance. For bounded measurable R, define $V^\theta_{t-1}(z, c) = \mathbb{E}[R \mid z_{t-1} = z, c]$ and $W^\theta_t(m, c) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}[V^\theta_{t-1}(m + \sigma_t\epsilon, c)]$. The channel-summed local log density ℓ^{dens} satisfies

$$\mathbb{E}[R\nabla_\theta \ell^{\mathrm{dens}}_{\theta,t}(u) \mid z_t, c] = (D_\theta \mu_{\theta,t,u})^\top \nabla_{m_u} W^\theta_t(m, c)\big|_{m=\mu_{\theta,t}(z_t,c)}. \tag{14}$$

The score derivative holds sampled states fixed; the m-derivative holds the future policy fixed.

The transition Jacobian and continuation sensitivity motivate the time-varying maps in Eq. (8). STAR uses text–image attention to prioritize prompt-selected locations; Theorem 4 gives a sufficient structural model for useful alignment. Theorem 1 is proved in Appendix B.2.

Theorem 2 (Exact local gradient reweighting)**.** Assume a batch of N trajectories with T transitions each, with sampled states, stored old log probabilities, normalized rewards A_i, and maps fixed during differentiation. Use identical rewards, probability reductions, and clipping in STAR and the scalar surrogate. At $\theta = \theta_{\mathrm{old}}$ with positive clipping width, Eqs. (10)–(12) yield the reward-part gradients

$$\begin{aligned} q_{i,t,u} &= \frac{A_i}{NT|\Omega_t|}\nabla_\theta \ell^i_{\theta,t}(u)\big|_{\theta=\theta_{\mathrm{old}}}, \\ g_{\mathrm{scalar}} &= \sum_{i,t,u} q_{i,t,u}, \qquad g_{\mathrm{STAR}} = \sum_{i,t,u} C^i_t(u) q_{i,t,u}. \end{aligned} \tag{15}$$

The stop-gradient construction thus retains $C^i_t(u)$ on each local contribution. Using an ordinary scalar ratio would collapse these weights to their per-step spatial mean $\bar{C}_{i,t} = |\Omega_t|^{-1}\sum_u C^i_t(u)$ (Appendix B.3).

At $\theta = \theta_{\mathrm{old}}$, assume the matched scalar surrogate F in Eq. (35) and its reference KL are differentiable. Write $q_{\mathrm{ref}} = -\beta\nabla_\theta\mathcal{D}_{\mathrm{KL}}$, $f = \nabla F = g_{\mathrm{scalar}} + q_{\mathrm{ref}}$, and $h = g_{\mathrm{STAR}} + q_{\mathrm{ref}}$. Using Euclidean norms and finite $\lambda \geq 1$, define $B_{\mathrm{tot}} = \sum_{i,t,u}\|q_{i,t,u}\| + \|q_{\mathrm{ref}}\|$, $m_\lambda = (\lambda+1)/2$, and $\delta_\lambda = (\lambda-1)/(\lambda+1)$.

Theorem 3 (Gradient coherence under a dominant common direction)**.** Let v_j collect the local reward contributions and the reference contribution. Assume $v_j = b_j d + e_j$, with $\|d\| = 1$, $b_j \geq 0$, $B_0 = \sum_j b_j > 0$, and $E_0 = \sum_j \|e_j\| < B_0/\lambda$. Then

$$\frac{\|f\|}{B_{\mathrm{tot}}} \geq \frac{B_0 - E_0}{B_0 + E_0} > \delta_\lambda. \tag{16}$$

Thus a dominant shared component gives sufficient gradient coherence. For $\lambda = 1.5$, the required residual ratio is $E_0/B_0 < 2/3$. The proof and a less conservative ascent criterion are in Appendix B.6.1.

Theorem 4 (Attention–gradient alignment from semantic separation)**.** Assume $\lambda > 1$. In each block $b = (i, t)$, write $H_{b,u} = H^i_t(u)$ and $a_{b,u} = f^\top q_{i,t,u}$. Partition positions into prompt-relevant and other regions using a binary $Z_{b,u}$ specified from semantics before inspecting a_b. Let

$p_b \in (0,1)$ be the relevant fraction and $\Delta H_b, \Delta a_b$ the relevant-minus-other mean differences. Set $w_b = |\Omega_t|/(\max_u H_{b,u} - \min_u H_{b,u} + \epsilon_{\text{map}})$, with $\epsilon_{\text{map}} > 0$ denoting the stabilizer in Eq. (9). Define the weighted between-region signal $\mathcal{P} = \sum_b w_b p_b(1-p_b)\Delta H_b \Delta a_b$ and within-region covariance $\mathcal{U} = \sum_b w_b \mathbb{E}_{Z_b}[\mathrm{Cov}_u(H_b, a_b \mid Z_b)]$, using uniform spatial averaging. If $\mathcal{P} > 0$ and $\mathcal{U} \geq -\zeta\mathcal{P}$ for some $0 \leq \zeta < 1$, then

$$\Gamma = \sum_{i,t,u} \big(C_t^i(u) - \bar{C}_{i,t}\big) f^\top q_{i,t,u} \geq (\lambda - 1)(1-\zeta)\mathcal{P} > 0. \tag{17}$$

Semantic separation therefore yields positive alignment when its signal exceeds the adverse within-region covariance. The proof is in Appendix B.6.2; localization evidence and existing routing ablations motivate the model in Appendix B.6.3.

Theorem 5 (Local optimization benefit of STAR)**.** Under Theorem 2's assumptions, Eq. (9) ensures

$$\left\| \frac{h}{m_\lambda} - f \right\| \leq \delta_\lambda B_{\text{tot}}. \tag{18}$$

If $C_t^i(u) \equiv 1$, $h = f$. Under the structural hypotheses of Theorems 3–4, for all sufficiently small $\eta > 0$,

$$\begin{aligned} F(\theta + \eta h) &> F(\theta), \\ F(\theta + \eta h) - F(\theta + \eta h_{\text{uniform}}) &= \eta\Gamma + o(\eta) > 0, \end{aligned} \tag{19}$$

where $h_{\text{uniform}} = \sum_{i,t,u} \bar{C}_{i,t} q_{i,t,u} + q_{\text{ref}}$ uses the same mean reward coefficient for each trajectory and timestep, and the same reference penalty. Proofs are in Appendices B.4–B.5. The first inequality in Eq. (19) establishes a positive local optimization gain from STAR's use of the reward signal; the second shows that aligned allocation weights yield an additional gain over spatially uniform weighting by emphasizing useful local gradient contributions. These results provide a theoretical basis for STAR's adaptive reward allocation in policy optimization.

5 Experiments

5.1 Implementation Details

Evaluation tasks. The experiments include three T2I post-training tasks: GenEval, OCR text rendering, and PickScore. GenEval evaluates compositional generation ability, including object counting, attribute binding, colors, and spatial relations (Ghosh et al., 2023); training prompts are generated from the official GenEval templates, and rewards are computed by the official rule-based evaluation pipeline. OCR text rendering (Chen et al., 2023) requires the model to generate a specified text string in the image, and the reward is computed from the edit distance between the rendered text and the target text. PickScore (Kirstain et al., 2023) uses a learned preference reward model to score each prompt-image pair.

Baselines. We compare against representative T2I baselines: diffusion models, including LDM, SD1.5, SD2.1 (Rombach et al., 2022), SD-XL (Podell et al., 2024), DALLE-2 (Ramesh et al., 2022), DALLE-3 (Betker et al., 2023); autoregressive models, including Show-o (Xie et al., 2025b), Emu3-Gen (Wang et al., 2024), JanusFlow (Ma et al., 2025), Janus-Pro-7B (Chen et al., 2025); and recent diffusion/flow models, including FLUX.1 Dev (Black Forest Labs, 2024), SD3.5-L and SD3.5-M (Esser et al., 2024), SANA-1.5 4.8B (Xie et al., 2025a), SD3.5-M+Flow-GRPO (Liu et al., 2025).

Experimental setup. We implement STAR based on the Flow-GRPO (Liu et al., 2025) training pipeline and use Stable Diffusion 3.5 Medium (SD3.5-M) as the base model. Training uses ODE-to-SDE rollouts for online sampling, with the number of sampling steps set to $T = 10$, the SDE noise level set to $a = 0.7$, and the image resolution set to 512×512; evaluation uses $T = 40$ steps. For each prompt, we sample $G = 24$ images to form a group, and update the model with a group-relative policy-gradient objective, clipped ratio, and KL regularization. The KL coefficient is set to $\beta = 0.04$ for GenEval and OCR text rendering, and $\beta = 0.01$ for PickScore. We train the model with LoRA (Hu et al., 2022), using rank $r = 32$ and scaling factor $\alpha = 64$. STAR records text-image attention during rollout and uses the spatial objective in Section 3 to allocate group-relative advantages. Attention routing uses the joint attention of SD3.5-M, reads the attention layers from transformer blocks 10–15, and sets the allocation strength to $\lambda = 1.5$.

Table 1: GenEval results. Baseline scores follow Flow-GRPO (Liu et al., 2025). All values are shown to two decimal places; Overall may differ from the mean of rounded sub-task scores. Obj. and Attr. denote object and attribute. Blue and green indicate the best and second-best results.

Model	Overall	Single Obj.	Two Obj.	Counting	Colors	Position	Attr. Binding
Diffusion Models							
LDM	0.37	0.92	0.29	0.23	0.70	0.02	0.05
SD1.5	0.43	0.97	0.38	0.35	0.76	0.04	0.06
SD2.1	0.50	0.98	0.51	0.44	0.85	0.07	0.17
SD-XL	0.55	0.98	0.74	0.39	0.85	0.15	0.23
DALLE-2	0.52	0.94	0.66	0.49	0.77	0.10	0.19
DALLE-3	0.67	0.96	0.87	0.47	0.83	0.43	0.45
Autoregressive Models							
Show-o	0.53	0.95	0.52	0.49	0.82	0.11	0.28
Emu3-Gen	0.54	0.98	0.71	0.34	0.81	0.17	0.21
JanusFlow	0.63	0.97	0.59	0.45	0.83	0.53	0.42
Janus-Pro-7B	0.80	0.99	0.89	0.59	0.90	0.79	0.66
Diffusion / Flow Models							
FLUX.1 Dev	0.66	0.98	0.81	0.74	0.79	0.22	0.45
SD3.5-L	0.71	0.98	0.89	0.73	0.83	0.34	0.47
SANA-1.5 4.8B	0.81	0.99	0.93	0.86	0.84	0.59	0.65
SD3.5-M	0.63	0.98	0.78	0.50	0.81	0.24	0.52
SD3.5-M+Flow-GRPO	0.95	**1.00**	0.99	0.95	0.92	**0.99**	0.86
SD3.5-M+Flow-GRPO+STAR	**0.98**	**1.00**	**1.00**	**0.98**	**0.96**	**0.98**	**0.94**

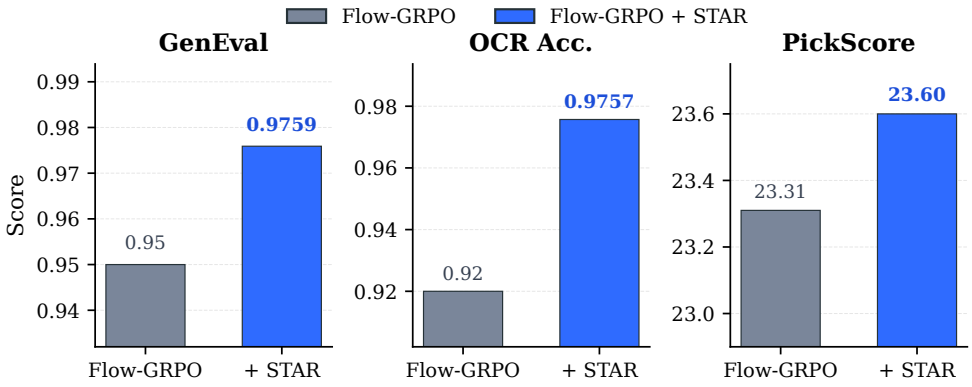
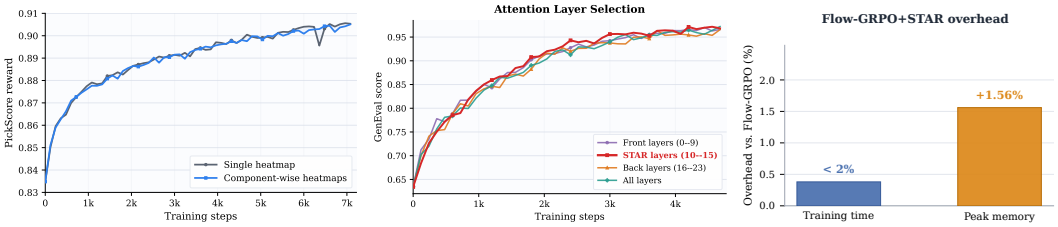


Figure 2: Main results on three T2I post-training benchmarks. STAR consistently improves over the Flow-GRPO baseline on compositional generation, text rendering, and preference optimization.

5.2 Main Results

GenEval results. Table 1 presents the main results on GenEval (Ghosh et al., 2023). STAR achieves an overall score of 0.9759, outperforming SD3.5-M at 0.63 and Flow-GRPO (Liu et al., 2025) at 0.95. This improvement mainly comes from more fine-grained compositional semantic constraints: on counting, colors, and attribute binding, STAR reaches 0.98, 0.96, and 0.94, respectively, clearly improving over 0.95, 0.92, and 0.86. Two-object generation further reaches 1.00. On single-object generation and position, two sub-tasks that are already close to saturation, STAR still maintains near-optimal performance, reaching 1.00 and 0.98, respectively.

These results show that the gains of STAR come from more accurate spatiotemporal credit assignment. By allocating the same image-level advantage to prompt-relevant latent regions, STAR can more selectively update the generative regions that affect object counts, colors, and attribute binding, thereby improving compositional generation performance.



(a) Prompt heatmap construction. (b) Attention layer selection. (c) Computational overhead.

Figure 3: Ablation and efficiency analysis for STAR, covering reward routing design choices and additional training overhead.

Results across benchmarks. Figure 2 further compares Flow-GRPO (Liu et al., 2025) and STAR on the three post-training tasks. STAR achieves 0.9759 on GenEval (Ghosh et al., 2023), 0.9757 on OCR text rendering (Chen et al., 2023), and 23.60 on PickScore (Kirstain et al., 2023), all outperforming the corresponding Flow-GRPO baselines. The improvements on GenEval and OCR indicate that spatiotemporal adaptive reward allocation can more effectively optimize verifiable text-alignment objectives; the improvement on PickScore shows that this allocation scheme can also be used with continuous rewards provided by a preference model. Overall, without changing the form of the task reward, STAR allocates the same sample-level feedback to more relevant generative regions and thus obtains consistent gains in compositional generation, text rendering, and preference optimization.

5.3 Ablation Study

5.3.1 Prompt Heatmap Construction

We compare two ways of constructing prompt heatmaps. The single-heatmap variant first aggregates the attention of multiple key components in the prompt into one heatmap and then obtains reward allocation coefficients; the component-wise heatmaps variant first generates a heatmap for each key component separately and then stacks and normalizes them. Both experiments use the same PickScore reward, training setting, and evaluation interval.

Figure 3a shows that the training curves of the two variants are overall very close, with the single heatmap showing a slight advantage in the later stage of training. This result indicates that directly aggregating multiple components that the user cares about in the prompt into a shared heatmap already provides a stable routing signal. Generating heatmaps component by component and then stacking them does not bring additional gains and increases implementation complexity; therefore, we use single-heatmap construction in the main experiments.

5.3.2 Attention Layer Selection

We further analyze the choice of text-image attention layers in STAR. We compare four settings: front layers use transformer blocks 0–9, STAR layers use blocks 10–15, back layers use blocks 16–23, and all layers use all attention layers.

Figure 3b shows that layer selection affects the stability of attention routing. Front layers and back layers both bring effective improvements, but their later-stage curves are slightly lower; all layers reach the highest level in the final stage, but require reading more attention maps. STAR layers remain in the best-performing group for most of training and are overall close to all layers, while using only a small number of middle blocks. This indicates that blocks 10–15 already contain sufficiently clear text-image alignment signals. Therefore, we use this layer set in the main experiments to achieve a better balance between effectiveness and implementation overhead.

5.3.3 Computational Overhead

We further compare the training overhead of Flow-GRPO (Liu et al., 2025) and Flow-GRPO+STAR. STAR reuses the text-image attention from the forward pass of the generative model and only

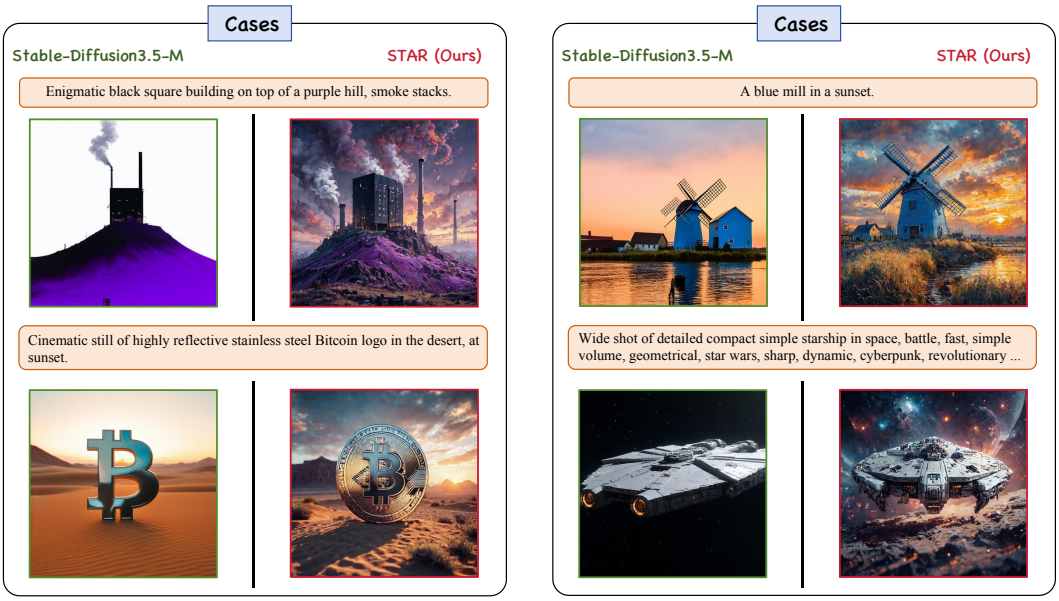


Figure 4: Qualitative comparison between the original SD3.5-M and SD3.5-M after Flow-GRPO+STAR post-training. Within every prompt pair, the left image is produced by SD3.5-M and the right image by SD3.5-M+Flow-GRPO+STAR.

performs heatmap aggregation and advantage rescaling on the selected layers; it does not require training an additional reward model or adding new image evaluation steps.

Figure 3c shows that STAR introduces less than 2% additional training time overhead, and increases peak memory by only 1.56%. This result shows that the main computation of STAR still comes from the original diffusion/flow rollout and GRPO update; attention-based allocation is only a lightweight reward-routing step. Therefore, we can obtain more fine-grained spatiotemporal credit assignment with almost no change in training cost.

5.4 Qualitative Comparison

To qualitatively examine the effect of post-training with STAR, we compare the original SD3.5-M base model with SD3.5-M after Flow-GRPO+STAR post-training on a diverse set of prompts.

As shown in Figure 4, both models generate images that capture the main concepts in the prompts. Across the displayed examples, Flow-GRPO+STAR post-training changes how the requested subjects, attributes, and surrounding scenes are realized. For example, the post-trained model depicts the black square building, reflective stainless-steel Bitcoin logo, blue windmill, and compact starship with more complete prompt-related elements and more coherent scene composition.

These qualitative results complement the quantitative evaluations and illustrate the effect of applying Flow-GRPO+STAR post-training to SD3.5-M across different types of prompts. By allocating the reward signal to prompt-relevant latent regions, STAR encourages the model to better incorporate the textual conditions during generation.

6 Conclusion

We propose **STAR** for spatiotemporal reward allocation in RL post-training of T2I diffusion and flow models. STAR uses internal text-image interactions to distribute image-level reward across denoising steps and latent regions, focusing updates on prompt-relevant content. Experiments show consistent gains over Flow-GRPO in compositional generation, text rendering, and preference alignment, using the same external reward sources with minimal additional training overhead.

AI USE STATEMENT

Generative AI tools were used solely to aid and polish the writing of this paper. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of the paper.

REFERENCES

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. Technical report, OpenAI, 2023. URL https://cdn.openai.com/papers/dall-e-3.pdf.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

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

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-Excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics*, 42(4):148:1–148:10, July 2023. ISSN 1557-7368. doi: 10.1145/3592116. URL http://dx.doi.org/10.1145/3592116.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. TextDiffuser: Diffusion models as text painters. In *Advances in Neural Information Processing Systems*, volume 36, pp. 9353–9387, 2023.

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.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in neural information processing systems*, volume 34, pp. 8780–8794, 2021.

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.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 79858–79885, 2023.

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

Yuxiang Guo, Jiang Liu, Ze Wang, Hao Chen, Ximeng Sun, Yang Zhao, Jialian Wu, Xiaodong Yu, Zicheng Liu, and Emad Barsoum. ImageDoctor: Diagnosing text-to-image generation via grounded image reasoning. *CoRR*, abs/2510.01010, 2025. doi: 10.48550/ARXIV.2510.01010. URL https://doi.org/10.48550/ARXIV.2510.01010.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in neural information processing systems*, volume 33, pp. 6840–6851, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

Zijing Hu, Fengda Zhang, Long Chen, Kun Kuang, Jiahui Li, Kaifeng Gao, Jun Xiao, Xin Wang, and Wenwu Zhu. Towards better alignment: Training diffusion models with reinforcement learning against sparse rewards. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23604–23614, 2025. doi: 10.1109/CVPR52734.2025.02198.

URL https://openaccess.thecvf.com/content/CVPR2025/html/Hu_Towards_Better_Alignment_Training_Diffusion_Models_with_Reinforcement_Learning_Against_CVPR_2025_paper.html.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78723–78747, 2023.

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. In *Advances in Neural Information Processing Systems*, volume 36, pp. 36652–36663, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

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. In *Advances in Neural Information Processing Systems*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. JanusFlow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 7739–7751, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023.

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. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in neural information processing systems*, volume 35, pp. 36479–36494, 2022.

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

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

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021b.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT Press, 1998.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5644–5659, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.310. URL https://aclanthology.org/2023.acl-long.310/.

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*, pp. 8228–8238, 2024.

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.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report. *CoRR*, abs/2508.02324, 2025. doi: 10.48550/ARXIV.2508.02324. URL https://arxiv.org/abs/2508.02324.

Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, Bingchen Liu, Daquan Zhou, and Song Han. SANA 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, 2025a.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025b.

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. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Shentao Yang, Tianqi Chen, and Mingyuan Zhou. A dense reward view on aligning text-to-image diffusion with preference. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, pp. 55998–56032, 2024.

Sashuai Zhou, Qiang Zhou, Junpeng Ma, Yue Cao, Ruofan Hu, Ziang Zhang, Xiaoda Yang, Zhibin Wang, Jun Song, Cheng Yu, Bo Zheng, and Zhou Zhao. SpatialReward: Verifiable spatial reward modeling for fine-grained spatial consistency in text-to-image generation. *CoRR*, abs/2603.22228, 2026. doi: 10.48550/ARXIV.2603.22228. URL https://doi.org/10.48550/ARXIV.2603.22228.

A TEXT-UNIT EXTRACTION PROMPTS

This appendix reports the task-specific text-unit extraction procedures used by STAR. These text units are used only to locate the corresponding text-token indices and aggregate text-image attention maps; the external image-level reward for each task remains unchanged.

A.1 GENEVAL TEXT UNITS

For GenEval, STAR does not require an LLM for text-unit extraction. Instead, it applies a deterministic focus-span rule: remove a boilerplate image prefix such as "a photo of", "an image of", or "a painting of", and use the remaining phrase as a single text unit. This matches the GenEval prompt format, where the suffix already contains the objects, attributes, counts, and spatial relations to be grounded.

```
GenEval structured-phrase extraction

Input:  c = original GenEval prompt
Rule:   remove prefix matching
        ^(?:a|an|the)?\s*(?:photo|picture|image|painting|drawing)\s+of\s+
Output: K(c) = [remaining focus span]

Examples:
Input:  "a photo of a blue book"
Output: ["a blue book"]

Input:  "a photo of a bus below a cat"
Output: ["a bus below a cat"]

Input:  "a photo of a yellow bus and an orange handbag"
Output: ["a yellow bus and an orange handbag"]
```

A.2 OCR TEXT UNITS

OCR prompts contain exact quoted strings that must be rendered in the image. We therefore extract visually localizable text units in advance and associate them with the corresponding training prompts. These units include the exact rendered text, its carrier or surface, visible text attributes, and other concrete scene elements. When pre-extracted units are unavailable, STAR falls back to the deterministic focus-span rule described above. The extraction prompt is shown below.

```
You are an expert at analyzing OCR/text-to-image prompts. Extract only the important,
     visually localizable semantic components that should correspond to concrete
    image regions and cross-attention heatmaps.

Priority for OCR/text-rendering prompts:
1. Exact rendered text: every quoted word/phrase that must appear legibly in the
    image. Preserve spelling and capitalization.
2. Text carrier/surface: the object or region that contains the text (label,
    billboard, screen, sign, page, ribbon, poster, display, book cover).
3. Text style/appearance: concrete visible style/color/layout of the rendered text or
     carrier (red warning label, bold letters, glowing green text).
4. Main objects/subjects: entities that must appear (medicine bottle, robot chest
    panel, spellbook, astronaut boot print).
5. Concrete scene/background: visually localizable environment (industrial background,
     Martian surface, neutral background).
6. Composition/viewpoint/style only when visually important and concise (close-up,
    textbook diagram, realistic photograph).

Include:
- Objects/subjects and their key visible attributes.
```

```
- Actions/interactions only when they create a visible region.
- Exact quoted OCR text as a standalone component.
- Text carrier + text content as separate components when both exist.
- Short phrases that can plausibly map to a region in the image.

Exclude:
- Abstract intent/emotion/judgment with no clear visual region: urgency, achievement,
     accomplishment, vibe, feeling, mood, intrigued, dismissive.
- Generic filler: a photo of, a picture of, an image of, featuring, set against, that
     reads, clearly reading, prominently displayed.
- Standalone connectors/prepositions: in, on, with, of, that are, a couple of.
- Quality tags: 8k, UHD, highly detailed, masterpiece, trending on Artstation.
- Artist-name modifiers.
- The full prompt itself.

Output format:
Return a JSON array of strings. Each string is one important component, as a short
     phrase (1-7 words). Prefer 3-7 components per prompt. Keep the original wording
     when possible. For OCR/text-rendering prompts, include every exact quoted text as
      one component. Do NOT include the full prompt as a component.

Examples:
Prompt: "A close-up of a medicine bottle with a clear, red warning label that reads
     \"Take With Food\" prominently displayed, set against a neutral background."
-> ["close-up", "medicine bottle", "clear red warning label", "Take With Food", "
     neutral background"]

Prompt: "A close-up of a robot's chest panel, with a digital display blinking \"
     System Override Active\" in red, set against a dimly lit industrial background."
-> ["close-up", "robot's chest panel", "digital display", "System Override Active", "
     red blinking text", "dimly lit industrial background"]

Batch-format instruction:
Extract the important semantic components from each of the following prompts. Return
     a JSON array for each prompt, in order. Format: return a JSON array of JSON
     arrays, one inner array per prompt.

Prompts:
1. {prompt_1}
2. {prompt_2}
...

Return ONLY a JSON array of arrays, e.g.:
[["comp1", "comp2"], ["comp3"], ...]
```

A.3 PICKSCORE TEXT UNITS

For PickScore, STAR uses the GenEval focus-span rule by default. We additionally consider component-level decomposition, in which a vision-language model extracts concise semantic units describing objects, attributes, quantities, and spatial relations. The decomposition prompt is provided below.

```
You are a precise text analyzer. Given an image generation prompt, decompose it into
     semantic components.

Rules:
- Extract each distinct object with its attributes (e.g., "red cat", "wooden table")
- Extract spatial relationships (e.g., "sitting on", "next to", "above")
- Extract quantities if specified (e.g., "two cups", "three dogs")
- Each component should be a short phrase (1-4 words)
- Return ONLY a JSON list of strings, no explanation
```

```
Example:
Input: "a photo of a red cat sitting on a wooden table"
Output: ["red cat", "wooden table", "sitting on"]

Example:
Input: "a photo of two blue cups on a white shelf"
Output: ["two blue cups", "white shelf", "on"]

Example:
Input: "a photo of a black car parked next to a red bicycle"
Output: ["black car", "red bicycle", "parked next to"]

Now decompose:
Input: "{prompt}"
Output:
```

B Proofs and Supporting Derivations

B.1 From Flow Matching to a Spatial Transition Density

Fix a prompt c and suppress it when unambiguous. Use the noise-increasing time convention $x_t = (1-t)x_0 + t\epsilon$, where x_0 is a data latent and $\epsilon \sim \mathcal{N}(0, I)$ is independent of x_0. The population optimum of the squared conditional flow-matching loss is

$$v_t^*(x) = \mathbb{E}[\epsilon - x_0 \mid x_t = x]. \tag{20}$$

This conditional-expectation construction connects the conditional and marginal vector fields (Lipman et al., 2023; Liu et al., 2023). Solving the conditional-mean equations and differentiating the Gaussian conditional density yield

$$\mathbb{E}[\epsilon \mid x_t = x] = x + (1-t)v_t^*(x), \qquad \nabla_x \log p_t(x) = -\frac{x + (1-t)v_t^*(x)}{t}, \quad 0 < t < 1. \tag{21}$$

To specify the direction of stochastic integration, let $s = 1 - t$ increase during generation, set $q_s = p_{1-s}$, and write $f_s = -v_{1-s}^*$ for the generating ODE velocity. For a diffusion coefficient a_s depending only on time, the SDE

$$dY_s = \left[f_s(Y_s) + \frac{a_s^2}{2} \nabla \log q_s(Y_s) \right] ds + a_s \, dW_s \tag{22}$$

has the same marginal probability path under compatible exact fields and the regularity conditions for the Fokker–Planck equation. Indeed,

$$-\nabla \cdot \left[(f_s + \tfrac{1}{2} a_s^2 \nabla \log q_s) q_s \right] + \tfrac{1}{2} a_s^2 \Delta q_s = -\nabla \cdot (f_s q_s). \tag{23}$$

This cancellation is the basis of ODE-to-SDE conversion in stochastic flow sampling (Liu et al., 2025).

Euler discretization at s_k, with $t_k = 1 - s_k \in (0, 1)$ and $h_k = s_{k+1} - s_k > 0$, gives

$$Y_{k+1} = \mu_{\theta,k}(Y_k) + \sigma_k \epsilon_k, \qquad \sigma_k^2 = a_{s_k}^2 h_k > 0, \tag{24}$$

$$\mu_{\theta,k}(y) = y - h_k \left[\gamma_k v_\theta(y, t_k) + \frac{a_{s_k}^2}{2t_k} y \right], \qquad \gamma_k = 1 + \frac{a_{s_k}^2 (1 - t_k)}{2t_k}. \tag{25}$$

The following derivation uses this Gaussian transition and also applies to fixed-variance Gaussian reverse transitions in diffusion models.

Return to the manuscript's discrete denoising index z_t, with $Y_k = z_{T-k}$ and $Y_{k+1} = z_{T-k-1}$. Here the integer label in z_t is distinct from the continuous noise time t_k above. For these density identities, assume a fixed latent dimension and spatial grid across consecutive transitions, so that u indexes

the corresponding output coordinates. For d_{ch} channels at location u, the log density summed over channels and its parameter gradient (score function) are

$$\ell^{\mathrm{dens}}_{\theta,t}(u) = -\frac{\|z_{t-1,u} - \mu_{\theta,t,u}\|^2}{2\sigma_t^2} - \frac{d_{\mathrm{ch}}}{2}\log(2\pi\sigma_t^2), \tag{26}$$

$$s_t(u) = \nabla_\theta \ell^{\mathrm{dens}}_{\theta,t}(u) = \frac{(D_\theta \mu_{\theta,t,u})^\top (z_{t-1,u} - \mu_{\theta,t,u})}{\sigma_t^2}. \tag{27}$$

The transition mean is differentiable in θ, the variance is strictly positive and parameter-independent, and the sampled input and output are held fixed when taking this score derivative. Diagonal transition covariance implies $\log p_\theta(z_{t-1} \mid z_t, c) = \sum_u \ell^{\mathrm{dens}}_{\theta,t}(u)$, while parameters may be shared across locations. We use ℓ^{dens} for the local log density and ℓ for the local quantity used in the reward optimization objective below.

B.2 Sensitivity of the Final Reward

Let R be the final-image reward and write $J(\theta) = \mathbb{E}[R(z_0, c)]$. Suppose the initial distribution and reward have no explicit parameter dependence and differentiation can be interchanged with integration. Factorizing the trajectory density gives

$$\nabla_\theta \mathbb{E}[R(z_0, c)] = \mathbb{E}\left[R \sum_{t,u} s_t(u)\right]. \tag{28}$$

A prompt-dependent baseline independent of the sampled trajectory can be subtracted because the transition scores have conditional mean zero. No derivative of R is required for this score-function identity.

Define $V^\theta_{t-1}(z, c) = \mathbb{E}[R(z_0, c) \mid z_{t-1} = z, c]$ as the expected reward from continuing generation from state z under the current policy, with $V^\theta_0 = R$. For fixed θ, take the expectation of V^θ_{t-1} over the next Gaussian transition:

$$W^\theta_t(m, c) = \int V^\theta_{t-1}(x, c)\, \varphi_{\sigma_t^2 I}(x - m)\, dx, \tag{29}$$

where $\varphi_{\sigma_t^2 I}$ is the centered Gaussian density. For bounded measurable R, V is bounded, and the derivative of the Gaussian kernel is locally dominated by an integrable function of x. Thus

$$\nabla_{m_u} W^\theta_t(m, c) = \int V^\theta_{t-1}(x, c) \frac{x_u - m_u}{\sigma_t^2} \varphi_{\sigma_t^2 I}(x - m)\, dx. \tag{30}$$

At $m = \mu_{\theta,t}(z_t, c)$, the Markov property and iterated expectation rewrite the expectation weighted by $V^\theta_{t-1}(z_{t-1}, c)$ above as an expectation weighted by the final reward. Together with Eq. (27), this gives the local reward-sensitivity identity

$$\mathbb{E}[R \nabla_\theta \ell^{\mathrm{dens}}_{\theta,t}(u) \mid z_t, c] = (D_\theta \mu_{\theta,t,u})^\top \left.\nabla_{m_u} W^\theta_t(m, c)\right|_{m = \mu_{\theta,t}(z_t,c)}. \tag{31}$$

This establishes Theorem 1. Unbounded rewards are also covered when the derivative is dominated by an integrable function. Taking an additional expectation over states requires the corresponding Jacobian-weighted terms to be integrable. The future policy defining V^θ is held fixed in this m-derivative. Its other parameter effects are accounted for by the remaining transition-score terms in Eq. (28). This derivation requires neither a differentiable reward nor a reward that decomposes as a sum over spatial positions.

If V^θ_{t-1} is continuously differentiable and both it and its gradient have at most polynomial growth, Gaussian integration by parts further gives

$$\nabla_{m_u} W^\theta_t(m, c) = \mathbb{E}_\epsilon[\nabla_u V^\theta_{t-1}(m + \sigma_t \epsilon, c)]. \tag{32}$$

Appropriate Gaussian-weighted weak-derivative conditions also yield this form. For example, suppose $V = V^\theta_{t-1}(\cdot, c)$ is locally Lipschitz and $|V(x)| + \|\nabla V(x)\| \le K(1 + \|x\|^p)$ almost everywhere for some $K > 0$ and $p \ge 0$. Continuous piecewise affine functions with finitely many pieces have at most linear growth and bounded weak gradients, and therefore satisfy this assumption. The weak

derivative agrees with the almost-everywhere gradient, and finite Gaussian polynomial moments make V, ∇V, and $(x-m)V$ integrable under the Gaussian weight. Write $\varphi_m(x) = \varphi_{\sigma_t^2 I}(x-m)$ and choose smooth cutoff functions $0 \le \chi_n \le 1$ that equal 1 inside the ball of radius n, vanish outside the ball of radius $2n$, and satisfy $\|\nabla\chi_n\| \le K_\chi/n$ for a constant K_χ independent of n. Apply the weak-derivative identity to the compactly supported function $\chi_n\varphi_m$ and let $n \to \infty$. The cutoff term vanishes by the integrability above, giving

$$\begin{aligned}\int \nabla_u V(x)\varphi_m(x)\,dx &= -\lim_{n\to\infty}\int V(x)\nabla_{x_u}\big(\chi_n(x)\varphi_m(x)\big)\,dx \\ &= \int V(x)\frac{x_u - m_u}{\sigma_t^2}\varphi_m(x)\,dx = \nabla_{m_u} W_t^\theta(m,c).\end{aligned}$$

The sensitivity can change with time because W_t^θ averages over the current transition and a different remaining generative process at each step. For deterministic generation, define the objective $J_{\mathrm{ODE}}(\theta) = \mathbb{E}[R(y_1^\theta, c)]$ with $\dot{y}_s = f_\theta(y_s, s)$ and a parameter-independent initial distribution. Assume that the flow and the final reward allow chain-rule differentiation, and that the resulting path derivative can be interchanged with expectation; continuously differentiable fields and reward with suitable integrable gradient domination are sufficient. The corresponding dependence is explicit in the adjoint $p_s = (D\Phi_{s\to 1})^\top \nabla R(y_1)$, where $\Phi_{s\to 1}$ is the remaining flow:

$$\dot{p}_s = -(D_y f_\theta)^\top p_s, \quad p_1 = \nabla R(y_1), \qquad \nabla_\theta J_{\mathrm{ODE}} = \mathbb{E}\int_0^1 (D_\theta f_\theta)^\top p_s\,ds. \tag{33}$$

These identities describe how the state and remaining generative process at each timestep affect reward, motivating spatial allocation that varies across timesteps.

B.3 Gradient Derivation for STAR Reward Allocation

Consider a batch of N rollout trajectories, each with T transitions and local position sets Ω_t. During backpropagation, treat the sampled input and output states, stored old log probabilities, advantages A_i, and allocation maps as constants. Here A_i denotes the advantage obtained by reward normalization or baseline subtraction as defined in the method. The derivation allows any fixed choice of group normalization or reward clipping used to compute A_i. STAR and the comparison objective using the scalar ratio use the same local quantity ℓ. For example, $\ell = \kappa_{\mathrm{ch}}\ell^{\mathrm{dens}}$, where $\kappa_{\mathrm{ch}} = 1$ for a channel sum and $\kappa_{\mathrm{ch}} = 1/d_{\mathrm{ch}}$ for a channel mean. This factor must be applied consistently in both current and stored old log probabilities, their spatial averages, and the local derivatives. Let

$$r_{i,t} = \exp(\bar{\ell}^i_{\theta,t} - \bar{\ell}^i_{\mathrm{old},t}), \qquad \rho_{i,t,u} = \mathrm{sg}(r_{i,t})\exp\big(\ell^i_{\theta,t}(u) - \mathrm{sg}(\ell^i_{\theta,t}(u))\big). \tag{34}$$

The second expression is algebraically equivalent to the ratio in the method. Let $\psi_A(r) = \min\{rA, \mathrm{clip}(r, 1-\varepsilon, 1+\varepsilon)A\}$. For comparison, define the reward optimization objective using the scalar ratio as

$$F_{\mathrm{scalar}}(\theta) = \frac{1}{NT}\sum_{i,t}\psi_{A_i}(r_{i,t}) - \beta\mathcal{D}(\theta). \tag{35}$$

This objective and STAR use the same sampled batch, advantages A_i, summation or averaging rules over channels and space, clipping width, and reference regularization. The spatial ratio's forward value and automatic-differentiation rule are

$$\rho_{i,t,u}\big|_{\mathrm{forward}} = r_{i,t}, \qquad \nabla_{\mathrm{AD}}\rho_{i,t,u} = r_{i,t}\nabla_\theta \ell^i_{\theta,t}(u). \tag{36}$$

Since $C_t^i(u) > 0$, the sign of $A_i C_t^i(u)$ is the sign of A_i. Clipping therefore retains the reward-gradient terms at all positions in a step or sets them all to zero. Away from clipping boundaries, write

$$\chi_{i,t} = \mathbf{1}\{A_i > 0,\ r_{i,t} < 1+\varepsilon\} + \mathbf{1}\{A_i < 0,\ r_{i,t} > 1-\varepsilon\}. \tag{37}$$

Zero advantages give zero contributions. The reward terms in F_{scalar} and STAR then have gradients

$$q_{i,t,u}(\theta) = \frac{\chi_{i,t} r_{i,t} A_i}{NT|\Omega_t|}\nabla_\theta \ell^i_{\theta,t}(u), \qquad g_{\mathrm{scalar}} = \sum_{i,t,u} q_{i,t,u}, \quad g_{\mathrm{STAR}} = \sum_{i,t,u} C_t^i(u) q_{i,t,u}. \tag{38}$$

At the rollout policy $r = 1$ and for a positive clipping width, these identities reduce to

$$q_{i,t,u} = \frac{A_i}{NT|\Omega_t|} \nabla_\theta \ell^i_{\theta,t}(u)\big|_{\theta=\theta_{\text{old}}},$$
$$g_{\text{scalar}} = \sum_{i,t,u} q_{i,t,u}, \qquad g_{\text{STAR}} = \sum_{i,t,u} C^i_t(u) q_{i,t,u}. \tag{39}$$

These identities prove Theorem 2. These identities hold exactly for the sampled batch and its advantages A_i. The decomposition also holds inside any fixed clipping region with the local contributions in Eq. (38). At nonsmooth boundaries, use the subgradient selected by the implementation.

STAR's gradients follow the specified stop-gradient operations, which preserve the forward values but change how gradients are propagated. For $\ell = \kappa_{\text{ch}} \ell^{\text{dens}}$, the scalar ratio is

$$r_{i,t} = \left(\frac{p_\theta(z^i_{t-1} \mid z^i_t, c)}{p_{\text{old}}(z^i_{t-1} \mid z^i_t, c)} \right)^{\kappa_{\text{ch}}/|\Omega_t|}. \tag{40}$$

This is the transition density ratio raised to the power $\kappa_{\text{ch}}/|\Omega_t|$. Summation or averaging over channels determines the ratio and its clipping status, so the comparison uses the same aggregation rule. Equation (38) follows directly by differentiating the clipped objective with the specified stop-gradient operations.

Holding allocation maps fixed during differentiation. If only the allocation maps are instead differentiated as $C_{i,t,u}(\theta)$, the gradient of the reward term contains the additional contribution

$$\Delta_{\text{route}} = \frac{1}{NT} \sum_{i,t} \frac{\psi_{A_i}(r_{i,t})}{|\Omega_t|} \sum_u \nabla_\theta C_{i,t,u}(\theta). \tag{41}$$

STAR holds the allocation map fixed, giving $\Delta_{\text{route}} = 0$. Sampled states and advantages are also held fixed, with gradients propagated through the local log probabilities of the current policy. When differentiating F_{scalar}, stored old log probabilities must also be treated as constants. In STAR, $\text{sg}(r)$ already blocks their gradient path through r.

Why the spatial ratio is necessary. Under the same differentiation assumptions, replace the specified spatial ratio by the ordinary scalar ratio $r_{i,t}$ while retaining $A_i C^i_t(u)$. Since $C^i_t(u) > 0$, this weight can be factored out of the clipped objective, giving

$$\frac{1}{|\Omega_t|} \sum_u \psi_{A_i C^i_t(u)}(r_{i,t}) = \bar{C}_{i,t} \psi_{A_i}(r_{i,t}), \qquad \bar{C}_{i,t} = \frac{1}{|\Omega_t|} \sum_u C^i_t(u). \tag{42}$$

Thus this objective has the same forward value as STAR, but its ordinary gradient, including the same reference penalty, is

$$h_{\text{uniform}} = \sum_{i,t,u} \bar{C}_{i,t} q_{i,t,u} + q_{\text{ref}}, \qquad q_{\text{ref}} = -\beta \nabla_\theta \mathcal{D}. \tag{43}$$

Equation (36) makes STAR retain the allocation weight at each position during backpropagation. With the ordinary scalar ratio, multiplying the advantage A_i by the allocation map affects the gradient only through the spatial mean of the map at each trajectory and timestep.

B.4 Bounding the Gradient Change from Reward Allocation

The min–max allocation rule, with $\lambda \geq 1$ and $\epsilon_{\text{map}} > 0$, ensures $1 \leq C_j \leq \lambda$ for every $j = (i, t, u)$; a constant heatmap gives $C_j = 1$. Set $m_\lambda = (\lambda + 1)/2$, $\delta_\lambda = (\lambda - 1)/(\lambda + 1)$, and $B = \sum_j \|q_j\|$. For any finite vectors q_j, the gradients in Eq. (39) satisfy

$$\left\| \frac{g_{\text{STAR}}}{m_\lambda} - g_{\text{scalar}} \right\| \leq \delta_\lambda B. \tag{44}$$

To prove this deterministic bound, expand the difference and apply the triangle inequality:

$$\frac{g_{\mathrm{STAR}}}{m_\lambda} - g_{\mathrm{scalar}} = \sum_j \left(\frac{C_j}{m_\lambda} - 1\right) q_j, \tag{45}$$

$$\left\| \frac{g_{\mathrm{STAR}}}{m_\lambda} - g_{\mathrm{scalar}} \right\| \le \sum_j \left| \frac{C_j}{m_\lambda} - 1 \right| \|q_j\| \le \delta_\lambda \sum_j \|q_j\|. \tag{46}$$

This gives a deterministic bound on the change in the gradient of the reward term for each sampled batch. It also implies $\|g_{\mathrm{STAR}} - g_{\mathrm{scalar}}\| \le (\lambda - 1) \sum_j \|q_j\|$, and exact baseline recovery at $\lambda = 1$.

The scaling factor m_λ minimizes the worst-case coefficient deviation $\max_{C \in [1,\lambda]} |C/m - 1|$ over $m > 0$. For $m \in [1, \lambda]$, the maximum is $\max(1 - 1/m, \lambda/m - 1)$, which is minimized by equating its two terms; outside this interval, the deviation is at least its value at the corresponding boundary. The comparison scales the STAR gradient by $1/m_\lambda$. For $\lambda = 1.5$, the compared coefficients lie in $[0.8, 1.2]$ and the bound is $0.2 \sum_j \|q_j\|$.

Including reference regularization. For a differentiable reference regularizer $\mathcal{D}$ and $\beta \ge 0$, introduce one additional contribution $q_{\mathrm{ref}} = -\beta \nabla_\theta \mathcal{D}$ with coefficient $C_{\mathrm{ref}} = 1$. Let f be the full gradient of F_{scalar} and h the STAR direction including reference regularization. The same proof gives

$$\left\| \frac{h}{m_\lambda} - f \right\| \le \delta_\lambda \left(\sum_j \|q_j\| + \beta \|\nabla_\theta \mathcal{D}\| \right). \tag{47}$$

This proves Eq. (18) in Theorem 5. Both directions are evaluated at the same parameters and regularization coefficient.

B.5 Local Gains in Reward Optimization with STAR

We now prove the local optimization statements in Theorem 5. All directions and allocation coefficients below are evaluated on the same sampled batch and at the same parameter θ. We compare updates of the form $\theta + \eta h$ with the same step size η and reference regularization.

Reward optimization under gradient coherence. Let $F = F_{\mathrm{scalar}}$ be the reward optimization objective defined in Eq. (35), including any reference regularization, so $f = \nabla F$ at a differentiability point. Let B_{tot} be the sum of contribution norms, including the reference term if present. The bound implies

$$f^\top \frac{h}{m_\lambda} \ge \|f\| \big(\|f\| - \delta_\lambda B_{\mathrm{tot}} \big). \tag{48}$$

Under the assumptions of Theorem 3, $\|f\| > \delta_\lambda B_{\mathrm{tot}}$, so the right-hand side is strictly positive and $f^\top h > 0$. Differentiability gives $F(\theta + \eta h) - F(\theta) = \eta f^\top h + o(\eta) > 0$ for all sufficiently small $\eta > 0$. For a quantitative finite-step bound, additionally assume that F has an L-Lipschitz gradient on a neighborhood of the entire segment connecting the parameters before and after the update, with $L > 0$ and $\eta \ge 0$. Then

$$F\left(\theta + \eta \frac{h}{m_\lambda}\right) - F(\theta) \ge \eta \|f\| \big(\|f\| - \delta_\lambda B_{\mathrm{tot}} \big) - \frac{L\eta^2}{2} \big(\|f\| + \delta_\lambda B_{\mathrm{tot}} \big)^2. \tag{49}$$

The gain from spatial allocation. Write $n_t = |\Omega_t|$, $a_{i,t,u} = f^\top q_{i,t,u}$ and $\bar{a}_{i,t} = n_t^{-1} \sum_u a_{i,t,u}$. For any two arrays on the spatial grid, define their covariance over positions as $\mathrm{Cov}_u(x, y) = n_t^{-1} \sum_u (x_u - \bar{x})(y_u - \bar{y})$. Subtracting the direction in Eq. (43) from STAR gives

$$\begin{aligned} f^\top (h - h_{\mathrm{uniform}}) &= \sum_{i,t,u} \big(C_t^i(u) - \bar{C}_{i,t} \big) a_{i,t,u} \\ &= \sum_{i,t} n_t \, \mathrm{Cov}_u(C_t^i, a_{i,t,\cdot}) = \Gamma. \end{aligned} \tag{50}$$

The average is taken over spatial positions in the sampled batch. By differentiability at θ,

$$F(\theta + \eta h) - F(\theta + \eta h_{\text{uniform}}) = \eta\Gamma + o(\eta). \tag{51}$$

Under the assumptions of Theorem 4, $\Gamma > 0$, making this difference strictly positive for all sufficiently small $\eta > 0$. Together with the local increase in F, this proves Eq. (19).

To express alignment directly in terms of Eq. (9), denote the normalized attention score by

$$S_t^i(u) = \frac{H_t^i(u) - \min_v H_t^i(v)}{\max_v H_t^i(v) - \min_v H_t^i(v) + \epsilon_{\text{map}}}, \qquad C_t^i(u) = 1 + (\lambda - 1)S_t^i(u). \tag{52}$$

Here $\epsilon_{\text{map}} > 0$ is the numerical stabilizer denoted by ϵ in the method. The allocation gain becomes

$$\Gamma = (\lambda - 1) \sum_{i,t} n_t \operatorname{Cov}_u(S_t^i, a_{i,t,\cdot}). \tag{53}$$

Thus $\Gamma > 0$ means that allocation weights and local gradient projections onto f have positive aggregate spatial covariance, giving more weight to contributions that help increase F. Another sufficient assumption is that attention weights and local gradient contributions have the same ordering within each timestep: $(S_t^i(u) - S_t^i(v))(a_{i,t,u} - a_{i,t,v}) \geq 0$ for every pair u, v, with at least one strict pair and $\lambda > 1$. Indeed,

$$n_t \operatorname{Cov}_u(S_t^i, a_{i,t,\cdot}) = \frac{1}{2n_t} \sum_{u,v} \big(S_t^i(u) - S_t^i(v)\big)\big(a_{i,t,u} - a_{i,t,v}\big). \tag{54}$$

Theorem 4 permits individual ordering errors and negative covariances at some timesteps, provided the aggregate between-region contributions and within-region covariances satisfy its assumptions. Constant spatial maps or $\lambda = 1$ give $\Gamma = 0$. Positive alignment $\Gamma > 0$ guarantees the strict first-order gain above, while baseline recovery follows directly from the gradient identity.

What the comparison establishes. The uniform update replaces the weight at each position by the spatial mean of the map for the same trajectory and timestep, and uses the same reference regularization. This is the update obtained with the same allocation map and the ordinary scalar ratio, so the comparison measures the benefit of retaining the weight at each position during differentiation. Both updates use the same parameter step size η. For the original unscaled scalar direction f, writing $Q_{i,t} = \sum_u q_{i,t,u}$ gives

$$f^\top(h - f) = \sum_{i,t} (\bar{C}_{i,t} - 1) f^\top Q_{i,t} + \Gamma. \tag{55}$$

The right-hand side separates the contribution from changes in the mean weight at each step from the spatial allocation gain Γ. The coherence result of Theorem 3 ensures a local increase in F independently of that comparison.

B.6 Deriving Gradient Coherence and Attention Alignment

This section proves Theorems 3–4 under their stated assumptions and discusses the evidence supporting those assumptions.

B.6.1 Proof and Further Analysis of Gradient Coherence

Let v_j collect the local gradient contributions weighted by the advantages and the gradient contribution from reference regularization, so $f = \sum_j v_j$ and $B_{\text{tot}} = \sum_j \|v_j\|$.

Proof of Theorem 3. Under the stated decomposition $v_j = b_j d + e_j$, the triangle inequality gives $\|f\| \geq B_0 - E_0$ and $B_{\text{tot}} \leq B_0 + E_0$. Since $E_0 < B_0/\lambda$ and finite $\lambda \geq 1$, the first lower bound is positive. Hence

$$\frac{\|f\|}{B_{\text{tot}}} \geq \frac{B_0 - E_0}{B_0 + E_0} > \frac{\lambda - 1}{\lambda + 1} = \delta_\lambda. \tag{56}$$

Cross-multiplication shows that the final strict inequality is equivalent to $E_0/B_0 < 1/\lambda$, proving Eq. (16). □

Here d is the common direction, while e_j includes sampling variation, conflicts between objectives, and any reference-gradient component opposing d. For $\lambda = 1.5$, the theorem assumes $E_0/B_0 < 2/3$. This assumption describes a dominant common direction.

In high dimensions, projection analysis gives a less restrictive criterion for local ascent. For M orthogonal unit local gradient vectors and zero reference gradient, $\|f\|/B_{\text{tot}} = M^{-1/2}$. At $\lambda = 1.5$ and $M \geq 25$, $\|f\|/B_{\text{tot}} \leq \delta_\lambda$, while $f^\top h = \sum_j C_j > 0$ for all positive weights. Thus STAR remains an ascent direction for F in this orthogonal case. A weaker sufficient assumption need only bound the negative projections of local gradients onto f. Define $N_- = \sum_{i,t,u}[-f^\top q_{i,t,u}]_+$, where $[x]_+ = \max(x, 0)$. Since $0 \leq C_t^i(u) - 1 \leq \lambda - 1$,

$$\begin{aligned} f^\top h &= \|f\|^2 + \sum_{i,t,u}(C_t^i(u) - 1) f^\top q_{i,t,u} \\ &\geq \|f\|^2 - (\lambda - 1)N_-. \end{aligned} \tag{57}$$

Consequently, $(\lambda - 1)N_- < \|f\|^2$ also suffices for a local increase in F and covers the orthogonal example. The reference contribution is accounted for in f, while N_- sums the negative projections of the local gradients from the reward term. The inner products of f with all contributions sum to $\|f\|^2$, while their absolute values sum to at most $\|f\|B_{\text{tot}}$. Hence $N_- \leq (\|f\|B_{\text{tot}} - \|f\|^2)/2$, so the coherence inequality $\|f\| > \delta_\lambda B_{\text{tot}}$ implies this weaker assumption; the orthogonal example illustrates its coverage of additional gradient structures.

B.6.2 From Semantic-Region Separation to Attention Alignment

For timestep t of trajectory i, write $b = (i, t)$ and $n_b = |\Omega_t|$ for its number of spatial positions. Denote the heatmap by $H_{b,u}$ and the local gradient contribution along f by $a_{b,u} = f^\top q_{i,t,u}$. Let $d_b = \max_u H_{b,u} - \min_u H_{b,u} + \epsilon_{\text{map}} > 0$, so that $w_b = n_b/d_b$ in Theorem 4. The exact allocation rule yields

$$\Gamma = (\lambda - 1) \sum_b \frac{n_b}{d_b} \operatorname{Cov}_u(H_b, a_b). \tag{58}$$

Covariance uses uniform averaging over the finite spatial grid. For $\lambda > 1$, the allocation preserves the sign of the spatial covariance within each trajectory and timestep; a spatially constant heatmap gives zero covariance. Normalization factors across trajectories and timesteps change the relative weights of positive and negative covariances in the sum.

For the binary partition $Z_{b,u} \in \{0, 1\}$ in Theorem 4, let $Z_b = 1$ denote the prompt-relevant region, with fraction $p_b \in (0, 1)$. The partition is specified from prompt semantics and image content before examining a_b. Define the differences between the means in the prompt-relevant region and the other region as $\Delta H_b = \mathbb{E}_u[H_b \mid Z_b = 1] - \mathbb{E}_u[H_b \mid Z_b = 0]$ and $\Delta a_b = \mathbb{E}_u[a_b \mid Z_b = 1] - \mathbb{E}_u[a_b \mid Z_b = 0]$. Let $P_b = p_b(1-p_b)\Delta H_b \Delta a_b$ and $U_b = p_b \operatorname{Cov}_u(H_b, a_b \mid Z_b = 1) + (1-p_b)\operatorname{Cov}_u(H_b, a_b \mid Z_b = 0)$. Thus $\mathcal{P} = \sum_b (n_b/d_b)P_b$ and $\mathcal{U} = \sum_b (n_b/d_b)U_b$ are the weighted sums of the contributions from between-region mean differences and the within-region covariances, respectively.

Proof of Theorem 4. Decomposing the finite spatial covariance into within-region and between-region components gives $\operatorname{Cov}_u(H_b, a_b) = P_b + U_b$. Substituting into Eq. (58) and using the hypotheses $\mathcal{P} > 0, \mathcal{U} \geq -\zeta\mathcal{P}$, $0 \leq \zeta < 1$, and $\lambda > 1$ gives

$$\Gamma = (\lambda - 1) \sum_b \frac{n_b}{d_b}(P_b + U_b) \geq (\lambda - 1)(1 - \zeta)\mathcal{P} > 0. \tag{59}$$

This proves Eq. (17). □

The model uses ΔH_b to describe how attention localizes the relevant region and Δa_b to describe its mean gradient contribution to increasing F relative to the other region. Positive mean differences in both quantities give a positive between-region contribution. Within-region covariances may be negative, provided their weighted sum does not offset the positive between-region contribution. For rewards based on objects and text, named entities and requested text provide a basis for this region partition.

B.6.3 Mechanistic and Empirical Evidence

At one attention head, let k range over the attended tokens. Holding attention weights fixed gives

$$o_u = \sum_k \alpha_{uk} v_k, \qquad \frac{\partial o_u}{\partial v_k} = \alpha_{uk} I. \tag{60}$$

With these weights fixed, α_{uk} determines how strongly the value vector v_k of a text token affects the output o_u at an image position. The direction of the local gradient contribution also depends on downstream Jacobians, value-vector directions, gradient paths through other parameters, and the sign of the advantage A_i. Reference regularization also affects its projection onto f. Theorem 4 further characterizes attention–gradient alignment through between-region mean differences and within-region covariances.

DAAM uses noun segmentation and human judgments in Stable Diffusion to evaluate how aggregated cross-attention maps locate image regions corresponding to words (Tang et al., 2023). Attend-and-Excite shows that attention-guided latent updates during inference can reduce omitted subjects and improve text-image alignment in its tested Stable Diffusion setting (Chefer et al., 2023). These studies support the localization premise and the relevance of attention to semantic generation.

Our existing layer-selection and heatmap-construction ablations in Figures 3b and 3a show how attention layer selection and allocation map construction affect training performance. The derivations establish coherence and alignment under the stated assumptions, while the existing experiments provide evidence about the resulting method's performance.

C Model License Statement

Stable Diffusion 3.5 Medium is used under the Stability AI Community License for academic research purposes. The authors confirm that the use of this model is limited to academic research purposes and has not been used for any commercial activities.