Extended Abstract

Motivation Diffusion-based generative text-to-image (T2I) models can produce high-quality images given an instructional user prompt. However, even though images are often photorealistic, modern T2I models often fail to adhere to fine-grained user specifications regarding spatial relationships and specific compositional details. Reinforcement learning based approaches that have been taken to remedy this failure mode include using image-caption scoring metrics such as CLIPScore or VLMs; however, these approaches often fall short due to VLMs and CLIP-like models outputting noisy rewards which are biased towards image quality rather than spatial consistency. We aim to address this through a novel baselining protocol we term caption keyword replacement (CKR), a simple adjustment that can be applied to previous CLIP and VLM-based approaches without additional reward model fine-tuning.

Method We propose Caption Keyword Replacement (CKR), a baselining protocol that sharpens a CLIP-based reward without any additional reward-model fine-tuning. Given a ground-truth prompt c_i for a generated image I, CKR constructs a set of counterfactual captions that each perturb one single compositional feature of c_i while leaving the rest of the caption fixed. We consider two kinds of perturbation: swapping the values of two bound attributes or objects ("a red book and a yellow vase" becomes "a yellow book and a red vase"), and modifying one attribute to an alternative value in isolation ("a red book and a yellow vase" becomes "a yellow book and a yellow vase"). The resulting captions act as near-misses that differ from the target only along the compositional axis under test. We then define the reward as the margin between the CLIPScore of the true caption and the mean CLIPScore of its counterfactuals, $R_i = \text{CLIP-S}(c_i, I) - \frac{1}{n-1} \sum_{j \neq i} \text{CLIP-S}(c_j, I)$. We use R_i directly as the reward in the DDPO objective during fine-tuning.

Implementation We generate counterfactual captions with GPT-5.4-mini, producing ten variants for complex compositions and six for all other categories. We fine-tune the caption-conditioned UNet of Stable Diffusion 1.5 with DDPO for a single epoch on the 5,600-prompt training split of T2I-CompBench++, using a learning rate of 1e-6, five gradient steps per batch of rollouts, mini-batches of 36 prompts, and a rollout length of $T = 25$ under DDIM sampling. Because T2I-CompBench++ provides prompts but not paired ground-truth images, we validate the reward signal separately on 100 image-caption pairs drawn from Microsoft COCO and SugarCrepe, spanning attribute binding, numeracy, and spatial relationships. For each pair we treat the original caption as positive, generate six CKR counterfactuals, and compare CLIP-S scores across all candidates.

Results On the reward-validation set of 100 examples and 600 counterfactuals, raw CLIP-S ranks the true caption highest only 36.0% of the time, and the positive caption wins just 69.8% of pairwise comparisons against individual counterfactuals. On the downstream T2I-CompBench++ evaluation, the CKR-fine-tuned model scores marginally highest on attribute binding (0.392) and complex compositions (0.641), the CLIPScore baseline is best on 2D/3D spatial relations (0.288), and untuned SD1.5 remains best on numeracy (0.451). No system significantly outperforms untuned SD1.5 across the board.

Discussion The counterfactuals in CKR expose real weaknesses in raw CLIP-S, and using them as a local baseline measurably improves discrimination of correct compositions. But this does not translate into a significant downstream gain. We attribute this partly to a small training set and a single training epoch, and more fundamentally to CLIP itself, which is weak at spatial and directional reasoning that a baseline can mitigate but not fix. CKR is further limited by its assumption that the compositional unit under test is identifiable in isolation, which breaks down when multiple attributes interact.

Conclusion CKR sharpens a CLIP-based reward by baselining against minimally perturbed counterfactuals, and our validation study confirms it improves discrimination of correct compositions. Under DDPO fine-tuning, however, this gain does not carry through to T2I-CompBench++, bottlenecked by CLIP's own weakness at spatial and multi-object reasoning. Future work should test CKR on stronger VLM- or preference-based reward models and extend it to multi-object scenes where several constraints interact.

CKR: A Novel Baselining Improvement for T2I-Model Reinforcement Learning Using Noisy Rewards

Stanford CS224R Custom Project

Ethan Zhang
Dept. of Computer Science
Stanford University

Jenny Wei
Dept. of Computer Science
Stanford University

Tatiana Zhang
Dept. of Classics
Stanford University

Abstract

Diffusion-based generative text-to-image (T2I) models can produce high-quality images given an instructional user prompt. However, though images are often photo-realistic, modern T2I models often fail to adhere to fine-grained user specifications regarding spatial relationships and specific compositional details. Reinforcement learning based approaches that have been taken to remedy this failure mode include using image-caption scoring metrics such as CLIPScore or VLMs; however, these approaches often fall short due to VLMs and CLIP-like models outputting noisy rewards which are biased towards image quality rather than spatial consistency. In this project, we aim to address this problem through a novel baselining protocol we term caption keyword replacement (CKR), a simple adjustment that can be applied to previous CLIP and VLM-based approaches without additional reward model fine-tuning. Given a ground-truth prompt, CKR produces a set of decoy prompts by replacing compositional keyword(s) while keeping the rest of the prompt identical to the original. By using the average CLIPScore over these decoy prompts as a per-prompt baseline, subtracted from the score of the ground-truth prompt, CKR isolates the portion of the reward that reflects compositional correctness rather than overall image quality. The resulting reward is positive only when the generated image matches the true prompt more closely than its near-miss alternatives, which suppresses the quality-biased offset shared across all prompts for a given image. We validate this reward signal in isolation and then use it to fine-tune Stable Diffusion 1.5 with DDPO, evaluating on the T2I-CompBench++ benchmark. We find that CKR improves discrimination of correct compositions over raw CLIPScore, most clearly on spatial-relationship prompts, but that this gain translates into only marginal downstream improvement, which we attribute to the underlying CLIP reward remaining too noisy for either signal to drive substantial updates.

1 Introduction

Recent advancements in artificial intelligence have led to the advent of text-to-image (T2I) diffusion models which can generate images given an instructional user prompt. However, though generated images are often photorealistic, models often fail to adhere to fine-grained user instructions concerning spatial relationships and specific compositional details. Reinforcement learning based approaches that have been taken to remedy this failure mode include using image-caption scoring metrics such as CLIPScore or VLMs - however, these approaches often fall short due to VLMs and CLIP-like models outputting noisy rewards which are biased towards image quality rather than spatial consistency. **In this project, we aim to address this problem through a simple baselining-like adjustment to previous CLIP-based approaches**.

2 Related Work

Training Diffusion Models with Reinforcement Learning [1]: Introduces DDPO, an RL approach targeted at prompt-image alignment. DDPO reframes the denoising process as a multi-step MDP. Demonstrates that this multi-step formulation outperforms reward-weighted regression (which treats the full sampling process as a single step), and shows that vision-language models can serve as automated reward signals in place of human labels, including for prompt-image alignment. While the authors apply their algorithm to T2I model fine-tuning, their work still has limitations. Specifically, the prompts that the authors include in their fine-tuning set are relatively coarse compared to fine-grained spatial relationship requirements. Thus, while we employ the DDPO objective for our model training work, our project differs in that we specifically target fine-grained instructions, instead of broad image alignment.

T2I-CompBench++ [2]: A benchmark for compositional text-to-image generation consisting of 8,000 text prompts. Prompts are split into four categories (attribute binding, object relationship, numeracy, and complex compositions). In order to evaluate fine-grained text-image performance, rather than using traditional CLIP scores, the benchmark introduces its own metrics: BLIP-VQA for attribute binding, UniDet for spatial relationship and numeracy, and MLLM for nonspatial relationship and complex prompts. We plan to benchmark our approach on the four categories of T2I-CompBench.

SpatialReward [3]: Introduces a spatial reward model combining prompt decomposition, expert detectors (object detection, OCR, depth estimation), and VLM chain-of-thought reasoning. They find that CLIP- and preference-based rewards (PickScore, ImageReward) produce only modest spatial improvements under RL fine-tuning, while their structured composite reward yields substantially larger gains. This motivates revisiting how CLIP-style rewards are used in compositional and spatial settings, rather than treating them as a fixed signal.

3 Methods

3.1 Caption Keyword Replacement

Let $\mathbf{c}$ denote a string (caption) and $A_\theta(\mathbf{c})$ be the T2I agent-predicted image. CLIPScore [4] is defined as

$$\text{Clip-S}(\mathbf{c}, A_\theta(\mathbf{c})) = \max\left(\cos(t_\phi(\mathbf{c}), i_\psi(A_\theta(\mathbf{c}))), 0\right),$$

where $t_\phi(\mathbf{c})$, $i_\psi(A_\theta(\mathbf{c})) \in \mathbb{R}^{512}$ are the embeddings produced by CLIP's text and image networks, respectively. However, naively using CLIP as a reward model is problematic due to its tendencies to overestimate rewards for high-quality (but spatially incorrect) images. To mitigate this, we propose a baselining approach:

$$R_{\text{Clip-BL}}(\mathbf{c}, A_\theta(\mathbf{c})) = \text{Clip-S}(\mathbf{c}, A_\theta(\mathbf{c})) - \mathbb{E}_{\mathbf{c}' \sim \text{CKR}(\mathbf{c})}\left[\cos(t_\phi(\mathbf{c}'), i_\psi(A_\theta(\mathbf{c}')))\right],$$

where $\text{CKR}(\mathbf{c})$ denotes a "caption keyword replacement" operation. Concretely, given a prompt $\mathbf{c}$ (e.g. "A cat *on top of* a *blue* chair"), $\text{CKR}(\mathbf{c})$ identifies key words and phrases concerning fine-grained details (e.g. "on top of", "blue") and permutes them to create a set of related prompts (e.g. "A cat *below* a *yellow* chair", etc) - we hypothesize that by doing so, we can more robustly extract image compositional signal from CLIP. Following the work in [2], the specific types of keywords we focus on are *attribute binding* keywords, *spatial relationship* keywords, and *numeracy* keywords.

3.2 Baseline Reward Calculation

Raw CLIPScore assigns globally similar values to an image whether its composition matches the target or a near-miss, making its absolute magnitude a weak, prompt-dependent training signal. To mitigate this, we define the reward for the true caption c_i and generated image I as the margin between the CLIPScore of the true caption and the mean CLIPScore of its counterfactuals:

$$R_i = \text{CLIP-S}(c_i, I) - \frac{1}{n-1}\sum_{j \neq i} \text{CLIP-S}(c_j, I), \tag{1}$$

where CLIP-S$(\cdot, I)$ is the cosine similarity between the CLIP embeddings of a caption and the image I, and n is the total number of captions (the target plus its $n-1$ counterfactuals). We use R_i directly as the reward $r(x_0, c)$ in the DDPO objective during fine-tuning.

3.3 Model Fine-Tuning

3.3.1 Background

We choose to fine-tune Stable Diffusion 1.5 (SD1.5), a latent diffusion-based T2I model [5]. Concretely, given a caption $\mathbf{c}$ and a desired final image size $h_{\text{image}} \times w_{\text{image}}$, SD1.5 generates an image by doing the following:

1) Randomly sampling the prior distribution over latent space

$$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \sigma_T \mathbf{I}),\ \mathbf{x}_T \in \mathbb{R}^{c_{\text{latent}} \times h_{\text{latent}} \times w_{\text{latent}}},$$

2) Iteratively predicting and removing noise via a caption-conditioned deep neural network, producing a T-step denoising trajectory $\{\mathbf{x}_{T-1}, \ldots, \mathbf{x}_0\}$,

3) Passing the fully denoised latent sample through the decoder portion of a variational autoencoder to map it back to image space

$$\hat{\mathbf{y}} = \text{VAE}_{\phi}^{\text{dec}}(\mathbf{x}_0).$$

The value of σ_T in step 1, as well as the dynamics of the denoising process, are determined by the user's selection of a noise schedule and sampling algorithm, respectively. In our work, we use the DDIM sampling algorithm [6], under which the denoising process is a Markov chain governed by the following dynamics:

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{\mathbf{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\mathbf{x}_t, t, c)}{\sqrt{\alpha_t}} \right) + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \epsilon_\theta(\mathbf{x}_t, t, c) + \sigma_t W_t,$$

where $W_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is standard Gaussian noise sampled at every timestep and α_t is an additional schedule-dependent parameter.

3.3.2 Fine-Tuning Details

We fine-tune the diffusion module of SD1.5—specificaly, in SD1.5, the neural network $\epsilon_\theta(\mathbf{x}_t, t, c)$ is a caption-conditioned UNet. We elect to use the denoising diffusion policy optimization (DDPO) algorithm introduced by [1]:

$$\nabla_\theta \mathcal{J}_{\text{DDPO}} = \mathbb{E}\left[\sum_{t=0}^{T} \frac{p_\theta(x_{t-1}|x_t, c)}{p_{\theta_{\text{old}}}(x_{t-1}|x_t, c)} \nabla_\theta \log p_\theta(x_{t-1}|x_t, c) r(x_0, c) \right],$$

the surrogate objective for which is

$$\mathcal{J}_{\text{surr}} = \mathbb{E}\left[\sum_{t=0}^{T} \text{Clip}\left(\frac{p_\theta(x_{t-1}|x_t, c)}{p_{\theta_{\text{old}}}(x_{t-1}|x_t, c)}, 1+\varepsilon, 1-\varepsilon \right) r(x_0, c) \right].$$

Crucially, as the denoising process is a Markov chain where each transition is given by a Gaussian conditional distribution, the per-step transition likelihoods are both computable and differentiable, allowing the application of the above objective.

We fine-tune for a single epoch on the 5,600-prompt training portion of the T2I-CompBench++ dataset, described further in the Methods section. We use a learning rate of 1e-6. As DDPO is an off-policy algorithm, for each batch of policy rollouts, we perform five gradient steps. We train using mini-batches containing 36 text prompts and set $T = 25$ for rollout length. Training was conducted on a B200 GPU through Modal.

4 Experiments

4.1 Dataset Overview: T2I-CompBench++

We train and benchmark on T2I-CompBench++ ([2]), a benchmark for compositional text-to-image generation consisting of 8,000 total text prompts. The 8,000 prompts are divided in a 70/30 train-test split, which we use for fine-tuning and evaluation, respectively. The benchmark splits its prompts into four overall categories: 1) Attribute Binding, which evaluates whether attributes such as colors, shapes, and textures are correctly associated with their intended objects; (2) Numeracy, which evaluates the model's ability to generate the correct number of objects; (3) Object Relationships, which evaluates 2D spatial relations (e.g., left/right), 3D spatial relations (e.g., in front of/behind), and non-spatial semantic relationships between objects; and (4) Complex Compositions, which combine multiple compositional constraints within a single prompt.

4.2 Dataset Augmentation via CKR

Given a prompt c_i from T2I-CompBench++, which we treat as the ground-truth caption for an image I generated under that prompt, we use a LLM to construct a set of counterfactual captions $\{c_j\}_{j \neq i}$ that each perturb one single compositional feature of c_i while leaving the rest of the caption the same.

For example, "a red book and a yellow vase" may be modified to "a blue book and a yellow vase," while "a cat to the left of a dog" may be modified to "a cat to the right of a dog." These counterfactuals act as near-miss alternatives that differ from the target caption only along the compositional axis under test.

We use GPT-5.4-mini to generate these counterfactuals. We generate 10 variants for complex compositions, and 6 variants for all other categories; these numbers were determined empirically based on preliminary experiments to provide a reasonable trade-off between counterfactual diversity and computational cost.

4.3 CKR Reward Validation

Before incorporating CKR into the DDPO training pipeline, we first evaluate whether CKR-generated counterfactual captions provide meaningful local alternatives for compositional reward modeling. Note that this validation experiment is separate from our final T2I-CompBench++ evaluation: since T2I-CompBench++ provides compositional text prompts but not paired ground-truth images, it is not directly suitable for testing whether a reward model prefers a correct caption over incorrect captions for a fixed image. Instead, we construct a reward validation set containing 100 image-caption pairs from Microsoft COCO ([7]) and SugarCrepe ([8]), spanning three compositional categories: attribute binding, numeracy, and spatial relationships.

For each image-caption pair, we treat the original caption as the positive caption and apply our CKR augmentation procedure to generate six counterfactual captions. For each image I, we compute CLIP-S scores for the positive caption and all associated counterfactual captions. We evaluate whether CLIP-S distinguishes the correct caption from CKR hard negatives using three ranking-based metrics. **Top-rank accuracy** measures the fraction of examples for which the positive caption receives the highest CLIP-S score among all candidates. **Pairwise win rate** measures the fraction of positive-counterfactual pairs for which the positive caption scores higher than the counterfactual, i.e., $S(c, I) > S(c', I)$. **Hardest-negative margin** measures the difference between the positive caption score and the highest-scoring counterfactual, defined as $S(c, I) - \max_{c'} S(c', I)$. Negative margins indicate that at least one counterfactual caption outscored the correct caption.

5 Results & Discussion

5.1 CKR Reward Validation

Table 1 shows that CKR-generated counterfactuals are challenging alternatives for a CLIP-based reward model. Across 100 validation examples and 600 counterfactual captions, CLIP-S ranks the true caption highest only 36.0% of the time. The average CLIP-S of the positive caption is 0.3038,

while the average score of the highest-scoring CKR variant is slightly higher at 0.3051; raw CLIP-S often assigns comparable or greater reward to plausible but semantically incorrect near-miss captions.

Category	n	**Top-rank Acc.**	**Pairwise Win**	**Hardest Margin**
Attribute binding	35	40.0%	77.6%	-0.0009
Numeracy	30	53.3%	77.8%	0.0011
Spatial relationship	35	17.1%	55.2%	-0.0038
Overall	100	36.0%	69.8%	-0.0013

Table 1: Reward validation results for CKR counterfactual captions.

Note the gap between pairwise comparisons and hardest-negative ranking. Although the true caption beats individual CKR variants in 69.8% of pairwise comparisons, it is top-ranked in only 36.0% of examples. This indicates that, for many images, at least one minimally perturbed counterfactual receives a higher CLIP-S score than the correct caption. Thus, while CLIP-S often captures some local differences, it is not reliably robust to the hardest CKR negative.

Performance also varies substantially across compositional categories. Numeracy performs best, with 53.3% top-rank accuracy and a small positive hardest-negative margin, while attribute binding achieves a similar pairwise win rate but lower top-rank accuracy. Spatial relationships are the weakest category, with only 17.1% top-rank accuracy and the most negative hardest-negative margin, suggesting that CLIP-based rewards are especially unreliable for relational distinctions. Overall, these results show that CKR-generated counterfactuals expose weaknesses in raw CLIP-S: minimally perturbed captions often receive scores comparable to, or higher than, the true caption. This motivates using CKR variants as local baselines during reward computation, but also suggests that downstream gains may be limited by the noisiness of CLIP itself.

5.2 Fine-tuning Stable Diffusion 1.5 with DDPO

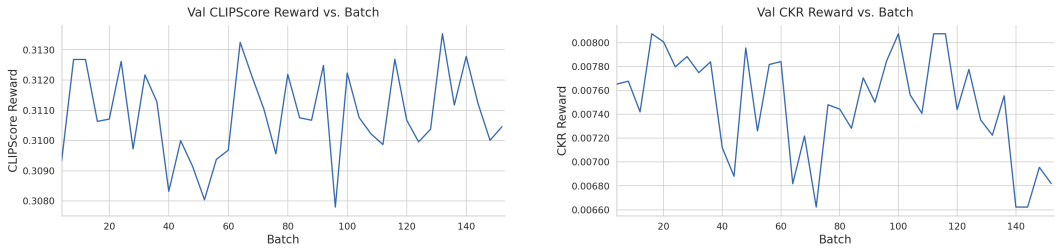

Figure 1: Reward curves for DDPO fin-tuning of Stable Diffusion 1.5. Rewards were evaluated every 4th batch of training. To reduce reward estimation noise, each prompt was rolled out four timed. (a) CLIPScore reward during CLIPScore fine-tuning. (b) CKR reward during CKR fine-tuning.

	Att. Binding	Num.	Object (2D/3D)	Object (Non-spatial)	Complex
SD1.5	0.387	**0.451**	0.274	0.762	0.636
SD1.5 + CLIPScore	0.389	0.449	**0.288**	0.775	0.632
SD1.5 + CKR	**0.392**	0.437	0.281	**0.788***	**0.641**
	B-VQA	UniDet	UniDet	CLIPScore	GPT4-V

Table 2: Stable Diffusion 1.5 + fine-tuned model performances on T2I-Compbench++.

* As we use a CLIPScore-based reward function to fine-tune our models, this result is likely biased.

As shown in Figure 1, the per-batch rewards remain largely unchanged throughout training, suggesting that fine-tuning does not produce substantial improvements in the rewards attained by the model. Likely as a result of this, our fine-tuned models (with both our baseline CLIPScore approach as well as CKR) do not significantly outperform the vanilla Stable Diffusion 1.5 model on T21-Compbench++ (Table 2). While our CKR approach does achieve the best benchmark score in some categories, the

performance increase is not substantial enough to suggest that our approach brings about genuine model improvement. Furthermore, one of the categories that our CKR fine-tuned model outperforms the others in (the non-spatial sub-category of the object relations category) uses a CLIP-based evaluation metric, likely meaning that the 'win' for our model is, in reality, a biased and unreliable result.

There are several factors which may have contributed to our results being less strong that anticipated. We believe that CLIP's noisiness and inherent model inaccuracies played a significant role in this regard—within our CKR testing, we observed that CLIP could often be unpredictably inaccurate. As such, it is likely that the reward signals we use to fine-tune Stable Diffusion 1.5 are noisy, rendering it difficult for the model to properly learn to generate reward-maximizing outputs. Relatedly, due to the computational expense of the DDPO procedure and the large parameter count of the Stable Diffusion 1.5 UNet, we could only conduct one epoch of training per fine-tuning approach. It is very possible that due to the low-training regime that we operated in, the model simply was not able to extract sufficient learning signal from the small dataset we used. Finally, although CKR demonstrated promising performance on our toy dataset consisting of 100 high-quality, human-curated image-caption pairs, its intended deployment setting differed substantially from this evaluation environment. Specifically, during fine-tuning, the images are generated by Stable Diffusion 1.5 rather than being high-quality, human-curated images. This discrepancy introduces a distribution shift between the data used to develop CKR and the data on which it is ultimately deployed. Consequently, CKR may provide less reliable reward estimates in practice, potentially contributing to the limited performance gains observed during fine-tuning.

6 Conclusion

We introduced Caption Keyword Replacement (CKR), a counterfactual baselining method for improving CLIP-based rewards in text-to-image reinforcement learning. Rather than treating raw CLIPScore as an absolute measure of prompt-image alignment, CKR compares the target caption against minimally perturbed alternatives that differ along specific compositional axes. Our reward validation experiments show that these counterfactuals expose weaknesses in raw CLIP-S, especially for spatial relationships, where near-miss captions often receive scores comparable to or higher than the correct caption.

When applied to DDPO fine-tuning of Stable Diffusion 1.5, CKR produced only marginal downstream improvements on T2I-CompBench++. This suggests that while counterfactual baselining can make reward evaluation more compositionally targeted, it is still limited by the quality of the underlying reward model. In particular, CLIP remains weak at spatial, directional, and multi-object compositional reasoning, and baseline subtraction can only partially mitigate this noise.

Future work should apply CKR to stronger reward models, such as VLM-based or preference-based evaluators, to test whether the counterfactual baseline generalizes beyond CLIP. Another promising direction is extending CKR to more complex multi-object scenes, where several compositional constraints interact and single-feature perturbations may underestimate the ambiguity of the reward signal.

7 Team Contributions

- **Ethan Zhang:** Implemented DDPO training setup + DDPO experiments.
- **Jenny Wei:** Implemented LLM-based CKR approach + CKR reward validation experiments + dataset cleaning and preprocessing.
- **Tatiana Zhang:** Implemented rule-based CKR protocol + LLM-based CKR approach + dataset cleaning and preprocessing.

Changes from Proposal In our original proposal, we planned to implement CKR using a rule-based keyword replacement system with fixed vocabularies for attributes, counts, and spatial relations. During preliminary experiments, we found that this approach was too limited: it missed many valid compositional phrases and often failed to generate enough useful counterfactual variants. We therefore replaced the rule-based generator with an LLM-based CKR pipeline, which allows us to

produce grammatical, semantically close counterfactual captions across a broader range of prompt types while preserving the original goal of perturbing a single compositional feature at a time.

References

[1] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301, 2024.

[2] 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, 2025.

[3] 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, 2026.

[4] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022.

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[6] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[8] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality, 2023.