This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SpatialReward: Verifiable Spatial Reward Modeling for Fine-Grained Spatial Consistency in Text-to-Image Generation

Sashuai Zhou[1,2*], Qiang Zhou[2,3*], Junpeng Ma[2,4*], Yue Cao[2], Ruofan Hu[1], Ziang Zhang[1], Xiaoda Yang[1], Zhibin Wang[2], Jun Song[2,3†], Cheng Yu[2,3], Bo Zheng[2,3], Zhou Zhao[1†]

[1]Zhejiang University , [2]Alibaba Group Holding Limited, [3]Future Living Lab of Alibaba, [4]Fudan University

Abstract

Recent advances in text-to-image (T2I) generation via reinforcement learning (RL) have benefited from reward models that assess semantic alignment and visual quality. However, most existing reward models pay limited attention to fine-grained spatial relationships, often producing images that appear plausible overall yet contain inaccuracies in object positioning. In this work, we present ***SpatialReward****, a verifiable reward model explicitly designed to evaluate spatial layouts in generated images. SpatialReward adopts a multi-stage pipeline: a* Prompt Decomposer *extracts entities, attributes, and spatial metadata from free-form prompts; expert detectors provide accurate visual grounding of object positions and attributes; and a vision-language model applies chain-of-thought reasoning over grounded observations to assess complex spatial relations that are challenging for rule-based methods. To more comprehensively evaluate spatial relationships in generated images, we introduce* ***SpatRelBench****, a benchmark covering object attributes, orientation, inter-object relations, and rendered text placement. Experiments on Stable Diffusion and FLUX show that incorporating SpatialReward into RL training consistently improves spatial consistency and overall generation quality, with results aligned more closely to human judgments. These findings indicate that verifiable reward models hold considerable potential for enabling more accurate and controllable optimization in text-to-image generation models. The project page is available at:* https://github.com/LivingFutureLab/SpatialReward

1. Introduction

Recent advances in text-to-image generation [5, 16, 42, 56, 58, 70] have been increasingly fueled by reinforcement learning techniques [15, 18, 37, 43, 65, 83, 93]. Among these, GRPO-based approaches [37, 43, 83] have demonstrated notable effectiveness in improving the generative performance. A central component in these methods is the pre-trained reward model (RM) [21, 33, 73, 79, 80], which assesses both visual quality and semantic alignment, providing crucial feedback for policy-gradient optimization. Such reward-driven training has led to images that better match human preferences in global appearance and content.

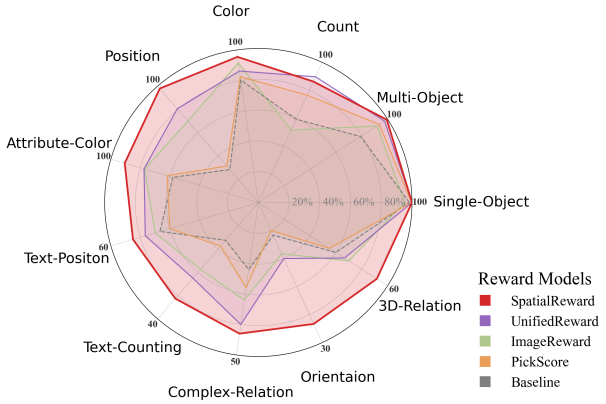


Figure 1. Performance comparison of SD3.5-M [17] optimized via RL using SpatialReward versus Baseline Rewards.

Despite these advances, existing RMs primarily focus on global semantics and coarse visual quality, while neglecting fine-grained spatial relationships. As a result, current T2I models still struggle to preserve spatial consistency among objects within a scene [1, 3, 7, 22, 30, 39]. These spatial inconsistencies degrade the realism of generated images and compromise faithful adherence to prompt semantics.

We hypothesize that further improvements in T2I spatial generation depend more on verifiable, spatially-aware reward models than on refinements to RL training strategies. In particular, existing approaches suffer from two major deficiencies: *Prompt-side rigidity*—Structured evaluation methods [21, 29, 30], such as Geneval [21], rely on fixed-format prompts and predefined object detectors. For instance, they can handle template-based inputs like "a photo of a purple backpack" but fail to generalize to

*Equal contribution.
†Corresponding author.

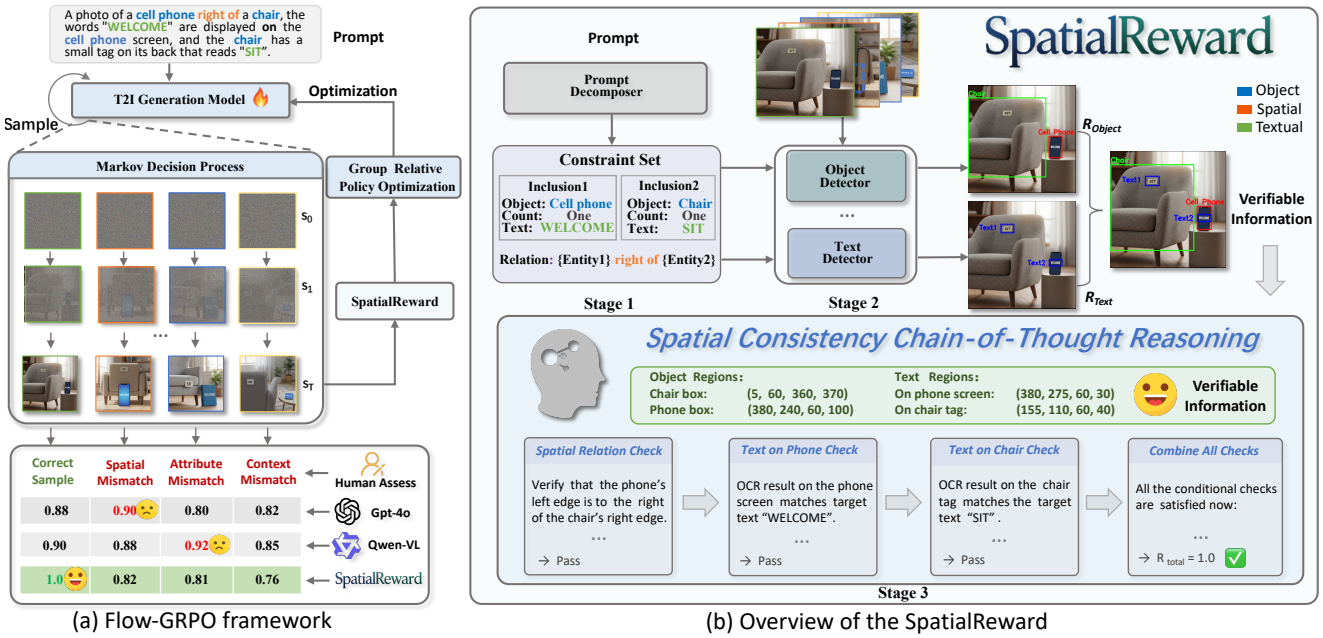


Figure 2. **Overall framework of our approach.** (a) Standard Flow-GRPO [43] reinforcement learning pipeline for text-to-image generation. (b) The proposed SpatialReward, which parses prompts into structured spatial and attribute constraints, verifies them on generated images via expert detection, and uses vision–language chain-of-thought reasoning to produce the final reward score.

complex, compositional prompts frequently found in open-ended generation tasks. *Vision-side overlooking*—Holistic evaluation methods based on CLIPScore [33, 55, 79, 80] or vision–language models [2, 72, 73, 82] can handle arbitrary prompts and capture global semantics, but without fine-grained spatial verification they fail to detect positional errors, often rewarding scenes that are visually plausible yet spatially wrong [10, 41, 50, 67, 89].

In this paper, we propose **SpatialReward**, a verifiable reward model explicitly built for fine-grained evaluation of spatial layouts. *Regarding the prompt side*, we introduce a Prompt Decomposer to extract core entities, attributes, and spatial relations from arbitrary free-form prompts. This normalization enables expert detectors to operate regardless of prompt format, thereby enriching the model's textual perception and supporting performance improvements in diverse generation scenarios [8, 63, 87, 91]. *Regarding the visual side*, we draw inspiration from the success of rule-based, verifiable rewards in logical reasoning [24, 45, 86], where explicit and checkable rewards have been shown to significantly improve complex inference performance. We extend this concept to visual spatial evaluation with a collaborative verification mechanism. First, object-related metadata extracted by the Prompt Decomposer is passed to open-set detectors [12, 13, 36, 44], which produce factual and highly verifiable information on object attributes and locations, thereby reducing hallucinations. Since spatial relationship assessment requires a sequence of reasoning steps that link detected facts to relative positions and overall layouts, we incorporate the verified grounding into a chain-of-thought (CoT) process [66, 68, 77] within a vision–language model. This explicit reasoning enables robust reward estimation for complex layouts and achieves greater flexibility than conventional rule-based checks.

To enrich the evaluation dimensions for spatial consistency in T2I models, we introduce **SpatRelBench**. This benchmark extends assessment beyond simple positional or color attributes to include object orientation, multi-object 3D positioning, complex spatial arrangements, and the placement of rendered text. We integrate our SpatialReward model into the Flow-GRPO [43] framework, an RL approach built upon GRPO, using Stable Diffusion [17] and FLUX [34] as base models. Experimental results show that our approach significantly enhances the spatial consistency of generated images.

Our main contributions are as follows:

- We present **SpatialReward**, a verifiable spatial reward model, combining prompt decomposition, expert detection, and chain-of-thought reasoning.
- We introduce **SpatRelBench**, a benchmark extending spatial evaluation to fine-grained object attributes, inter-object relations, and rendered-text placement.
- We conduct extensive experiments to verify that reinforcement learning with a verifiable reward model can significantly enhance spatial consistency in T2I generation.

2. Related Work

2.1. RL-based T2I Optimization

Diffusion-based T2I models, such as Stable Diffusion [17] and FLUX [34], have achieved high visual quality, while reinforcement learning has been shown to effectively improve generation performance across diverse prompts. Typical approaches train preference-based reward models to guide generation [4, 81], and recent work has further adapted GRPO-style optimization to text-to-image generation [43, 83, 88], improving text-image alignment and generalization. These advances are broadly relevant to multimodal applications [27, 47, 48, 64, 90] that require fine-grained understanding, semantic alignment, and knowledge-enhanced reasoning [28, 46, 49, 69, 85, 92]. This trend motivates our design of a verifiable reward model specialized for fine-grained spatial relationship assessment in generated images.

2.2. Reward Models for T2I Generation

Existing reward models can be broadly classified into structured methods [21, 29, 30] and holistic scorer methods [33, 55, 73, 79, 80, 82]. Structured approaches, such as Geneval [21], rely on fixed-format prompts and predefined object detectors to verify attributes and spatial arrangements, achieving high precision within narrow templates but generalizing poorly to free-form inputs. Holistic scorers include CLIP-based regressors such as ImageReward [80], PickScore [33], and HPSv2 [79], which fine-tune CLIP to predict human-preference scores. More recently, vision-language model (VLM) backbones [2, 35] have been adopted to capture richer semantic and spatial cues, with representative methods including VisionReward [82] and UnifiedReward [71, 73]. While VLM-based RMs improve flexibility over structured pipelines, they often overlook detailed spatial inconsistencies, highlighting the importance of verifiable reward models with robust spatial reasoning to ensure trustworthy evaluation across diverse prompts.

2.3. Benchmarks for T2I Evaluation

Early evaluation of image generation models relied on generic metrics such as Fréchet Inception Distance (FID) [25], Inception Score (IS) [59], and CLIPScore [33, 55, 79], which, while effective for assessing overall image quality, fall short in capturing fine-grained image–text alignment and complex spatial relations. To remedy these issues, specialised benchmarks have emerged. GenEval [21] and T2I-CompBench [29, 30] are object-centric benchmarks focusing on fundamental aspects such as object attributes and positions, typically using object detectors for automated scoring. More recent benchmarks [6, 19, 23, 51, 52] employ advanced VLMs [2, 82] to evaluate specific reasoning skills. However, current spatial benchmarks tend to ignore fine-grained inter-object relations, including orientation, 3D spatial positioning, and text placement. These gaps motivated us to develop SpatRelBench, a benchmark with richer spatial relation coverage.

3. SpatialReward

SpatialReward operates in three stages: Section 3.1 parses the prompt into structured spatial and attribute constraints; Section 3.2 verifies these constraints on the generated image using expert detection models, yielding verified rewards; Section 3.3 leverages a vision–language model with chain-of-thought reasoning to infer spatial relations and aggregate the results into the final reward score.

3.1. Prompt Decomposition

For RL-based training to generalize effectively, it is essential to handle prompts describing complex spatial arrangements of multiple objects [21, 22, 29]. The spatial reward in our framework aims to evaluate such cases by extracting accurate inter-object positional relations for subsequent verification using expert detection models. However, reliable detection requires that the prompt content be first decomposed into structured elements: separating each subject, its attributes, and the spatial relations involved.

Free-form prompts in the wild often contain irrelevant context or merge descriptions of different objects, introducing ambiguity that undermines detection accuracy. To mitigate this, we introduce a **Prompt Decomposer** $\mathcal{D}$ that transforms a free-form prompt P into structured constraint set $\mathcal{C}$:

$$\mathcal{C} = \mathcal{D}(P) = (\text{tag}, \mathcal{C}_{\text{inc}}, \mathcal{C}_{\text{exc}}), \tag{1}$$

where *tag* denotes the primary evaluation category (e.g., counting, orientation, spatial relation), $\mathcal{C}_{\text{inc}}$ contains inclusion constraints, and $\mathcal{C}_{\text{exc}}$ contains exclusion constraints. Each atomic constraint specifies object category, quantity, attributes, spatial relations, or textual inscriptions.

Inspired by metadata-driven evaluation frameworks [21, 29], we constructed a dataset of approximately 100k multi-object metadata instances, explicitly defining attributes, counts, spatial relations, and associated text for each subject. Using GPT-4o [53], we generated diverse natural language prompts from these metadatas, yielding (prompt, metadata) pairs for supervised training. We fine-tuned a Qwen2.5-VL-7B [2] model to accurately extract core meta-attributes from unrestricted prompts. This structured decomposition provides a reliable foundation for subsequent fine-grained image evaluation in our pipeline.

3.2. Fine-grained Verifiable Rewards

While the capabilities of VLMs are advancing rapidly, existing work [32, 41, 50, 67, 75, 89] has shown that even state-of-the-art models struggle with compositional text prompts involving multiple objects, attribute binding,

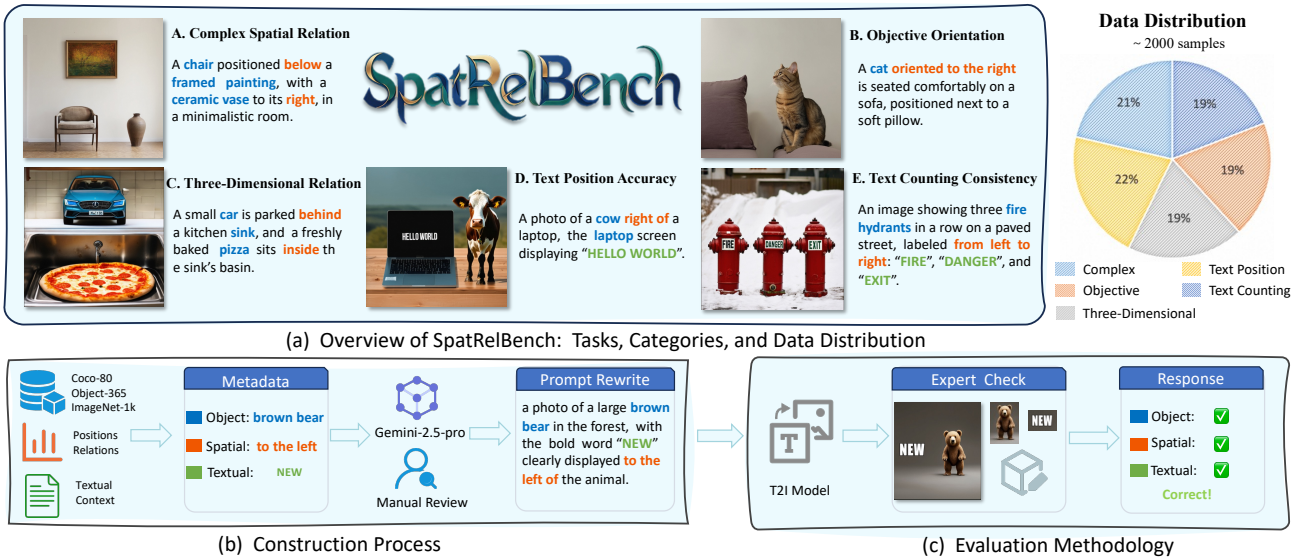


Figure 3. Overview of SpatRelBench, depicting benchmark tasks and their data distribution (a), the construction pipeline (b), and the evaluation methodology (c) designed to assess spatial relation understanding in text-to-image models.

spatial-action relationships, counting, and logical reasoning. Relying solely on VLMs for these tasks makes it difficult to obtain stable and verifiable objective reward scores. Fortunately, modern open-domain object detection [11, 38, 44, 57, 74] and Optical Character Recognition (OCR) models [13, 36, 76] demonstrate accuracy that significantly surpasses the judgmental capabilities of VLMs, providing objective scores that closely align with human evaluation standards. Leveraging the constraints extracted by the Prompt Decomposer in Section 3.1, we integrate these specialized detectors to perform precise, criterion-specific verification. This process yields a sub-reward for each positive constraint $c_i \in \mathcal{C}_{\text{inc}}$, providing a verifiable and quantitatively accurate signal for assessing spatial positions, object attributes, and other fine-grained relationships.

Object Attribute and Presence Reward. For each inclusion constraint $c \in \mathcal{C}_{\text{inc}}$ that refers to an object, we evaluate whether the generated image I satisfies the specified visual and spatial properties, including object category, color, target count, orientation, and depth ordering. These properties are extracted via specialized detection models [11, 44, 74, 84].

Given a target category in c, object detector $\mathcal{F}_{\text{det}}$ [11, 44] is applied to the generated image I, yielding candidate bounding boxes

$$D_c = \{(B_j, s_j)\}_{j=1}^{k}, \tag{2}$$

where B_j denotes a bounding box and $s_j \in [0, 1]$ its confidence score. Applying a confidence threshold τ_{det} produces the verified set of detections

$$\mathcal{B}_c = \{B_j \mid (B_j, s_j) \in D_c \ \wedge \ s_j \geq \tau_{\text{det}}\}, \tag{3}$$

whose cardinality $\hat{N}_c = |\mathcal{B}_c|$ forms the *presence reward*

$$\mathcal{R}_{\text{presence}}(c) = \mathbb{I}(\hat{N}_c > 0), \tag{4}$$

as well as the *count reward*

$$\mathcal{R}_{\text{count}}(c) = \exp\left(-\left|\hat{N}_c - N_c^*\right|\right), \tag{5}$$

in which N_c^* is the target count given in c.

Beyond category and quantity, object attributes are also verified. The *color reward* is defined as

$$\mathcal{R}_{\text{color}}(c) = \text{sim}_{\text{color}}(C_{\text{det}}, C^*), \tag{6}$$

where C_{det} is obtained via a CLIP-based [55] classifier that evaluates cropped object regions against prompt templates combining each candidate color with the object class name. The top-scoring color is then compared to the target C^* using $\text{sim}_{\text{color}}(\cdot)$. Similarly, the *orientation reward* assesses angular consistency as

$$\mathcal{R}_{\text{ori}}(c) = \mathbb{I}(|\theta_{\text{det}} - \theta^*| \leq \delta_\theta), \tag{7}$$

where the detected orientation θ_{det} is obtained from an orientation-sensitive model [74], θ^* is the target in c, and δ_θ specifies the tolerance.

For 3D spatial reasoning, the *depth reward* evaluates whether the relative depth ordering matches the target:

$$\mathcal{R}_{\text{depth}}(c) = \exp(-\left|d_{\text{rank}} - d_{\text{rank}}^*\right|), \tag{8}$$

where d_{rank} is the rank order inferred via monocular depth estimation [84], d^*_{rank} is the target ordering.

This formulation enforces comprehensive verification of every inclusion constraint $c \in \mathcal{C}_{\text{inc}}$ against the generated image, capturing both appearance fidelity and fine-grained spatial consistency.

Text Content and Localization Reward. For prompts that require rendering specific textual content within objects, the reward must jointly assess semantic correctness and spatial placement. Given a target object B_{obj} and required text T^*, a global OCR model $\mathcal{F}_{\text{ocr}}$ [13, 76] extracts a set of detected text–box pairs $\mathcal{T}_{\text{rec}} = \{(T'_j, B'_j)\}_{j=1}^{m}$, where B'_j and B_{obj} denote bounding boxes for detected text regions and the target object, respectively. The textual reward is defined by identifying the text instance that best matches T^* in content and is correctly localized within B_{obj}:

$$\mathcal{R}_{\text{text}}(T^*, B_{\text{obj}}) = \max_{(T'_j, B'_j) \in \mathcal{T}_{\text{rec}}} \left[\text{sim}(T^*, T'_j) \cdot \text{IoA}(B'_j, B_{\text{obj}}) \right], \tag{9}$$

where $\text{sim}(\cdot)$ measures normalized lexical similarity and IoA quantifies the degree of spatial containment:

$$\text{IoA}(B_{\text{text}}, B_{\text{obj}}) = \frac{\text{Area}(B_{\text{text}} \cap B_{\text{obj}})}{\text{Area}(B_{\text{text}})}, \tag{10}$$

High *text reward* is given only when generated text matches the target string and appears within the correct object bounding box, ensuring prompt fidelity in tasks involving embedded text.

3.3. Spatial Chain-of-Thought Reasoning

Although the fine-grained object and text rewards provide reliable verification for individual attributes, determining complex spatial relations between multiple entities requires higher-level reasoning [9, 22, 50]. Simple rule-based geometric checks often struggle with nuanced semantics (e.g., distinguishing "on" from "above") and with context-dependent layouts that cannot be resolved from geometry alone. To address this, we adopt Qwen2.5-VL [2] as our Chain-of-Thought reasoning backbone, using verified detection-based signals as grounding to reduce hallucination.

For the spatial relation between entities e_A and e_B, we construct a CoT prompt P_{CoT} comprising: (1) the target relation r, (2) their detected bounding boxes B_A and B_B, and (3) the set of attribute rewards for each object from previous stages $\{\mathcal{R}_{\text{pres}}, \ldots, \mathcal{R}_{\text{ori}}, \mathcal{R}_{\text{depth}}, \mathcal{R}_{\text{text}}\}$. By explicitly providing these verifiable signals, the VLM is guided to reason step-by-step: first interpreting each attribute reward relative to the bounding boxes, then performing geometric analysis, and finally inferring whether the relation r holds.

The output of $\mathcal{F}_{\text{vlm}}$ is restricted to a structured format containing the reasoning trace and a final score. We parse the score using a function $\mathcal{P}_{\text{score}}$ to yield the spatial-consistency reward:

$$\mathcal{R}_{\text{spatial}} = \mathcal{P}_{\text{score}}\big(\mathcal{F}_{\text{vlm}}(P_{\text{CoT}}(r, B_A, B_B, \text{attributes}))\big), \tag{11}$$

where attributes denotes the verified property scores (e.g., presence, orientation, depth, text) for both entities from earlier detection stages.

To enhance robustness and avoid overfitting to positive cases, we incorporate explicit penalties for satisfied exclusion constraints. For each inclusion constraint $c \in \mathcal{C}_{\text{inc}}$ and exclusion constraint $c \in \mathcal{C}_{\text{exc}}$, both derived from the earlier Prompt Decomposer, the CoT module produces $\mathcal{R}_{\text{spatial}}(c)$, and calculates the aggregated spatial score:

$$\mathcal{R}_{\text{total}} = \sum_{c \in \mathcal{C}_{\text{inc}}} \mathcal{R}^{+}_{\text{spatial}}(c) - \sum_{c \in \mathcal{C}_{\text{exc}}} \mathcal{R}^{-}_{\text{spatial}}(c), \tag{12}$$

This formulation rewards satisfaction of required relations while penalizing the presence of undesired ones, producing a spatial score that is both semantically informed and grounded in verifiable evidence.

4. SpatRelBench

To enable more fine-grained and comprehensive evaluation of spatial relationships in generated images, we introduce **SpatRelBench**, a benchmark specifically designed to assess spatial consistency in complex scenarios. The evaluation protocol covers five primary dimensions: (1) *Complex spatial relations*, which assess the relative positioning and arrangement of multiple objects within a scene; (2) *Object orientation*, which checks whether each object is depicted facing the correct direction; (3) *Three-dimensional relations*, which evaluate depth ordering and 3D layout consistency among objects; (4) *Text-position accuracy*, which verifies whether rendered text appears at the correct location relative to the associated object, using optical character recognition; and (5) *Text-counting consistency*, which determines whether the quantity of rendered text across multiple objects matches the prompt specification.

As illustrated in Fig. 3, we extend the evaluation category set beyond the COCO-80 [40] classes to include ImageNet-1k [14] and Objects365 [61] categories, covering both common objects and fine-grained subcategories. Prompts are generated using Gemini-2.5-Pro [31] and then manually validated to ensure both diversity and correctness. During evaluation, domain-specific expert models are employed to score each dimension. For every sub-task, a binary decision (*correct* or *incorrect*) is recorded, and the overall accuracy is computed by normalizing the number of correct judgments over the total number of requirements. The current release contains approximately 2,000 annotated

Table 1. **Quantitative comparison of T2I generation models aligned with different reward models.** Results are on GenEval(80-Obj) and SpatRelBench(1k-Obj), where parentheses indicate the number of object categories. S-Obj: Single object, T-Obj: Two objects, Cnt: Counting, Pos: Positions, Attr-C: Attribute (Color), P-Text: Position-Text OCR, C-Text: Counting-Text OCR, Cpx: Complex spatial relations, Ori: Orientation, 3DRel: 3D spatial relations, Overall: average score over all metrics in each dataset. **Bold** denotes the best score, and underline denotes the second best.

Reward Model	GenEval (80-Obj)							SpatRelBench (1k-Obj)					
	S-Obj.	T-Obj.	Cnt.	Color	Pos.	Attr-C.	**Overall**	P-Text.	C-Text.	Cpx.	Ori.	3DRel.	**Overall**
Proprietary T2I models													
GPT Image 1	0.99	0.92	0.85	0.92	0.75	0.61	0.84	0.50	0.22	0.53	0.15	0.45	0.37
Seedream 3.0	0.99	0.96	0.91	0.93	0.47	0.80	0.84	0.25	0.20	0.11	0.07	0.41	0.21
Qwen-Image	0.99	0.92	0.90	0.88	0.76	0.77	0.87	0.22	0.21	0.32	0.07	0.32	0.23
Based on Stable Diffusion 3.5													
SD3.5-M	0.98	0.79	0.59	0.80	0.28	0.58	0.67	0.40	0.13	0.22	0.07	0.36	0.23
+ TextOCR	0.99	0.87	0.59	0.81	0.33	0.58	0.70	0.48	0.24	0.23	0.09	0.38	0.28
+ PickScore	0.98	0.94	0.77	0.83	0.31	0.62	0.74	0.36	0.15	0.28	0.06	0.33	0.24
+ QwenVL	0.99	0.58	0.48	0.81	0.24	0.55	0.61	0.32	0.21	0.30	0.12	0.37	0.26
+ ImageReward	0.95	0.92	0.52	0.92	0.70	0.77	0.80	0.42	0.23	0.32	0.11	0.42	0.30
+ UnifiedReward	**1.00**	0.98	**0.90**	0.86	0.81	0.78	0.89	0.46	0.26	0.40	0.12	0.40	0.33
+ SpatialReward	**1.00**	**0.99**	0.86	**0.93**	**0.97**	**0.91**	**0.94**	**0.51**	**0.33**	**0.43**	**0.26**	**0.55**	**0.42**
Imp. over Baseline	+0.02	+0.20	+0.27	+0.16	+0.70	+0.33	+0.28	+0.11	+0.20	+0.21	+0.19	+0.19	+0.19
Based on FLUX													
FLUX1-dev	0.97	0.88	0.48	0.82	0.67	0.76	0.76	0.49	0.25	0.23	0.04	0.38	0.28
+ SpatialReward	**1.00**	**0.99**	**0.89**	**0.98**	**0.99**	**0.94**	**0.97**	**0.63**	**0.40**	**0.52**	**0.32**	**0.45**	**0.46**
Imp. over Baseline	+0.03	+0.11	+0.41	+0.16	+0.32	+0.18	+0.21	+0.14	+0.15	+0.29	+0.28	+0.17	+0.18

entries. Built on a modular pipeline, SpatialRelBench is designed to facilitate future expansion to additional categories, tasks, or spatial dimensions, thus providing a challenging and discriminative benchmark for assessing complex spatial consistency in T2I models.

5. Experiments

5.1. Implementation Details

Training Configuration. We apply reinforcement learning to SD3.5-M [17] and FLUX1-dev [34], adopting the GRPO approach provided by the Flow-GRPO [43] framework. During training, we employ a sampling timestep of $T = 10$, a group size of $G = 24$, a noise level of $a = 0.7$, and a fixed image resolution of 512×512. For evaluation, we increase the timestep to $T = 40$. Parameter-efficient tuning is enabled by LoRA with a rank $r = 32$ and a scaling factor $\alpha = 64$. The KL regularization coefficient β is set to 0.04. All models are trained on 16 NVIDIA L20 GPUs.

Evaluation Models and Baselines. We evaluate the proposed SpatialReward model by comparing it against a set of established reward models under identical experimental conditions. The baseline set comprises TextOCR [13], PickScore [33], Qwen2.5-VL [2], ImageReward [80], and UnifiedReward [73]. To ensure fairness, all reward models are trained on the same dataset, consisting of 100k spatial-relation prompts automatically generated by GPT-4o and verified for correctness and diversity. The backbones are optimized via Flow-GRPO framework with identical hyperparameter settings. We further compare with advanced proprietary models [20, 78] to reflect real-world performance.

5.2. Quantitative Comparison

Evaluation on Spatial-Consistency Benchmarks As shown in Table 1, integrating the proposed SpatialReward into both Stable Diffusion and FLUX models yields consistent and substantial improvements across all evaluation dimensions in the standard GenEval benchmark and SpatialRelBench. On GenEval, SpatialReward enhances performance not only on common object-level metrics but also on

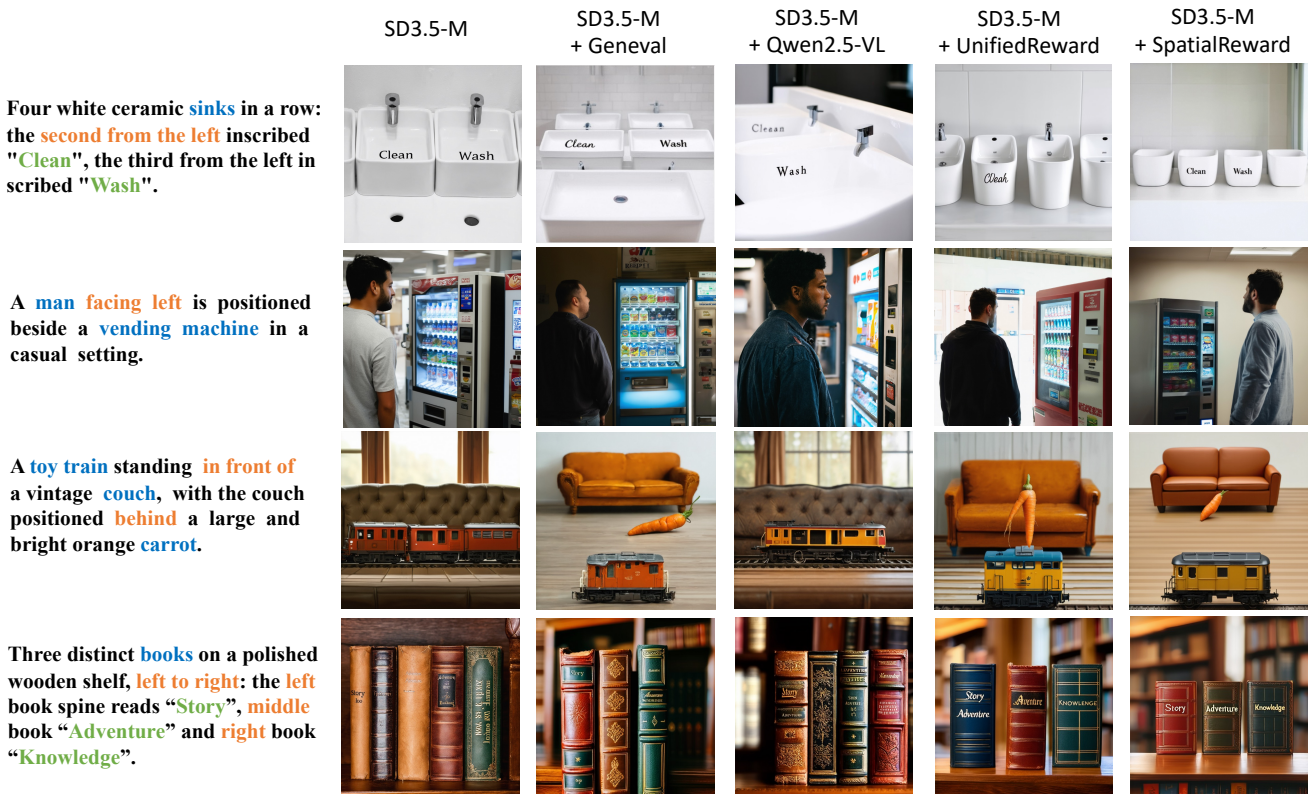


Figure 4. Qualitative comparison of generated image quality across different methods.

Table 2. General-Purpose comparison on Wise, DPG, Aesthetic, and PickScore metrics.

Model	Wise	DPG	Aesthetic	PickScore
SD3.5-M	0.45	83.96	**5.39**	22.34
+ SpatialReward	**0.46**	**84.08**	5.23	**22.52**
Flux1-dev	0.50	83.84	6.13	22.45
+ SpatialReward	**0.52**	**84.19**	**6.15**	**23.22**

Table 3. Correlation and accuracy with human spatial-consistency judgments; accuracy measured at threshold $\tau = 0.8$.

Reward Model	Spearman ρ	Pearson r	Accuracy
CLIPScore	0.42	0.40	0.68
ImageReward	0.48	0.45	0.70
UnifiedReward	0.51	0.49	0.72
VisionReward	0.55	0.53	0.74
SpatialReward	**0.63**	**0.61**	**0.79**

complex compositional tasks, while on SpatialRelBench the gains are particularly notable in the challenging dimensions that require fine-grained spatial reasoning, such as multi-object relation understanding, orientation accuracy, depth ordering, and text–position alignment. The improvements manifest across both benchmarks and for different backbone models, demonstrating that SpatialReward effectively enforces spatial coherence and semantic fidelity in generated images. These results confirm the robustness and generalizability of our approach, and highlight the discriminative power of SpatialRelBench in revealing performance gaps that remain hidden under single-dimensional evaluation protocols.

Evaluation on General-Purpose Metrics To assess the generalization of SpatialReward beyond spatial-focused benchmarks, we evaluated it using widely adopted metrics: overall fidelity (Wise [52], DPG [26]), visual appeal (Aesthetic [60]), and prompt–image alignment (PickScore [33]). With both Stable Diffusion and FLUX backbones, SpatialReward matched or exceeded baseline performance across all metrics, except for a slight drop in Aesthetic. This suggests reward optimization maintains overall visual quality and semantic alignment. Gains on Wise and PickScore reflect improved holistic prompt–image alignment, while stable Aesthetic and DPG scores indicate enhanced spatial reasoning without perceptual loss.

Table 4. Ablation results for SpatialReward. Scores (accuracy) are reported for GenEval, SpatRel: SpatialRelBench, and T2IComp: T2I-CompBench. T2I-CompBench values are averaged over its 2D and 3D spatial-consistency tasks.

Removed Component	GenEval	SpatRel.	T2IComp.
Full SpatialReward	**95.2**	**37.1**	**50.1**
– Exclusion Constraints	90.5	25.9	45.9
– Expert Detection	70.3	21.6	39.2
– CoT Reasoning	94.2	27.9	47.5

5.3. Human Alignment on Spatial Consistency

We evaluated whether SpatialReward aligns more closely with human perception of spatial consistency through a study on 500 prompt–image pairs from SpatialRelBench, each generated by GPT-4o and labeled by annotators as correct or incorrect in depicting the intended spatial relations. Compared with representative general-purpose reward models, SpatialReward showed the highest correlation with human judgments (Spearman's ρ [62], Pearson's r [54]) and the best classification accuracy at a fixed threshold $\tau = 0.8$, confirming that explicitly modeling fine-grained spatial relations produces assessments more consistent with human perception and serves as a reliable metric for complex spatial tasks in text-to-image generation.

5.4. Ablation Study

To assess the contribution of each component within SpatialReward, we conduct ablation experiments by selectively removing three key modules: *Exclusion Constraints*, *Expert Detection*, and *Chain-of-Thought Reasoning*. The resulting accuracy variations across GenEval, SpatialRelBench, and T2I-CompBench are summarized in Table 4.

Effect of Exclusion Constraints Removing the exclusion constraint, which introduces negative samples to penalize undesired spatial configurations, consistently reduces accuracy across all benchmarks. Incorporating such penalties improves reward robustness, mitigates over-optimization, and prevents reward hacking that could degrade real-world performance. This mechanism is particularly beneficial for prompts containing distractor objects, preventing performance degradation in realistic settings.

Effect of Expert Detection Omitting the expert detection stage, which performs fine-grained verification of object presence, attributes, and rendered text, results in the largest accuracy decline across all benchmarks. This outcome substantiates our core principle that spatial reasoning benefits significantly from verified reward signals, as domain-specific detectors provide reliable and interpretable evidence that general-purpose vision–language models alone cannot consistently ensure.

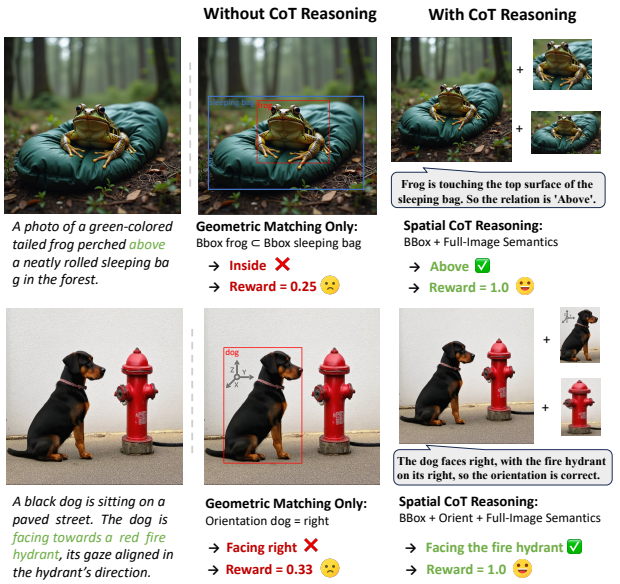


Figure 5. Effect of CoT reasoning in spatial relations. CoT combines bounding boxes, orientation, and scene semantics, yielding correct classifications where detected-only matching fails.

Effect of Chain-of-Thought Reasoning Eliminating the Chain-of-Thought reasoning process, which integrates verified object attributes with geometric analysis to infer spatial relations, results in moderate accuracy reductions. Although the quantitative decline is smaller than that observed for other components, qualitative case studies (Fig. 5) highlight CoT's critical role in complex scenarios. Examples include distinguishing "above" from "inside" and resolving intricate multi-object arrangements that rule-based heuristics cannot effectively handle.

6. Conclusion

We presented SpatialReward, a spatial-consistency verifiable reward model, and SpatialRelBench, a benchmark for fine-grained evaluation of spatial relations in text-to-image generation. By combining constraint parsing, expert-based verification, and chain-of-thought reasoning, SpatialReward effectively enforces spatial fidelity in complex scenes. Experiments on both SpatialRelBench and general evaluation metrics, along with human alignment studies, show that our method surpasses existing reward models in capturing spatial correctness without compromising overall image quality. The results support the use of verifiable spatial reward as a reliable and adaptable objective for future text-to-image generation models.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. U24A20326.

References

[1] Sumukh K Aithal, Pratyush Maini, Zachary Lipton, and J Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation. *Advances in Neural Information Processing Systems*, 37:134614–134644, 2024. 1

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 2, 3, 5, 6

[3] Lital Binyamin, Yoad Tewel, Hilit Segev, Eran Hirsch, Royi Rassin, and Gal Chechik. Make it count: Text-to-image generation with an accurate number of objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13242–13251, 2025. 1

[4] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 3

[5] Boyuan Cao, Jiaxin Ye, Yujie Wei, and Hongming Shan. RepLDM: Reprogramming pretrained latent diffusion models for high-quality, high-efficiency, high-resolution image generation. In *Adv. Neural Inform. Process. Syst.* 1

[6] Agneet Chatterjee, Yiran Luo, Tejas Gokhale, Yezhou Yang, and Chitta Baral. Revision: Rendering tools enable spatial fidelity in vision-language models, 2024. 3

[7] Agneet Chatterjee, Gabriela Ben Melech Stan, Estelle Aflalo, Sayak Paul, Dhruba Ghosh, Tejas Gokhale, Ludwig Schmidt, Hannaneh Hajishirzi, Vasudev Lal, Chitta Baral, et al. Getting it right: Improving spatial consistency in text-to-image models. In *European Conference on Computer Vision*, pages 204–222. Springer, 2024. 1

[8] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 2

[9] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025. 5

[10] Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. Unified hallucination detection for multimodal large language models. *arXiv preprint arXiv:2402.03190*, 2024. 2

[11] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16901–16911, 2024. 4

[12] Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. Depth-regularized optimization for 3d gaussian splatting in few-shot images. In *CVPR*, 2024. 2

[13] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. Paddleocr 3.0 technical report, 2025. 2, 4, 5, 6

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[15] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023. 1

[16] 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. 1

[17] 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 *ICML*, 2024. 1, 2, 3, 6

[18] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS) 2023*. Neural Information Processing Systems Foundation, 2023. 1

[19] Xingyu Fu, Muyu He, Yujie Lu, William Yang Wang, and Dan Roth. Commonsense-t2i challenge: Can text-to-image generation models understand commonsense? *arXiv preprint arXiv:2406.07546*, 2024. 3

[20] Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025. 6

[21] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *NIPS*, 36:52132–52152, 2023. 1, 3

[22] Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*, 2022. 1, 3, 5

[23] Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation, 2023. 3

[24] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning

capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2

[25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 3

[26] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 7

[27] Hai Huang, Yan Xia, Shengpeng Ji, Shulei Wang, Hanting Wang, Minghui Fang, Jieming Zhu, Zhenhua Dong, Sashuai Zhou, and Zhou Zhao. Enhancing multimodal unified representations for cross modal generalization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2353–2366, Vienna, Austria, 2025. Association for Computational Linguistics. 3

[28] Hai Huang, Yan Xia, Sashuai Zhou, Hanting Wang, Shulei Wang, and Zhou Zhao. Bridging domain generalization to multimodal domain generalization via unified representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22488–22498, 2025. 3

[29] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *NIPS*, 36:78723–78747, 2023. 1, 3

[30] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1, 3

[31] Yichen Huang and Lin F Yang. Gemini 2.5 pro capable of winning gold at imo 2025. *arXiv preprint arXiv:2507.15855*, 2025. 5

[32] Amita Kamath, Jack Hessel, and Kai-Wei Chang. Text encoders bottleneck compositionality in contrastive vision-language models. *arXiv preprint arXiv:2305.14897*, 2023. 3

[33] 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. *NIPS*, 36:36652–36663, 2023. 1, 2, 3, 6, 7

[34] Black Forest Labs. Flux. *https://github.com/black-forest-labs/flux*, 2024. 2, 3, 6

[35] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv:2408.03326*, 2024. 3

[36] Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, et al. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. *arXiv preprint arXiv:2206.03001*, 2022. 2, 4

[37] Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025. 1

[38] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, 2022. 4

[39] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023. 1

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[41] Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. *arXiv preprint arXiv:2306.01879*, 2023. 2, 3

[42] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 1

[43] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025. 1, 2, 3, 6

[44] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 2, 4

[45] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025. 2

[46] Wuyang Luo, Su Yang, Hong Wang, Bo Long, and Weishan Zhang. Context-consistent semantic image editing with style-preserved modulation. In *European conference on computer vision*, pages 561–578. Springer, 2022. 3

[47] Wuyang Luo, Su Yang, Xinjian Zhang, and Weishan Zhang. Siedob: Semantic image editing by disentangling object and background. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1868–1878, 2023. 3

[48] Wuyang Luo, Chengkai Tan, Chang Ge, Binye Hong, Su Yang, and Yongjiu Ma. Fontcrafter: High-fidelity element-driven artistic font creation with visual in-context generation. *arXiv preprint arXiv:2603.22054*, 2026. 3

[49] Junpeng Ma, Qizhe Zhang, Ming Lu, Zhibin Wang, Qiang Zhou, Jun Song, and Shanghang Zhang. Mmg-vid: Maximizing marginal gains at segment-level and token-level for efficient video llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 24253–24261, 2026. 3

[50] 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*, pages 10910–10921, 2023. 2, 3, 5

[51] Fanqing Meng, Wenqi Shao, Lixin Luo, Yahong Wang, Yiran Chen, Quanfeng Lu, Yue Yang, Tianshuo Yang, Kaipeng Zhang, Yu Qiao, et al. Phybench: A physical commonsense benchmark for evaluating text-to-image models. *arXiv preprint arXiv:2406.11802*, 2024. 3

[52] Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025. 3, 7

[53] OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[54] K. Pearson. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895. 8

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2, 3, 4

[56] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 1

[57] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 4

[58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[59] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 3

[60] Chrisoph Schuhmann. Laion aesthetics. *https://github.com/LAION-AI/aesthetic-predictor*, 2022. 7

[61] X. Shang, D. Wang, G. Chen, and et al. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019. 5

[62] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15 (1):72–101, 1904. 8

[63] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6619–6628, 2019. 2

[64] Gong Tianxiang, Gao Shiqi, Song Qi, Sun Qingyun, Zhou Haoyi, and Li Jianxin. Towards reliable multimodal intelligence via uncertainty-aware inference. *Chinese Journal of Electronics*, 35:1–16, 2026. 3

[65] 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 *CVPR*, pages 8228–8238, 2024. 1

[66] Qunzhong Wang, Jie Liu, Jiajun Liang, Yilei Jiang, Yuanxing Zhang, Jinyuan Chen, Yaozhi Zheng, Xintao Wang, Pengfei Wan, Xiangyu Yue, et al. Vr-thinker: Boosting video reward models through thinking-with-image reasoning. *arXiv preprint arXiv:2510.10518*, 2025. 2

[67] Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11998–12008, 2023. 2, 3

[68] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. 2

[69] Xuqin Wang, Tao Wu, Yanfeng Zhang, Lu Liu, Dong Wang, Mingwei Sun, Yongliang Wang, Niclas Zeller, and Daniel Cremers. Ladb: Latent aligned diffusion bridges for semi-supervised domain translation. In *DAGM German Conference on Pattern Recognition*, pages 221–236. Springer, 2025. 3

[70] Xuqin Wang, Tao Wu, Yanfeng Zhang, Lu Liu, Mingwei Sun, Yongliang Wang, Niclas Zeller, and Daniel Cremers. Geodesicnvs: Probability density geodesic flow matching for novel view synthesis. *arXiv preprint arXiv:2603.01010*, 2026. 1

[71] Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *arXiv preprint arXiv:2505.03318*, 2025. 3

[72] Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *arXiv preprint arXiv:2505.03318*, 2025. 2

[73] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025. 1, 2, 3, 6

[74] Zehan Wang, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything: Learning robust object orientation estimation from rendering 3d models, 2024. 4

[75] Zehan Wang, Sashuai Zhou, Shaoxuan He, Haifeng Huang, Lihe Yang, Ziang Zhang, Xize Cheng, Shengpeng Ji, Tao Jin, Hengshuang Zhao, et al. Spatialclip: Learning 3d-aware image representations from spatially discriminative language. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29656–29666, 2025. 3

[76] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*, 2024. 4, 5

[77] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2

[78] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 6

[79] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 1, 2, 3

[80] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 1, 2, 3, 6

[81] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 3

[82] Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024. 2, 3

[83] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025. 1, 3

[84] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 4, 5

[85] Xiaoda Yang, Jiayang Xu, Kaixuan Luan, Xinyu Zhan, Hongshun Qiu, Shijun Shi, Hao Li, Shuai Yang, Li Zhang, Checheng Yu, et al. Omnicam: Unified multimodal video generation via camera control. *arXiv preprint arXiv:2504.02312*, 2025. 3

[86] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025. 2

[87] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14246–14255, 2023. 2

[88] Zilyu Ye, Zhiyang Chen, Tiancheng Li, Zemin Huang, Weijian Luo, and Guo-Jun Qi. Schedule on the fly: Diffusion time prediction for faster and better image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23412–23422, 2025. 3

[89] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022. 2, 3

[90] Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024. 3

[91] Yang Zhang, Teoh Tze Tzun, Lim Wei Hern, and Kenji Kawaguchi. Enhancing semantic fidelity in text-to-image synthesis: Attention regulation in diffusion models. In *European Conference on Computer Vision*, pages 70–86. Springer, 2024. 2

[92] Chunzheng Zhu, Yangfang Lin, Jialin Shao, Jianxin Lin, and Yijun Wang. Pathology-aware prototype evolution via llm-driven semantic disambiguation for multicenter diabetic retinopathy diagnosis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9196–9205, 2025. 3

[93] Chunzheng Zhu, Yangfang Lin, Shen Chen, Yijun Wang, and Jianxin Lin. Medeyes: Learning dynamic visual focus for medical progressive diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13916–13924, 2026. 1