

GETTING SPACE RIGHT: AGENTIC REWARDS FOR IMAGE GENERATION

Xiaofeng Tan^{1,2*} Jiaxin Lin² Haobo Yuan³ Tiancheng Shen³ Zihan Su⁴
 Rui Chen² Hongsong Wang^{1†} Ming-Hsuan Yang³

¹ Southeast University ² Tencent HY ³ UC Merced ⁴ Tsinghua University

ABSTRACT

Reinforcement learning (RL) is promising for spatial text-to-image generation, but existing work largely focuses on policy optimization, assuming accurate rewards and effective online rollouts. We observe in-domain gains with out-of-domain degradation, exposing two limitations: unreliable spatial rewards in open-world settings and rollouts lacking basic-to-complex progression. We propose a curriculum-guided RL framework. First, we introduce SPATIALCRITIC, an agentic reward model equipped with a spatial perception harness that combines reusable verification guidance with specialized vision tools. We then construct spatial preference datasets through targeted counterfactual perturbations and perform agent–harness alignment to improve tool selection and evidence integration efficiently. Second, curriculum learning organizes shared foundational capabilities and their compositional dependencies in a capability graph for mastery-guided rollouts, periodic review, and targeted remediation. On SpaRW-Eval, SPATIALCRITIC achieves 95.1% verification accuracy, outperforming 320B GLM-5.3-Flash with a 40× smaller 8B backbone and raising average baseline scores from 78.4 to 83.9 on SD3.5-Medium and 83.2 to 86.5 on BAGEL. See the [project page](#).

1 INTRODUCTION

Enhancing spatial intelligent in text-to-image (T2I) generation, i.e., generating accurate spatial relationships faithful to text prompts (Ghosh et al., 2023; Wang et al., 2026c), remains a critical and challenging task in T2I post-training. Recently, reinforcement learning (RL) has emerged as a prevailing paradigm for aligning pre-trained models with task-specific goals.

RL aligns a T2I model by generating images from training prompts, evaluating rewards, and updating the policy. Thus, its effectiveness depends on three aspects: (1) rollout organization, (2) reward accuracy, and (3) policy optimization. Existing work mainly advances the third (Black et al., 2024; Clark et al., 2024; Liu et al., 2025a; Zheng et al., 2026; Tan et al., 2026a;b), often assuming that rollouts support effective learning and rewards provide reliable feedback. This leaves two questions: (Q1) Are reward models accurate and robust enough in open-world scenarios? (Q2) Can rollout organization support progressive learning from basic spatial capabilities to complex spatial compositions?

Unfortunately, these assumptions fail in spatial text-to-image generation: RL-aligned models improve steadily on in-domain (ID) benchmarks but *generalize poorly out of domain (OOD)*, eventually falling below the pre-trained model (Fig. 1). This failure stems from repurposing benchmarks such as GenEval as training prompts and their rule-based scoring pipelines as on-line rewards, exposing two critical flaws: (1) inaccurate and non-robust rewards, as benchmark-specific scoring generalizes poorly to open-world scenarios, especially under object occlusion (Tang et al., 2026); and (2) rollouts targeting complex compositions without progressive learning. Reused benchmark prompts emphasize joint constraint satisfaction, neglecting fundamental spatial capabilities and their progressive composition into complex ones.

To overcome the brittleness of rule-based rewards, recent efforts pursue learned reward modeling across two paradigms: *regression-based models* that map dual-encoder or VLM representations

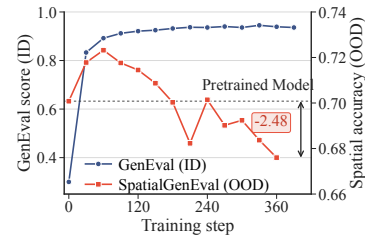


Figure 1: **Generalization collapse.** ID accuracy improves while OOD performance falls.

*Work done at Tencent HY. †Corresponding author.

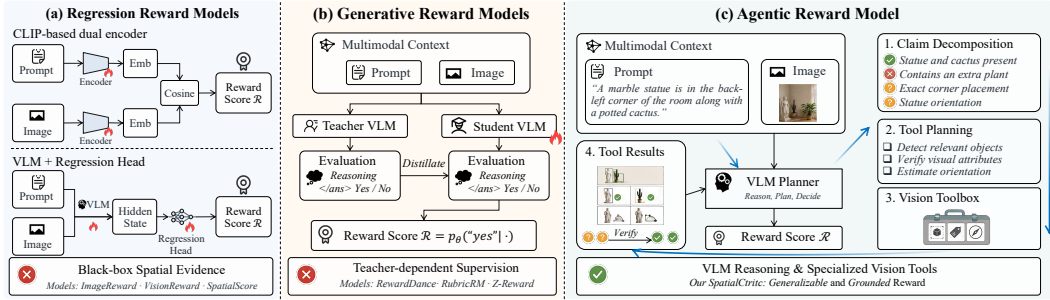


Figure 2: **Comparison of reward-modeling paradigms.** (a) Regression-based scoring; (b) teacher-guided generative evaluation; and (c) our agentic reward model, SPATIALCRITIC.

into scalar preference scores (Xu et al., 2026; Tang et al., 2026), and *generative models* that emulate deliberative evaluation via reasoning rationales or dynamic rubrics (Wang et al., 2026a; Kan et al., 2026), as shown in Fig. 2. However, neither ensures reliable spatial verification. Regression models treat foundation models as black-box feature extractors, underutilizing native reasoning and geometric depth priors. Meanwhile, generative models fundamentally rely on ungrounded teacher generation, and struggle to explicitly and reliably resolve geometric quantities such as occlusion and orientation. Beyond reward modeling, progressive capability learning remains largely unexplored. *These gaps motivate us to combine a reward model leveraging vision foundation models for verification with a progressive curriculum, jointly ensuring accuracy and generalizability.*

Contributions. We propose a curriculum-guided RL paradigm for spatial T2I post-training, comprising SPATIALCRITIC, an agentic reward model that leverages vision foundation models for robust supervision, and a progressive curriculum from basic spatial capabilities to complex compositions.

First, we design an agentic reward model, SPATIALCRITIC, that harnesses specialized vision foundation models for grounded spatial verification, optimized via Reinforcement Learning with Verifiable Rewards (RLVR). Its development comprises three components: (1) *Agent with a spatial perception harness.* To enhance the agent’s spatial perception capabilities, we design an agent equipped with a spatial harness that provides reusable verification guidance and executes specialized perceptual tools. Through this harness, the agent decomposes prompts into atomic claims and selectively queries specialized tools under uncertainty. (2) *Spatial preference dataset construction.* To enhance the agentic ability on the harness, the verifiable RL environment is necessary for post-training. We therefore construct a spatial preference dataset by counterfactually perturbing targeted spatial relations, such as occlusion, depth, and orientation, while keeping other visual content invariant. (3) *Agent–harness alignment via RLVR.* Building upon the agent and the preference dataset, we optimize the agent’s harness ability, such as tool-selection and evidence-integration, using verifiable outcome rewards and token-budget penalties to improve spatial judgment accuracy and efficiency.

Second, we introduce curriculum learning for spatial T2I post-training, which organizes policy roll-outs from basic spatial capabilities to complex compositions rather than directly optimizing over an unstructured mixture of complex tasks. Our key insight is that *higher-order spatial tasks involve different combinations of shared foundational capabilities*. For example, multi-object layout combines counting, and topology, while occlusion combines depth and topology, sharing topology capabilities. Accordingly, we introduce a Compositional Capability Graph (CCG) to capture these dependencies. Building on CCG, capability mastery guides progression, while periodic review identifies compositional weaknesses for targeted remediation. Together, these designs enable feedback-driven progression from basic capabilities to complex compositions.

Finally, *comprehensive experiments* validate accurate preference judgments, effective reward-guided image selection, and improved RL alignment. SPATIALCRITIC scores 95.1% on SpaRW-Eval. Curriculum RL raises baseline averages to 83.9 on SD3.5-Medium and 86.5 on BAGEL.

2 RELATED WORK

Reward Models for Visual Generation. Visual reward modeling for generation (Su et al., 2025; Fan et al., 2026; Su et al., 2026) has evolved from similarity heuristics (Radford et al., 2021) to

learned human preferences. Early dual encoders (Wu et al., 2023; Zhang et al., 2024) predict scalar preference scores via Bradley-Terry objectives. To overcome single-scalar bottlenecks, subsequent paradigms deployed multimodal VLMs to predict disentangled rewards across aesthetics, alignment, and visual artifacts (Xu et al., 2026; Liu et al., 2025b; Ma et al., 2025; Liu et al., 2026a; Tang et al., 2026). More recently, generative reward models emit dynamic rubrics, critique rationales, preference tokens, or calibrated score distributions (Wu et al., 2025b; Jin et al., 2026; Kan et al., 2026; Wang et al., 2026b;a). Closest to our work, SpatialReward (Zhou et al., 2026b) grounds decomposed spatial constraints with expert detectors and reasons over the resulting observations. Its verification procedure is nonetheless determined prior to observation, following a pre-defined routine per task category (Zhou et al., 2026a). We instead learn evidence acquisition as a per-claim decision over an open tool catalog, aligned to the harness under outcome supervision.

Spatial Intelligent Training Paradigm. Recent spatial intelligence studies reveal that learning complex spatial reasoning without grounded perceptual foundations induces severe shortcuts and optimization instability (Li et al., 2025; Liang et al., 2026; Yang et al., 2026). Consequently, structured spatial data and staged learning are explored in multimodal representation (Cai et al., 2026; Li et al., 2025; Liang et al., 2026) and embodied planning (Yang et al., 2026; Qian et al., 2026; Tan et al., 2025; Weng et al., 2026; Tan et al., 2026c;d). However, these works primarily target spatial understanding and embodied planning through data scaling, representation adaptation, or staged alignment. In contrast, our work centers on spatial T2I post-training, where a capability dependency graph and an absolute mastery probe guide online prompt scheduling, periodic review, and targeted remediation.

3 SPATIALCRITIC: TOOL-MEDIATED AGENTIC REWARD MODEL

3.1 AGENT WITH A SPATIAL PERCEPTION HARNESS

Problem Formulation. For spatial intelligence in text-to-image generation, the reward model requires to assess the spatial consistency of the generated image with the spatial constraints in the prompt, including object existence, attribute binding, counting, comparison, 2D position, background placement, distance, depth, orientation, topology, occlusion, multi-object layout, and 3D pose. Action, physical-plausibility, and image-quality checks provide auxiliary context. Preference judgments typically follow either pairwise comparison or pointwise scoring $S(\mathbf{I}, c)$ (Xu et al., 2023; Wang et al., 2025b). Our framework supports both, but primarily uses pairwise comparison. Given $\mathbf{x} = (\mathbf{I}_A, \mathbf{I}_B, c)$, SPATIALCRITIC predicts which image better satisfies the spatial constraints in c , yielding $\mathbf{I}_A \succ \mathbf{I}_B$. For score-wise setting, the preference is predicted by $S(\mathbf{I}_A, c) > S(\mathbf{I}_B, c)$.

Motivation: From Complex Spatial Relations to Atomic Verification. Spatial reward assessment benefits from complementary semantic and geometric capabilities. While VLMs excel at open-ended prompt comprehension, they fundamentally struggle with fine-grained spatial relations such as depth ordering and occlusion boundaries (Liu et al., 2026b). Conversely, fixed detector-based evaluators (Ghosh et al., 2023) rely on rigid and predefined heuristics and fail on complex open-world spatial relations. For example, consider the prompt “A palm tree is closer to the camera than a blue beach umbrella”. The VLM is hard to accurately locate the position of the tree and the umbrella, and estimates their depth ordering, while the detector-based evaluator can provide the precise information. In this case, the complex spatial relation is decomposed into three atomic verification tasks: checking the presence of both objects, the umbrella’s color, and their depth ordering.

Key Challenges. As discussed above, spatial reward modeling benefits from combining semantic reasoning with specialized perceptual evidence. Beyond integrating tool outputs within a predefined verification pipeline (Zhou et al., 2026b), we seek an agent that dynamically orchestrate vision foundation models for open-world spatial verification. This raises two complementary challenges: (1) *for high-frequency and canonical spatial relations*, how to make recurring verification patterns reliable and reusable? (2) *for complex and unseen spatial relations*, how to adapt these patterns to unfamiliar spatial compositions without being restricted to predefined tools pipeline or relying solely on unverified VLM judgments? Fundamentally, this requires balancing standardized execution on common and frequent relations with compositional flexibility on unseen and novel ones.

Spatial Perception Harness. The harness integrates a modular *perceptual toolkit*, covering grounding, counting, segmentation, depth, 3D geometry, orientation, and contextual checks, with structured

skill guidance, specifying evidence requirements and output interpretations. For claims flagged as uncertain during decomposition, the harness guides tool planning along two paths. (1) When a claim matches a *high-frequency relation*, the agent uses a *standardized skill* with a fixed local tool routine and evidence interpretation, addressing the first challenge. (2) Otherwise, for *complex or unfamiliar relations* requiring query-specific composition, the harness provides an open tool catalog with capability descriptions, guiding the agent to *compose and sequence targeted tool plans*. In Fig. 3(b), the plan grounds the instances, checks each dog’s orientation relative to the fountain, and aggregates these relations to verify the requested counts. This balances robust execution on high-frequency relations with flexible open-world composition. Toolkit details appear in App. C.

Agent Pipeline. As shown in Fig. 2(c), SPATIALCRITIC verifies spatial consistency through four interactive stages: (1) *Claim decomposition*: the VLM policy decomposes the prompt into atomic spatial claims and flags perceptual uncertainties across candidate images; (2) *Tool planning*: guided by the spatial perception harness, the agent identifies missing visual evidence and plans targeted tool calls; (3) *Tool execution*: the harness invokes the selected vision foundation models and collects structured perceptual observations; and (4) *Evidence integration*: the policy integrates these observations into its reasoning to resolve outstanding claims and produce the final preference verdict.

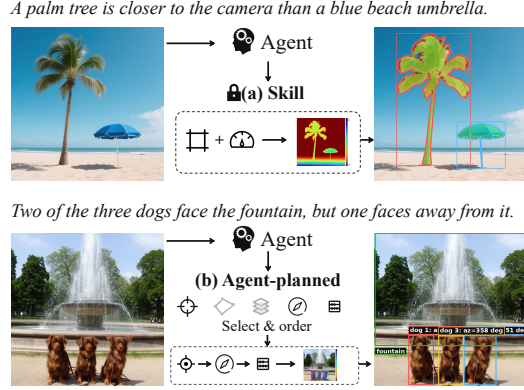


Figure 3: **Spatial verification.** (a) Fixed mask–depth routine; (b) agent-planned tool use.

3.2 VERIFIABLE SPATIAL PREFERENCE DATASET

Although the proposed spatial perception harness balances standardized execution for frequent tasks with flexibility for out-of-distribution cases, efficiency and accuracy remain limited by redundant or failed tool calls, conflicting tool outputs, and ambiguity among similar tools. Fundamentally, these issues reflect a *mismatch between the agent’s policy and the spatial perception harness*, requiring the agent to adapt its invocation, selection, and judgment policies to the harness.

This mismatch is difficult to resolve through harness design alone, as agent and tool capabilities vary across tasks and are difficult to align a priori. We therefore employ RLVR to learn adaptive tool selection and evidence integration through interaction, balancing verification accuracy and tool-use cost without predefined execution trajectories. To this end, we construct a spatial preference dataset of image pairs with explicit preference labels for judgment verification and reward computation.

Goal. To make these preferences effective reward signals for RL, the dataset should satisfy two requirements: *diversity* across spatial tasks and visual contexts, supporting generalizable use of the harness; and *accuracy* of the preference labels, providing reliable feedback for learning tool selection and evidence integration. We address diversity through *diverse task and scene coverage*, and accuracy through *constraint-targeted preference construction*, which combines explicit prompt constraints with text- and image-level validation.

Diverse task and scene coverage. Our construction covers a registry of 36 *scene settings* and a taxonomy of 13 *task categories*, with 3D pose held out from agent training. We balance target-category coverage and sample compatible scenes, supplying their objects, scales, and reference-frame rules to the LLM. Duplicate checks and multiple image generators further broaden textual and visual variation. App. B.2 lists the task and scene inventories, the complete generation system prompt, and the task-conditioned input fields.

Counterfactual preference construction & data pipeline. To obtain reliable spatial preference labels, we define unambiguous constraints and verify that the generated images form a valid contrast. We construct image preferences from reference prompts $\mathcal{D}_0 = \{c\}$ in four stages as shown in Fig. 4: (1) *Prompt-pair construction*. Targeted perturbations provide compositional negatives and counterfactual image pairs (Ma et al., 2023; Tang et al., 2026). Conditioned on the target category and scene constraints, DeepSeek-V4-Flash (DeepSeek-AI, 2026) jointly generates an unambiguous reference c^+ and three alternatives c^- , forming $\mathcal{D}_1 = \{(c^+, c^-)\}$ (Fig. 4(a)). Each alternative is intended to

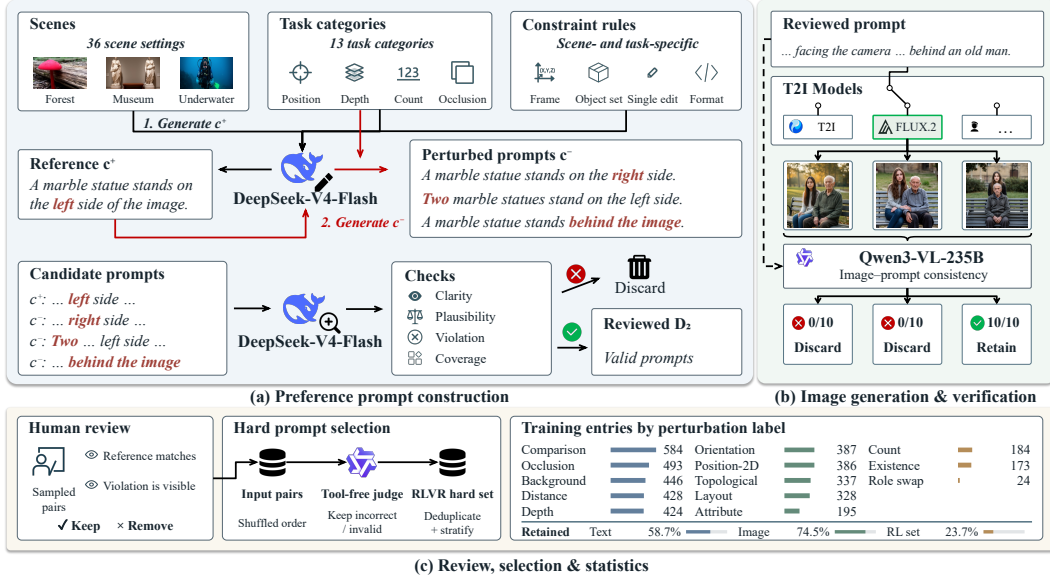


Figure 4: **Spatial preference dataset construction.** (a) Conditioned prompt generation, perturbation, and review. (b) Multiple image generation and VLM verification. (c) Human review, hard-set selection, and dataset statistics.

violate one constraint while preserving the others; at least one targets the designated category. The generation instructions are provided in App. B.2. (2) *Text-level validation.* A separate review call to DeepSeek-V4-Flash yields $\mathcal{D}_2 = \text{Filter}(\mathcal{D}_1)$ by filtering pairs for clarity, physical plausibility, and explicit violations, assigning perturbation categories, and checking target-category coverage. (3) *Image-level verification.* As illustrated in Fig. 4(b), we render both prompts of each pair in $\mathcal{D}_2$ with FLUX.2-dev (Black Forest Labs, 2025), Hunyuan-Image-3.0 (Tencent Hunyuan Foundation Model Team, 2025), and Qwen-Image (Wu et al., 2025a), using the same generator within a pair, and retain only pairs that pass a Qwen-VL-235B (Bai et al., 2025a) check of reference compliance and the intended violation, so that preferences reflect image content rather than textual changes alone. We apply best-of- n selection over the rendered candidates and keep the highest-scoring reference-compliant image as $\mathbf{I}^+$. (4) *Human inspection.* To mitigate automated verification noise, human annotators independently audit candidate pairs retained from image-level verification against identical criteria, confirming that $\mathbf{I}^+$ satisfies c^+ and $\mathbf{I}^-$ exhibits the intended violation; failing pairs are removed. The counterfactual prompt c^- is discarded after verification, yielding $\mathcal{D}_3 = \{(c^+, \mathbf{I}^+, \mathbf{I}^-)\}$, where $\mathbf{I}^+ \succ \mathbf{I}^-$ under c^+ . The inspection protocol and quality statistics are in App. B.1.

3.3 AGENT-HARNESS ALIGNMENT VIA RLVR

Building on the spatial perception harness and the spatial preference dataset, we align the agent’s tool-selection and evidence-integration policy with the harness through RLVR. We optimize the agent policy while keeping the harness and perceptual tools fixed, using verifiable outcome rewards and token-budget penalties to improve spatial judgment accuracy and token efficiency.

Hard set selection. The initial verification capability of the agent can induce an imbalance between positive and negative outcome rewards. To provide balanced feedback across spatial tasks, we identify hard where the tool-free base model fails. Combining this difficulty mining with category rebalancing prioritizes complex spatial relations over elementary recognition. Deduplication then yields a balanced training set as shown in Fig. 4(c), with complete protocols and counts in App. B.1. Based on this dataset, we define rewards for preference judgment and tool use.

Reward formulation. To improve token efficiency while preserving judgment accuracy, we combine an outcome reward with a token-budget penalty conditioned on correct judgments:

$$\mathcal{R}(\tau) = \mathcal{R}_{\text{out}}(\tau) + \mathbb{I}[\hat{y}(\tau) = y^{\text{gt}}] \text{clip}(\mathcal{R}_{\text{tok}}(\tau), -\kappa, 0). \quad (1)$$

where $\hat{y}(\tau)$ is the predicted preference, y^{gt} is its ground-truth label, and κ bounds the auxiliary penalty. We set $\mathcal{R}_{\text{out}}$ to +1, -1, and -1.5 for correct, incorrect, and invalid results, respectively.

The token penalty is $\mathcal{R}_{\text{tok}}(\tau) = -\alpha \max(0, n_{\text{tok}} - n_0)$, where n_{tok} counts agent-generated tokens, excluding input prompts and tool observations; n_0 is the free token budget and α penalizes each excess token. This objective encourages accurate judgments, targeted tool use, and concise evidence integration with fewer tokens and lower latency, as illustrated in Figs. 6, 7, and S1.

Policy optimization. With the dataset and reward, we optimize π_ϕ using GRPO (Shao et al., 2024). For each randomly ordered image pair, we sample G rollouts under $\pi_{\phi_{\text{old}}}$ following Sec. 3.1, then update the policy using group-relative reward advantages. Details appear in App. D.1.

4 CURRICULUM LEARNING FOR SPATIAL T2I POST-TRAINING

Building on SPATIALCRITIC, we introduce curriculum learning to address the lack of basic-to-complex progression. This requires (1) acquiring foundational spatial capabilities; and (2) composing them for combination. We use a *compositional capability graph* to define these capabilities and their dependencies, and a *progressive learning strategy* to schedule foundational and compositional training accordingly, guided by policy competence, review, and remediation.

4.1 COMPOSITIONAL CAPABILITY GRAPH

To guide foundational learning and composition, we examine how spatial capabilities relate. Our key insight is that *higher-order spatial tasks combine shared foundational capabilities in different ways*. CCG represents these foundations and dependencies beyond a single easy-to-hard ordering. Specifically, it is a directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with 13 capability nodes, where each edge (u, v) requires u to precede v in training. In Fig. 5, Layout and Occlusion share Topology; Layout additionally depends on Count and Position, and Occlusion on Depth. These prescribed dependencies provide an inductive bias for curriculum design, with the complete mapping in App. D.4. This graph facilitates training scheduling with a progressive learning strategy.

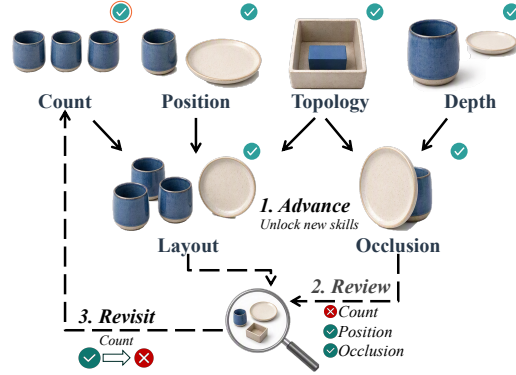


Figure 5: **CCG-guided progressive learning.** Solid arrows show capability dependencies.

4.2 PROGRESSIVE LEARNING WITH REVIEW AND REMEDIATION

Our progressive learning framework has two components in Fig. 5: (1) *forward progression*, which develops foundational capabilities and advances to dependent tasks based on mastery; and (2) *review and remediation*, which addresses forgetting and joint-constraint failures by reassessing learned capabilities individually and jointly.

Forward progression. Let $\mathcal{P}(v)$ denote the prerequisites of capability v in CCG. At assessment round t , we obtain the mean evaluation reward $\mathcal{S}_v^t$ for each assessed capability. To reduce sensitivity to outliers, we update its running score as $\bar{\mathcal{S}}_v^t = \text{Smooth}(\mathcal{S}_v^t, \bar{\mathcal{S}}_v^{t-1})$, where Smooth is an exponential moving average, detailed in App. D.4. Mastery requires $\bar{\mathcal{S}}_v \geq \tau$ and at least $N_{\min}$ image assessments. To prevent curriculum deadlock, unmastered capabilities permit fallback progression after $N_{\max}$ image assessments. We define the mastered and fallback sets as

$$\mathcal{M} = \underbrace{\{v \in \mathcal{V} \mid \bar{\mathcal{S}}_v \geq \tau, n_v \geq N_{\min}\}}_{\text{Mastered capabilities}}, \quad \mathcal{F} = \underbrace{\{v \in \mathcal{V} \mid v \notin \mathcal{M}, n_v \geq N_{\max}\}}_{\text{Unmastered at assessment limit}}. \quad (2)$$

where n_v counts image assessments for capability v . $\mathcal{M}$ contains mastered capabilities, while $\mathcal{F}$ contains those still unmastered at the assessment limit. Both sets permit downstream progression, but capabilities in $\mathcal{F}$ remain candidates for review and remediation. A successor capability v becomes *unlocked*, or eligible for training, when all its prerequisites belong to either set:

$$\text{Unlocked}(v) \iff \mathcal{P}(v) \subseteq \mathcal{M} \cup \mathcal{F}. \quad (3)$$

Fig. 5 illustrates progression from foundational to composite capabilities. Once Count, Position, Topology, and Depth meet the mastery criterion, they enter $\mathcal{M}$. Layout requires Count, Position, and

Topology, while Occlusion requires Topology and Depth. Both sets therefore lie in $\mathcal{M}$, unlocking Layout and Occlusion for training. The *Advance* phase shifts training toward these capabilities, with policy updates on the selected prompts using standard diffusion RL (Liu et al., 2025a).

Review and remediation. Forward progression alone does not ensure capability retention or compositional performance. (1) *Review* periodically evaluates capabilities in $\mathcal{M} \cup \mathcal{F}$ through individual and compositional tasks, updating $\mathcal{S}_v^t$ using the same evaluation and smoothing procedure. Capabilities failing the mastery criterion in Eq. (2) form the remediation set $\mathcal{C}$ and are removed from $\mathcal{M}$ via $\mathcal{M} \leftarrow \mathcal{M} \setminus \mathcal{C}$. (2) *Remediation* applies targeted training to $\mathcal{C}$ before forward progression resumes. Capabilities satisfying the mastery criterion in either stage enter $\mathcal{M}$ and leave $\mathcal{F}$; those remaining in $\mathcal{F}$ remain eligible for subsequent review and remediation. For example, jointly assessing Count, Position, and Occlusion in Fig. 5 yields $\mathcal{C} = \{\text{Count}\}$ when only Count fails the mastery criterion. The *Revisit* step then applies counting-specific training to improve performance under joint constraints. App. D.4 specifies review schedules and training budgets.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUPS

Implementation. We fine-tune Qwen3-VL-8B (Bai et al., 2025a) into SPATIALCRITIC with GRPO on 4,389 spatial preference pairs on 16 GPUs. We also evaluate Qwen3-VL-30B with the same spatial perception harness. We set $n_0 = 400$ free tokens and an excess-token penalty of $\alpha = 1.25 \times 10^{-3}$; see App. D.2 for reward coefficients. For T2I post-training, we use Flow-GRPO (Liu et al., 2025a) with SD3.5-Medium (Esser et al., 2024) and BAGEL (Deng et al., 2025). Our curriculum schedules prompts by capability dependencies and mastery; baselines use GenEval (Ghosh et al., 2023) or GenEval-2 (Kamath et al., 2025). Each Flow-GRPO iteration samples 24 images for each of 24 prompts, assessing four per group to track mastery with $\tau = 0.60$ and $N_{\min} = 200$. Further details appear in App. D.

Benchmarks and Datasets. We evaluate spatial preference judgments on SpaRW-Eval’s 5,572 pairs, reporting per-category and category-size-weighted overall accuracy. Construction follows Sec. 3.2, with evaluation details in App. B.3. We evaluate generation on T1IF-Bench (Wei et al., 2025), UniGenBench++ (Wang et al., 2025a), and SpatialGenEval (Wang et al., 2026c), whose 5,000, 600, and 1,230 prompts cover three instruction-following levels, 10 semantic dimensions, and 10 spatial sub-domains, respectively. The alignment table reports seven benchmark scores and their unweighted mean; ImageReward, PickScore, and HPSv2 separately measure preference quality. Following the benchmark’s recommendation to use the Qwen series, we adopt the stronger Qwen3-VL-235B to evaluate generation under the recommended setting.

Compared Methods. Reward baselines include Aesthetic Predictor V2.5 (discus0434, 2024), CLIP (Radford et al., 2021), ImageReward (Xu et al., 2023), PickScore (Kirstain et al., 2023), HPSv2 (Wu et al., 2023), HPSv3 (Ma et al., 2025), UnifiedReward (Wang et al., 2025c), and SpatialScore (Tang et al., 2026). We also compare zero-shot VLMs: Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a), Qwen3-Omni (Xu et al., 2025), DeepSeek (DeepSeek-AI, 2026), GLM (Z.ai, 2026), Doubao-Seed (Bytedance Seed, 2026), Gemini (Google DeepMind, 2026a;b), Qwen3.8 (Qiu et al., 2026), and Grok (xAI, 2026). For RL training, we compare three baselines, Flow-GRPO (Liu et al., 2025a), DiffusionNFT (Zheng et al., 2026), and AWM (Xue et al., 2026), with reward comparisons covering GenEval (Ghosh et al., 2023), GenEval-2 (Kamath et al., 2025), SpatialReward (Zhou et al., 2026b), and SpatialScore (Tang et al., 2026).

5.2 MAIN RESULTS

Accuracy of Preference Judgments. Tab. 1 shows that 8B SPATIALCRITIC achieves 95.1% verification accuracy, compared with 84.9% for Qwen3-VL-8B and 87.5% for SpatialScore. Agent-harness alignment raises accuracy from 92.9% to 95.1%, surpassing 320B GLM-5.3-Flash at 94.5%. The 30B variant reaches 95.5%, approaching Gemini 3.8 Flash at 96.7%.

Quality of Reward Signals. In Tab. 2, SPATIALCRITIC leads Best-of- N selection with 74.4% accuracy, exceeding GLM-5.3-Flash at 71.8% and the selection-free baseline at 70.8%. It leads in

Table 1: **Spatial preference accuracy (%) on SpaRW-Eval.** Per column, **bold** marks the best open-weight result, underline the second best, and ***bold italic underline*** the best across all models. Darker shading indicates higher accuracy. **Overall** subscripts: relative gain over their backbone VLM (Qwen3-VL-8B/-30B). †: released after July 2026.

Model	Params	Obj-Attr	Obj-Exist	Count	Bg	Comp	3D-Pose	Depth	Orient	Pos-2D	Topo	Layout	Occl	Dist	Overall†
<i>Open-weight: reward models</i>															
Aesthetic predictor	0.4B	54.3	51.4	46.8	46.7	45.0	47.8	48.1	42.4	51.9	47.5	46.7	55.4	51.3	51.6
CLIP	0.4B	72.7	85.8	84.4	52.0	66.7	76.5	52.8	62.0	50.6	63.3	66.7	63.1	59.0	75.3
PickScore	1.0B	82.3	86.4	85.3	49.3	65.0	79.3	72.2	78.3	50.6	72.5	73.3	61.5	71.8	81.1
HPSv2	1.0B	84.3	94.0	90.8	40.0	55.0	75.4	67.6	72.8	46.8	78.3	76.0	52.3	79.5	83.1
UnifiedReward	7B	85.7	91.4	89.9	61.3	73.3	87.1	63.9	71.7	57.0	84.2	81.3	66.2	89.7	85.7
SpatialScore	7B	89.1	98.4	92.7	98.7	78.3	61.7	76.9	84.8	<u>86.1</u>	95.0	90.7	80.0	<u>92.3</u>	87.5
<i>Open-weight: large vision-language models</i>															
Qwen3-VL-8B	8B	89.5	82.3	85.3	58.7	70.0	91.0	58.3	67.4	48.1	82.5	81.3	67.7	79.5	84.9
DeepSeek-V4.1-Flash†	552B	92.2	94.0	95.4	70.7	88.3	91.0	70.4	79.3	74.7	91.7	86.7	72.3	76.9	90.9
Qwen3-VL-235B	235B	95.3	94.6	93.6	88.0	81.7	95.7	68.5	76.1	67.1	95.0	90.7	70.8	92.3	93.3
Qwen3-Omni-30B	35B	95.3	95.5	95.4	69.3	88.3	95.1	74.1	81.5	68.4	91.7	89.3	80.0	89.7	93.5
Qwen3-VL-30B	30B	95.3	95.9	92.7	73.3	83.3	94.4	84.3	78.3	69.6	<u>95.8</u>	85.3	80.0	94.9	93.7
GLM-5.3-Flash†	320B	93.9	<u>96.7</u>	99.1	77.3	93.3	97.0	87.0	92.4	74.7	<u>95.8</u>	94.7	86.2	94.9	94.5
<i>Agentic Reward Model</i>															
SPATIALCRITIC	8B	94.2	93.6	90.8	<u>90.7</u>	88.3	92.6	<u>88.0</u>	82.6	79.7	93.3	<u>93.3</u>	<u>81.5</u>	89.7	92.9 _{+9.4%}
w/ AHA	8B	<u>96.0</u>	96.2	<u>98.2</u>	84.0	88.3	95.6	90.7	85.9	81.0	95.0	94.7	<u>81.5</u>	89.7	95.1 _{+12.0%}
SPATIALCRITIC	30B	96.6	96.0	95.4	89.3	95.0	<u>95.9</u>	<u>88.0</u>	<u>90.2</u>	91.1	<u>97.5</u>	90.7	78.5	<u>92.3</u>	95.3 _{+1.9%}
<i>Proprietary: large vision-language judges</i>															
Doubao-Seed-2.0-Lite	—	86.0	92.3	93.6	68.0	80.0	90.6	65.7	76.1	68.4	91.7	88.0	61.5	76.9	87.2
Gemini 3.5 Flash	—	90.3	93.1	92.7	70.7	65.0	88.3	56.5	76.1	63.3	90.0	81.3	52.3	76.9	88.4
Qwen3.8-Flash-Next†	—	97.0	97.9	99.1	84.0	81.7	96.6	76.9	94.6	79.7	96.7	97.3	76.9	92.3	95.9
Grok 4.6†	—	96.6	98.3	98.2	88.0	91.7	95.4	<u>93.5</u>	91.3	92.4	95.8	94.7	84.6	97.4	96.3
Doubao-Seed-2.0-Pro	—	96.9	98.5	97.2	93.3	96.7	96.6	84.3	95.7	83.5	95.8	97.3	83.1	100.0	96.6
Gemini 3.8 Flash†	—	96.4	98.9	98.2	86.7	95.0	97.4	91.7	87.0	97.5	95.0	94.7	87.7	94.9	96.7

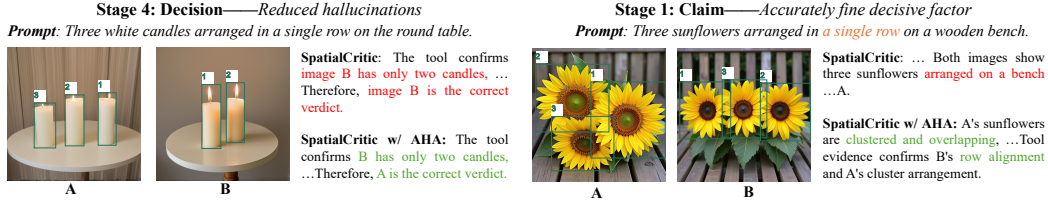


Figure 6: **Agent-harness alignment cases.** Left: alignment reduces final-judgment hallucinations, correctly selecting A because evidence shows only two candles in B. Right: it identifies the decisive single-row constraint and distinguishes B’s row from A’s cluster using tool evidence. App. Fig. S1 illustrates improved planning.

nine of ten categories, including all five geometry and spatial-relation categories. Position accuracy reaches 73.3%, versus 67.6% for the best baseline.

Effectiveness of Harness and Agent-Harness Alignment.

The harness provides verification guidance and grounded evidence, while alignment improves evidence integration to mitigate hallucinations and support reliable judgments. In Fig. 7, alignment rectifies the unaligned verdict error while reducing tokens from 67 to 49 and normalized latency from 1.000 to 0.653, yielding a more concise and accurate judgment. Figs. 6 and S1 illustrate targeted constraint identification, fewer irrelevant tool calls, and active verification of uncertain claims.

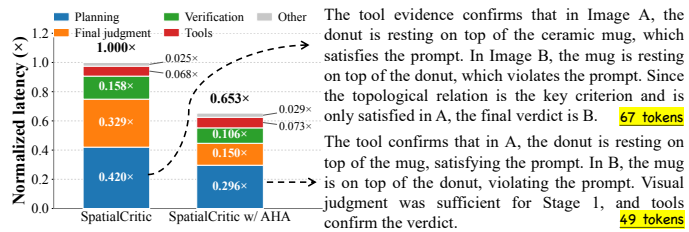


Figure 7: **Efficiency and decision correction.** Alignment reduces tokens and latency while rectifying judgment errors.

Table 2: **Reward-guided Best-of-N image generation.** Bold and underlined entries indicate the best and second-best results.

Selector	Attr.	Object	Comp.	Causal	Motion	Layout	Occl.	Orient.	Pos.	Prox.	Overall [†]
SD3.5-Medium (base)	87.89	94.63	38.62	80.57	72.68	72.68	59.35	65.77	66.59	68.86	70.76
Qwen2.5-VL-7B	88.10	94.40	38.50	80.60	72.20	72.50	<u>59.70</u>	65.30	65.60	68.50	70.54
PickScore	<u>89.43</u>	94.80	39.11	80.49	72.52	73.50	57.80	65.12	65.69	69.51	70.80
ImageReward	88.94	95.37	38.37	81.14	73.74	72.20	58.05	66.10	65.12	69.27	70.83
CLIP	89.19	94.80	<u>40.57</u>	81.14	72.76	72.85	59.51	63.90	65.45	69.59	70.98
HPSv2	89.11	94.88	39.84	80.49	73.33	72.20	58.70	<u>67.40</u>	65.53	69.59	71.11
HPSv3	88.94	94.07	40.24	80.81	74.23	73.33	58.05	65.85	66.42	<u>71.22</u>	71.32
UnifiedReward	88.70	94.88	39.35	<u>81.30</u>	74.96	72.93	58.21	66.18	<u>67.64</u>	70.00	71.41
GLM-5.3-Flash	89.27	94.39	39.92	<u>81.30</u>	<u>75.85</u>	<u>74.31</u>	57.72	66.42	<u>67.64</u>	70.73	<u>71.76</u>
SPATIALCRITIC	90.41	<u>94.96</u>	43.17	83.25	79.43	77.72	60.81	68.21	73.33	73.09	74.44

Table 3: **Reward-guided RL post-training.** Bold marks the best result within each backbone; CL denotes curriculum learning. [†]This non-curriculum setting uses GenEval-2 prompts without its benchmark-specific questions.

Method	Reward	TIIF-Bench			UniGenBench++		SpatialGenEval		Preference-model scores			Avg
		BR	AR	RR	2D	3D	Basic	Spatial	IR	PS	HPS	
SD3.5-Medium (base)	—	82.2	77.6	68.0	82.9	81.4	92.0	65.0	0.740	0.851	0.283	78.4
w/ Flow-GRPO	GenEval	57.8	53.7	45.4	43.6	38.7	75.9	48.1	-1.317	0.727	0.162	51.9
w/ DiffusionNFT	GenEval	68.5	57.7	56.8	66.9	44.1	65.9	53.7	-1.396	0.739	0.141	59.1
w/ AWM	GenEval	72.8	69.4	62.9	69.4	44.1	83.2	56.1	-1.116	0.757	0.165	65.4
w/ Flow-GRPO	SpatialReward	81.3	80.1	69.6	76.2	77.3	92.6	65.0	0.451	0.823	0.259	77.4
w/ Flow-GRPO	SpatialScore	83.4	80.9	69.4	92.1	78.7	93.7	67.3	0.885	0.857	0.306	80.8
w/ Flow-GRPO	GenEval-2	84.8	81.1	72.1	84.2	84.1	93.7	68.4	0.241	0.795	0.226	81.2
w/ DiffusionNFT	GenEval-2	85.1	83.7	73.4	89.0	82.8	94.0	68.3	0.588	0.823	0.271	82.3
w/ Flow-GRPO [†]	SPATIALCRITIC	83.3	80.3	69.8	93.3	85.3	94.4	67.0	0.857	0.855	0.304	81.9
w/ Flow-GRPO (CL)	SPATIALCRITIC	86.2	83.9	74.1	93.3	85.3	95.0	69.5	0.908	0.857	0.305	83.9
BAGEL (base)	—	86.4	82.5	73.8	91.8	82.8	94.2	70.9	0.753	0.855	0.283	83.2
w/ DiffusionNFT	GenEval	86.8	79.5	74.5	92.3	82.4	90.8	67.8	0.825	0.837	0.264	82.0
w/ DiffusionNFT	GenEval-2	87.6	80.7	77.3	90.6	87.3	92.2	70.0	0.764	0.826	0.257	83.7
w/ Flow-GRPO	GenEval	85.0	78.4	72.4	88.0	77.3	78.9	62.3	0.666	0.832	0.268	77.5
w/ Flow-GRPO	GenEval-2	86.9	80.9	74.9	90.8	84.0	80.5	65.6	0.624	0.817	0.233	80.5
w/ Flow-GRPO [†]	SPATIALCRITIC	86.1	82.2	74.8	93.2	82.7	93.6	73.1	0.775	0.890	0.294	83.7
w/ Flow-GRPO (CL)	SPATIALCRITIC	88.3	83.9	77.1	94.6	89.3	94.7	77.8	0.802	0.913	0.298	86.5

Effectiveness in RL Post-Training. Tab. 3 shows that SpatialCritic without CL raises averages from 78.4 to 81.9 on SD3.5-Medium and 83.2 to 83.7 on BAGEL, but does not consistently lead. On SD3.5-Medium, its average is below DiffusionNFT + GenEval-2 (82.3), and SpatialGenEval Spatial accuracy (67.0%) trails SpatialScore (67.3%) and Flow-GRPO + GenEval-2 (68.4%). Adding CL raises averages to 83.9 and 86.5 and Spatial accuracy to 69.5% and 77.8%, respectively. The ablation on curriculum dataset in App. A.3 confirms the effectiveness of curriculum scheduling.

6 CONCLUSION

In this work, we investigate spatial T2I post-training from the perspectives of reward accuracy and rollout organization. We identify two limitations beyond policy optimization: unreliable spatial rewards and rollouts that target complex compositions without progressive learning. First, we introduce SPATIALCRITIC, which combines a spatial perception harness with specialized vision tools for grounded preference judgments. Spatial preference data and agent-harness alignment via RLVR improve its tool selection and evidence integration. Second, we introduce curriculum learning based on a Compositional Capability Graph, guiding foundational capability development and subsequent composition through mastery-based progression, review, and remediation. Experiments demonstrate accurate preference judgments on SpaRW-Eval, effective reward signals for Best-of- N selection, and improved RL alignment on SD3.5-Medium and BAGEL.

Limitations. Our evaluation uses a fixed tool configuration, leaving transfer to updated or alternative tools unexplored. Additionally, robustness to temporary tool failures or missing outputs also remains untested. Future work will investigate cross-tool transfer and failure-aware fallback strategies.

AI USE STATEMENT

We used generative models to construct synthetic spatial prompts and images and to review candidate preference labels, as described in Sec. 3.2 and App. B.1. We used AI assistants for literature search and summarization to support related-work discovery. We also used AI tools to aid and polish paper writing, and for feedback on experimental analysis and inspection of data-processing code and metadata.

ETHICS STATEMENT

This work studies spatial verification and text-to-image generation using synthetic preference data and existing models and benchmarks. Synthetic data can inherit biases from its generators, and model-based verification can introduce systematic annotation errors. Human inspection provides an additional check but does not guarantee unbiased or error-free labels. Improved spatial fidelity may also make misleading synthetic images more convincing; spatial correctness should not be interpreted as evidence of authenticity or safety. Model and dataset reuse and redistribution remain subject to the applicable licenses and access conditions.

REPRODUCIBILITY STATEMENT

We describe the perceptual toolkit in App. C, preference-data construction and selection in App. B.1, and generation prompts in App. B.2. App. B.3 details the evaluation data and protocol. Training settings, reward coefficients, policy optimization, and curriculum parameters appear in App. D, D.2, D.1, and D.4. We will open-source our code upon acceptance.

REFERENCES

- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibao Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *CoRR*, abs/2511.21631, 2025a. doi: 10.48550/ARXIV.2511.21631. URL <https://doi.org/10.48550/arXiv.2511.21631>.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025b. doi: 10.48550/ARXIV.2502.13923. URL <https://doi.org/10.48550/arXiv.2502.13923>.
- 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*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=YCWjhGrJFD>.
- Black Forest Labs. FLUX.2: Frontier visual intelligence, 2025. URL <https://bfl.ai/blog/flux-2>.
- Bytedance Seed. Seed2.0 model card: Towards intelligence frontier for real-world complexity. *CoRR*, abs/2607.00248, 2026. doi: 10.48550/ARXIV.2607.00248. URL <https://doi.org/10.48550/arXiv.2607.00248>.
- Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, Tongxi Zhou, Jiaqi Li, Hui En Pang, Oscar Qian, Yukun Wei,

- Zhiqian Lin, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Xiangyu Fan, Hanming Deng, Lewei Lu, Liang Pan, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Scaling spatial intelligence with multimodal foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2026)*, 2026. doi: 10.48550/ARXIV.2511.13719. URL <https://doi.org/10.48550/arXiv.2511.13719>.
- Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=lvmSEVL19f>.
- DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026. URL <https://arxiv.org/abs/2606.19348>. Model variant: DeepSeek-V4-Flash.
- DeepSeek-AI. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression. *CoRR*, abs/2609.19969, 2026. doi: 10.48550/ARXIV.2609.19969. URL <https://doi.org/10.48550/arXiv.2609.19969>.
- Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025. URL <https://arxiv.org/abs/2505.14683>. Model: BAGEL.
- discus0434. Aesthetic Predictor V2.5. GitHub repository, 2024. URL <https://github.com/discus0434/aesthetic-predictor-v2-5>. SigLIP-based aesthetic score predictor; last commit 2024-12-18, accessed September 25, 2026.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning (ICML 2024)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024. doi: 10.48550/ARXIV.2403.03206. URL <https://proceedings.mlr.press/v235/esser24a.html>.
- Yuanting Fan, Jun Liu, Bin-Bin Gao, Xiaochen Chen, Yuhuan Lin, Zhewei Dai, Jiawei Zhan, and Chengjie Wang. UniDG: Universal defect generation via defect-context editing. *arXiv preprint arXiv:2604.08915*, 2026.
- Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL <https://openreview.net/forum?id=Wbr51vK331>.
- Google DeepMind. Gemini 3.5 Flash — model card, 2026a. URL <https://deepmind.google/models/model-cards/gemini-3-5-flash>.
- Google DeepMind. Gemini 3.8 Flash — model card, 2026b. URL <https://deepmind.google/models/model-cards/gemini-3-8-flash>.
- Xin Jin, Huanqia Cai, Zhen Li, Zechao Zhan, Dengyang Jiang, Aiming Hao, Yuming Jiang, Xiangpeng Yang, Chunle Guo, Peng Gao, Ming-Ming Cheng, and Steven C. H. Hoi. Z-Reward: Beyond scalar rewards by internalizing reasoning into score distributions. *CoRR*, abs/2606.09076, 2026. doi: 10.48550/ARXIV.2606.09076. URL <https://doi.org/10.48550/arXiv.2606.09076>.
- Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. GenEval 2: Addressing benchmark drift in text-to-image evaluation. *CoRR*, abs/2512.16853, 2025. doi: 10.48550/ARXIV.2512.16853. URL <https://doi.org/10.48550/arXiv.2512.16853>.
- Zijian Kan, Wei Wang, Long Luo, Bing Zhao, Xuan Ren, Weixu Qiao, Wenbo Li, Hu Wei, and Lin Qu. Rubricrm: Generative reward modeling via dynamic rubrics for image generation and editing. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*

-
- (EMNLP 2026), 2026. doi: 10.48550/ARXIV.2608.26956. URL <https://doi.org/10.48550/arXiv.2608.26956>.
- 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 Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/73aacd8b3b05b4b503d58310b523553c-Abstract-Conference.html.
- Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. *CoRR*, abs/2510.08531, 2025. doi: 10.48550/ARXIV.2510.08531. URL <https://doi.org/10.48550/arXiv.2510.08531>.
- Huizhi Liang, Yichao Shen, Yu Deng, Sicheng Xu, Zhiyuan Feng, Tong Zhang, Yaobo Liang, and Jiaolong Yang. Hispatial: Taming hierarchical 3d spatial understanding in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2026)*, 2026. doi: 10.48550/ARXIV.2603.25411. URL <https://doi.org/10.48550/arXiv.2603.25411>.
- 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. *CoRR*, abs/2505.05470, 2025a. doi: 10.48550/ARXIV.2505.05470. URL <https://doi.org/10.48550/arXiv.2505.05470>.
- Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025b. doi: 10.48550/ARXIV.2501.13918. URL https://papers.nips.cc/paper_files/paper/2025/hash/76227feb18ea0ee40bd15cf02c33e18e-Abstract-Conference.html.
- Yijun Liu, Jie Huang, Zeyue Xue, Yuming Li, Ruizhe He, Haoran Li, Shijia Ge, and Siming Fu. Hpsv3++: Scaling reward models across the full spectrum of diffusion model capabilities. *CoRR*, abs/2606.14657, 2026a. doi: 10.48550/ARXIV.2606.14657. URL <https://doi.org/10.48550/arXiv.2606.14657>.
- Yiqian Liu, Iuliia Kotseruba, and John K. Tsotsos. Disentangling pictorial cue understanding from language bias in VLMs via depth ordering task. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026b. URL <https://arxiv.org/abs/2607.01503>.
- Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 15086–15095. IEEE, 2025. doi: 10.1109/ICCV51701.2025.01400. URL <https://doi.org/10.1109/ICCV51701.2025.01400>.
- Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. CREPE: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10910–10921, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Ma_CREPE_Can_Vision-Language_Foundation_Models_Reason_Compositionally_CVPR_2023_paper.html.
- Kangan Qian, Chuchu Xie, Yang Zhong, Jingrui Pang, Siwen Jiao, Sicong Jiang, Zilin Huang, Yunlong Wang, Kun Jiang, Mengmeng Yang, Hao Ye, Guanghao Zhang, Hangjun Ye, Guang Chen, Long Chen, and Diange Yang. Xembodied: A foundation model with enhanced geometric and physical cues for large-scale embodied environments. *CoRR*, abs/2604.18484, 2026. doi: 10.48550/ARXIV.2604.18484. URL <https://doi.org/10.48550/arXiv.2604.18484>.

-
- Zihan Qiu, Zekun Wang, Xiao Li, Yanpeng Li, Yang Xu, Yixuan Wang, Huaqing Zhang, Rui Men, Bochao Mao, Chengruidong Zhang, Fan Zhou, Hao Luo, Haofeng Huang, Haoran Lian, Haoyan Huang, Hongqing Chen, Jianwei Zhang, Jing Xu, Junjie Wang, Langshi Chen, Liangyu Wang, Linlang Jiang, Man Yuan, Minmin Sun, Peng Jin, Siqi Zhang, Siyu Wang, Xingzhang Ren, Yakai Wang, Yi Zhang, Yiming Dong, Yizhong Cao, Yubo Ma, Yunfei Mao, Bo Zheng, and Dayiheng Liu. On the design of Qwen3.8-Next architecture: Evaluation, efficiency, and training stability. *CoRR*, abs/2608.30320, 2026. doi: 10.48550/ARXIV.2608.30320. URL <https://doi.org/10.48550/arXiv.2608.30320>.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL <http://proceedings.mlr.press/v139/radford21a.html>.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. URL <https://arxiv.org/abs/2402.03300>.
- Zihan Su, Xuerui Qiu, Hongbin Xu, Tangyu Jiang, Jun-hao Zhuang, Chun Yuan, Ming Li, Shengfeng He, and Fei Yu. Safe-sora: Safe text-to-video generation via graphical watermarking. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=pWzcPRIKp8>.
- Zihan Su, Hongyang Wei, Kangrui Cen, Yong Wang, Guanhua Chen, Chun Yuan, and Xiangxiang Chu. Generation enhances understanding in unified multimodal models via multi-representation generation. In *Forty-third International Conference on Machine Learning*, 2026. URL <https://openreview.net/forum?id=uYaNE2Cys2>.
- Xiaofeng Tan, Hongsong Wang, Xin Geng, and Pan Zhou. SoPo: Text-to-motion generation using semi-online preference optimization. In *Advances in Neural Information Processing Systems*, 2025. URL <https://arxiv.org/abs/2412.05095>.
- Xiaofeng Tan, Jun Liu, Yuanting Fan, Bin-Bin Gao, Xi Jiang, Xiaochen Chen, Jinlong Peng, Chengjie Wang, Hongsong Wang, and Feng Zheng. ConsistentRFT: reducing visual hallucinations in flow-based reinforcement fine-tuning. *CoRR*, abs/2602.03425, 2026a. doi: 10.48550/ARXIV.2602.03425. URL <https://doi.org/10.48550/arXiv.2602.03425>.
- Xiaofeng Tan, Jun Liu, Bin-Bin Gao, Yuanting Fan, Xi Jiang, Chengjie Wang, Hongsong Wang, and Feng Zheng. When policy entropy constraint fails: Preserving diversity in flow-based RLHF via perceptual entropy. In *Advances in Neural Information Processing Systems*, 2026b. URL <https://arxiv.org/abs/2605.12112>.
- Xiaofeng Tan, Wanjiang Weng, Haodong Lei, and Hongsong Wang. EasyTune: Efficient step-aware fine-tuning for diffusion-based motion generation. In *The Fourteenth International Conference on Learning Representations*, 2026c. URL <https://openreview.net/forum?id=FylEoIaAzQ>.
- Xiaofeng Tan, Wanjiang Weng, Hongsong Wang, Fang Zhao, Xin Geng, and Liang Wang. MotionRFT: Unified reinforcement fine-tuning for text-to-motion generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026d. URL <https://arxiv.org/abs/2603.27185>.
- Zhenyu Tang, Chaoran Feng, Yufan Deng, Jie Wu, Xiaojie Li, Rui Wang, Yunpeng Chen, and Daquan Zhou. Enhancing spatial understanding in image generation via reward modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2026)*, 2026. doi: 10.48550/ARXIV.2602.24233. URL <https://doi.org/10.48550/arXiv.2602.24233>.

-
- Tencent Hunyuan Foundation Model Team. HunyuanImage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025. URL <https://arxiv.org/abs/2509.23951>.
- Haozhe Wang, Cong Wei, Weiming Ren, Jiaming Liu, Fangzhen Lin, and Wenhui Chen. Rationalrewards: Reasoning rewards scale visual generation both training and test time. *CoRR*, abs/2604.11626, 2026a. doi: 10.48550/ARXIV.2604.11626. URL <https://doi.org/10.48550/arXiv.2604.11626>.
- Yibin Wang, Zhimin Li, Yuhang Zang, Jiazi Bu, Yujie Zhou, Yi Xin, Junjun He, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Unigenbench++: A unified semantic evaluation benchmark for text-to-image generation. *CoRR*, abs/2510.18701, 2025a. doi: 10.48550/ARXIV.2510.18701. URL <https://doi.org/10.48550/arXiv.2510.18701>.
- Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-grpo: Pairwise preference reward-based GRPO for stable text-to-image reinforcement learning. *CoRR*, abs/2508.20751, 2025b. doi: 10.48550/ARXIV.2508.20751. URL <https://doi.org/10.48550/arXiv.2508.20751>.
- Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *CoRR*, abs/2503.05236, 2025c. doi: 10.48550/ARXIV.2503.05236. URL <https://doi.org/10.48550/arXiv.2503.05236>.
- Yibin Wang, Yuhang Zang, Feng Han, Jiazi Bu, Yujie Zhou, Cheng Jin, and Jiaqi Wang. Unified personalized reward model for vision generation. *CoRR*, abs/2602.02380, 2026b. doi: 10.48550/ARXIV.2602.02380. URL <https://doi.org/10.48550/arXiv.2602.02380>.
- Zengbin Wang, Xuecai Hu, Yong Wang, Feng Xiong, Man Zhang, and Xiangxiang Chu. Everything in its place: Benchmarking spatial intelligence of text-to-image models. In *The Fourteenth International Conference on Learning Representations*, 2026c. URL <https://openreview.net/forum?id=ddFN3lWpIr>.
- Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, Bairui Li, and Lei Zhang. Tiif-bench: How does your T2I model follow your instructions? *CoRR*, abs/2506.02161, 2025. doi: 10.48550/ARXIV.2506.02161. URL <https://doi.org/10.48550/arXiv.2506.02161>.
- Wanjiang Weng, Xiaofeng Tan, Junbo Wang, Guo-Sen Xie, Pan Zhou, and Hongsong Wang. Re-Align: Text-to-motion generation via step-aware reward-guided alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. URL <https://arxiv.org/abs/2511.19217>.
- 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. *arXiv preprint arXiv:2508.02324*, 2025a. URL <https://arxiv.org/abs/2508.02324>.
- Jie Wu, Yu Gao, Zilyu Ye, Ming Li, Liang Li, Hanzhong Guo, Jie Liu, Zeyue Xue, Xiaoxia Hou, Wei Liu, Yan Zeng, and Weilin Huang. Rewarddance: Reward scaling in visual generation. *CoRR*, abs/2509.08826, 2025b. doi: 10.48550/ARXIV.2509.08826. URL <https://doi.org/10.48550/arXiv.2509.08826>.
- 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. *CoRR*, abs/2306.09341, 2023. doi: 10.48550/ARXIV.2306.09341. URL <https://doi.org/10.48550/arXiv.2306.09341>.
- xAI. Introducing Grok 4.6, 2026. URL <https://x.ai/news/grok-4-6>.

-
- 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 Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/33646ef0ed554145eab65f6250fab0c9-Abstract-Conference.html.
- Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Dan Zhang, Ming Ding, Xiaohan Zhang, Shiyu Huang, Xiaotao Gu, Minlie Huang, Jie Tang, and Yuxiao Dong. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 11269–11277. AAAI Press, 2026. doi: 10.1609/AAAI.V40I13.38107. URL <https://doi.org/10.1609/aaai.v40i13.38107>.
- Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-Omni technical report. *CoRR*, abs/2509.17765, 2025. doi: 10.48550/ARXIV.2509.17765. URL <https://doi.org/10.48550/arXiv.2509.17765>.
- Shuchen Xue, Chongjian Ge, Shilong Zhang, Yichen Li, and Zhi-Ming Ma. Advantage weighted matching: Aligning RL with pretraining in diffusion models. In *Forty-third International Conference on Machine Learning (ICML 2026)*, 2026. doi: 10.48550/ARXIV.2509.25050. URL <https://doi.org/10.48550/arXiv.2509.25050>.
- Yuyuan Yang, Junkun Hong, Hongrong Wang, Honghao Cai, Xunpeng Ren, Ge Wang, Mingcong Lei, Shenhao Yan, Jiahao Yang, Chengsi Yao, Xi Li, Yiming Zhao, Yatong Han, and Jinke Ren. SVLL: staged vision-language learning for physically grounded embodied task planning. *CoRR*, abs/2603.11563, 2026. doi: 10.48550/ARXIV.2603.11563. URL <https://doi.org/10.48550/arXiv.2603.11563>.
- Z.ai. GLM-5.3-Flash: Frontier Intelligence, Flash Cost, 2026. URL <https://z.ai/blog/glm-5.3-flash>.
- Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*, 2024. doi: 10.48550/ARXIV.2405.14705. URL <https://doi.org/10.48550/arXiv.2405.14705>.
- Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *The Fourteenth International Conference on Learning Representations (ICLR 2026)*, 2026. doi: 10.48550/ARXIV.2509.16117. URL <https://doi.org/10.48550/arXiv.2509.16117>.
- 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: Official implementation. Code repository, 2026a. URL <https://github.com/LivingFutureLab/SpatialReward>. Implementation of SpatialReward (CVPR 2026, pp. 647–658).
- 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, 2026b. doi: 10.48550/ARXIV.2603.22228. URL <https://doi.org/10.48550/arXiv.2603.22228>.