Structured Counterfactual Contrast for Compositional Text-to-Image Generation: A Pilot Study of Prompt Negation and Candidate Reranking

Anonymous authors
Paper under double-blind review

Abstract

Text-to-image diffusion models often fail on compositional prompts involving attribute binding, spatial relations, and multiple objects. Such failures can be near-miss compositions: an image is broadly plausible but violates one local structural constraint, such as an attribute assignment or a relation direction.

We study structured counterfactual contrast as an inference-time diagnostic and candidate-selection signal. We represent a prompt as a constraint graph and construct minimally perturbed counterfactuals through attribute swaps, relation inversions, and argument-role swaps. A frozen vision–language scorer then computes the margin between compatibility with the target prompt and compatibility with its nearest counterfactual alternative.

We evaluate two lightweight instantiations without modifying the diffusion model. First, prompt-level counterfactual augmentation lowers the diagnostic margin in a matched pilot, indicating that textual negation is not a reliable way to express structured exclusions. Second, in a compute-matched three-candidate setting, counterfactual-aware reranking selects images with higher values of the margin than first-candidate and random-selection baselines. Because this margin is also the selection objective, we treat this result as a selection diagnostic rather than independent evidence of improved compositional fidelity.

Our results establish a constrained empirical finding: explicit structured comparison is more promising than direct textual negation as an inference-time use of counterfactual information. We release a randomized method-blind evaluation protocol for assessing whether this signal translates into independently judged compositional fidelity, and identify latent-level counterfactual guidance as a direction for future work.

1 Introduction

Text-to-image diffusion models have substantially improved visual quality, but reliable compositional generation remains difficult. A prompt such as "a red cube inside a blue glass on a wooden table" requires correct object identity, attribute binding, and spatial or containment relations simultaneously. Strong models can nevertheless produce visually plausible images with a local compositional error: the cube may be outside the glass, colors may be assigned to the wrong objects, or a relation may be reversed.

Most methods for improving compositional fidelity strengthen alignment with the target prompt, for example through cross-attention control, layout conditioning, or prompt decomposition. These approaches emphasize the desired composition. We investigate a complementary question: can a generator or selector also benefit from explicitly representing plausible *incorrect* compositions that differ from the target by a small structural change?

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

We develop a counterfactual-contrastive formulation for this question. Given a prompt, we construct a constraint graph whose nodes represent entities and whose attributes and edges represent entity properties and relations. We then generate hard counterfactuals through minimal transformations, including attribute swaps, relation inversions, and argument swaps. For an image or intermediate latent state, a contrastive score compares compatibility with the target constraint graph against compatibility with these counterfactual alternatives. Intuitively, an image should score highly not only because it matches the target prompt, but also because it is distinguishable from nearby incorrect compositions.

This formulation suggests a latent-level guidance mechanism, in which the contrastive score is differentiated through the denoising process. Implementing and validating that mechanism requires a suitable latent-space verifier and is beyond the scope of the present experiments. Instead, we use two inference-time approximations to isolate the role of counterfactual contrast. First, we append textual counterfactual exclusions to the original prompt. Second, we generate multiple candidates and rerank them using a counterfactual-aware image–text margin.

Our pilot results show a clear negative result for prompt-level augmentation: expressing counterfactual constraints as natural-language exclusions lowers the measured margin relative to standard prompting. Counterfactual-aware reranking, by contrast, selects candidates with a higher average counterfactual margin and wins on most evaluated prompts, though the confidence interval on the mean gain is wide. Thus, the current evidence supports a narrower conclusion: explicit structured comparison can provide a useful selection signal, whereas directly injecting counterfactual descriptions into a prompt is not a reliable proxy for such comparison.

Our contributions are:

- We formulate compositional text-to-image fidelity in terms of structured counterfactual contrast between a target constraint graph and minimally perturbed alternatives.
- We define a counterfactual generator that produces attribute, relation, and argument-role perturbations corresponding to common near-miss compositional failures.
- We evaluate two lightweight inference-time approximations and find that prompt-level counterfactual augmentation degrades the counterfactual margin in our pilot, whereas counterfactual-aware reranking yields a promising but statistically inconclusive mean improvement at the current scale.
- We identify explicit structured comparison as a more promising use of counterfactual information than textual negation, and formulate a latent-level contrastive denoising direction for future evaluation.

2 Related Work

2.1 Compositional Text-to-Image Generation

Text-to-image models often fail on attribute binding, spatial relations, and multi-object compositions. Existing approaches improve compositional fidelity through layout control, scene representations, cross-attention manipulation, or grounding constraints. Despite these advances, models can produce visually plausible near-miss images with swapped relations or leaked attributes. Our work studies whether such failures can be addressed by contrasting the target prompt with minimally perturbed, structurally incorrect alternatives.

2.2 Guided Diffusion and Hard Negatives

Inference-time diffusion guidance typically increases compatibility with a target condition through classifier, classifier-free, energy-based, or attention-based signals. Separately, contrastive learning and vision-language models use hard negatives to distinguish semantically similar alternatives. We combine these perspectives at inference time: structured counterfactual prompts define hard negatives for assessing generated candidates. Unlike training-time contrastive learning, our current implementation requires no diffusion-model fine-tuning.

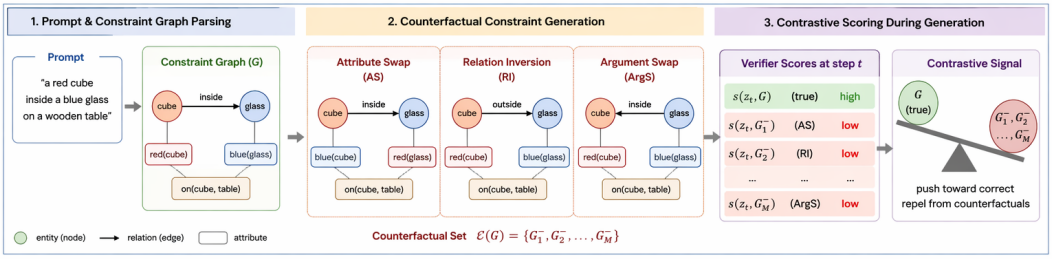

Figure 1: Overview of Counterfactual Constraint Diffusion (CCD). Given a prompt, we parse a structured constraint graph over entities, attributes, and relations. We then generate hard negative counterfactual graphs by minimally perturbing attributes, relations, or arguments. During diffusion, a verifier scores the compatibility of the current latent state with the true graph versus its counterfactual alternatives. The contrastive signal encourages alignment with the true constraints while repelling the trajectory from likely compositional failures.

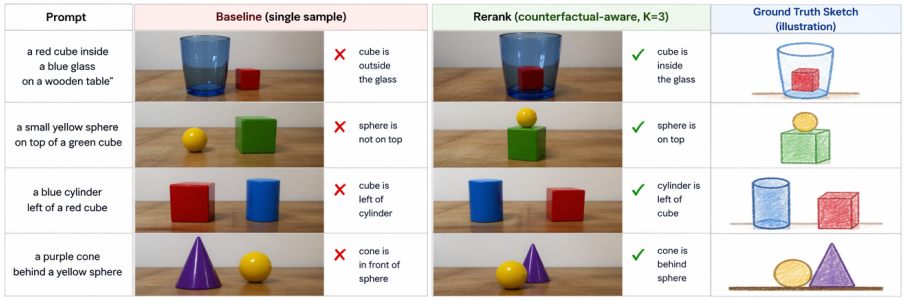

Figure 2: Qualitative comparison between baseline generation and counterfactual-aware reranking. Baseline samples often contain compositional errors such as wrong relations or attribute leakage. Reranking with the counterfactual-aware margin selects samples that better satisfy the intended structure.

Figure 1: CCD constructs a constraint graph from a prompt and generates counterfactual variants by perturbing attributes or relations. The current experiments evaluate prompt-level augmentation and candidate reranking; the latent-level guidance component is a formulation for future evaluation.

3 Method

We formulate **Counterfactual Constraint Diffusion (CCD)** as a counterfactual-contrastive framework for compositional text-to-image generation. CCD represents a prompt by explicit entities, attributes, and relations, then constructs minimally perturbed counterfactual alternatives. The full formulation suggests latent-level contrastive denoising; in this paper, we evaluate lightweight prompt-level and reranking approximations.

3.1 Constraint Graphs and Counterfactuals

Given a prompt p, we represent its composition as

$$G = (V, E, A), \tag{1}$$

where V contains entities, E contains pairwise relations, and A contains entity-specific attributes. For example, "a red cube inside a blue glass" is represented by

$$V = \{\texttt{cube}, \texttt{glass}\}, \quad A = \{\texttt{red(cube)}, \texttt{blue(glass)}\},$$
$$E = \{\texttt{inside(cube,glass)}\}.$$

We construct a counterfactual set

$$\mathcal{C}(G) = \{G_1^-, \ldots, G_M^-\} \tag{2}$$

using minimal structural transformations:

1. **Attribute swap:** assign an attribute to another entity.
2. **Relation inversion:** replace a relation with an incompatible or opposite relation.
3. **Argument swap:** exchange the subject and object of a relation.

These perturbations preserve most lexical content while creating plausible but structurally incorrect compositions.

3.2 Counterfactual Contrast Score

For an image x, let $s(x, p)$ denote a frozen vision-language similarity score between x and prompt p. We define the counterfactual margin

$$m(x, p) = s(x, p) - \max_{\tilde{p} \in \mathcal{C}(p)} s(x, \tilde{p}). \tag{3}$$

A high margin indicates that an image is more compatible with the target prompt than with its nearest counterfactual alternative. This score is used both as a diagnostic metric and as the selection criterion for reranking.

3.3 Inference-Time Instantiations

Prompt-level counterfactual augmentation. We append textual exclusions derived from counterfactual prompts to the original prompt. This tests whether standard language conditioning can express structured negative constraints without changing the diffusion model.

Counterfactual-aware reranking. Given K independently generated candidates $\{x_1, \ldots, x_K\}$, we select

$$x^* = \arg\max_{x_i} m(x_i, p). \tag{4}$$

This method applies structured counterfactual comparison after generation and does not modify the denoising trajectory.

3.4 Latent-Level CCD Formulation

For completeness, the same contrastive principle can be extended to a latent state z_t using a differentiable verifier $f_\phi(z_t, G)$. Let $s_t^+ = f_\phi(z_t, G)$ and $s_{t,j}^- = f_\phi(z_t, G_j^-)$, and define

$$\mathcal{L}_{cc}(z_t) = -\log \frac{\exp(s_t^+/\tau)}{\exp(s_t^+/\tau) + \sum_{j=1}^{M} \exp(s_{t,j}^-/\tau)}. \tag{5}$$

A potential guided update is

$$z_{t-1} = \text{Solver}\left(z_t, \hat{\epsilon}_\theta(z_t, p, t) - \lambda \nabla_{z_t} \mathcal{L}_{cc}(z_t)\right). \tag{6}$$

We do not instantiate this latent verifier or evaluate this update in the current experiments. It is included as a formulation that motivates future latent-level counterfactual guidance.

4 Experiments

4.1 Pilot Setup

We conduct a pilot study with SDXL-Turbo to examine two lightweight uses of counterfactual information: prompt-level augmentation and post-hoc candidate reranking. The prompt-level study uses 12 compositional prompts involving attribute binding and multi-object configurations, with three random seeds per prompt (36 prompt–seed instances). All methods use the same resolution, sampling steps, and seed for a matched prompt–seed instance.

For each prompt p, we construct attribute-focused counterfactual prompts $\tilde{p}$ by swapping attributes between entities. For example, a counterfactual may assign the color of one object to another object.

Prompt-level augmentation. The baseline uses the original prompt p. The prompt-level condition appends textual exclusions derived from its counterfactuals (e.g., "not $\tilde{p}$"). This experiment tests whether natural-language conditioning can express structured counterfactual exclusions.

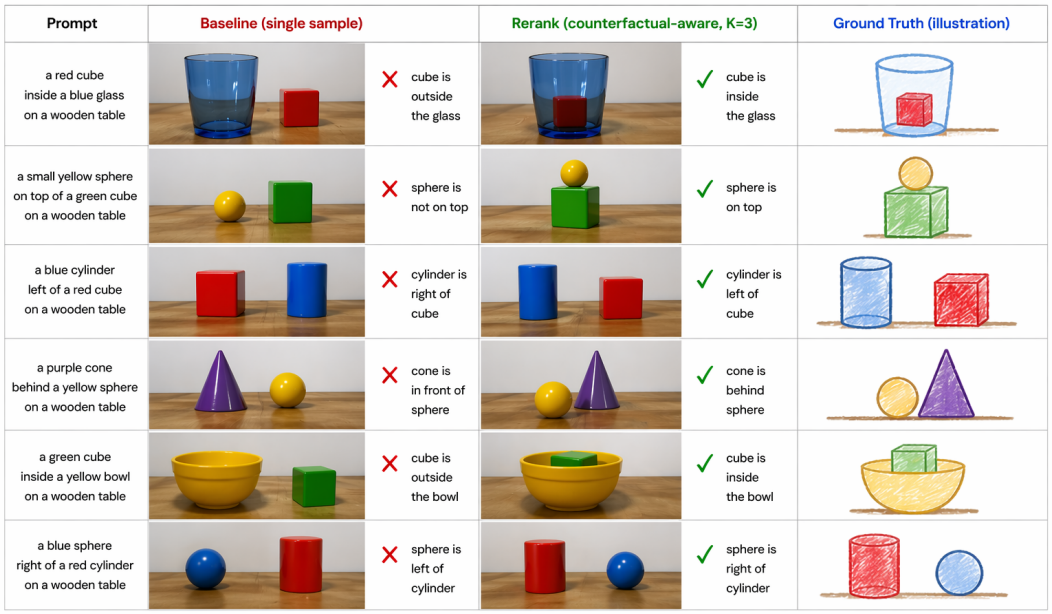

Figure 2: Qualitative comparison between baseline generation and counterfactual-aware reranking.
Baseline samples often contain compositional errors such as wrong relations or attribute leakage.
Reranking with the counterfactual-aware margin selects samples that better satisfy the intended structure.

Figure 2: Illustrative examples of standard single-sample generation and counterfactual-aware reranking. The displayed reranked candidates have larger counterfactual margins under the frozen vision-language scorer. These examples are qualitative and are not used as quantitative evidence.

Table 1: Prompt-level counterfactual augmentation pilot on 36 prompt–seed instances. Confidence intervals are bootstrap 95% CIs.

Method	Mean margin ↑	Gain	Win rate
Baseline	0.0066 [0.0024, 0.0109]	0.0000	0.000
Prompt-level CCD-Attr	0.0042 [0.0004, 0.0086]	-0.0023 [-0.0043, -0.0005]	0.361

Counterfactual-aware reranking. For each prompt, we generate $K = 3$ independent candidates and select the candidate with the largest counterfactual margin. This is a post-hoc selection procedure; it does not alter the diffusion trajectory.

4.2 Counterfactual Margin

Let $s(x, p)$ be the similarity between image x and prompt p under a frozen vision-language model. We define

$$m(x, p) = s(x, p) - \max_{\tilde{p} \in \mathcal{C}(p)} s(x, \tilde{p}), \tag{7}$$

where $\mathcal{C}(p)$ is the set of attribute-focused counterfactual prompts. A larger value means that the image is more distinguishable from its counterfactual alternatives according to this scorer.

The margin is a diagnostic proxy rather than an independent measure of visual compositional correctness. In particular, reranking directly optimizes Eq. 7; therefore, margin results are not interpreted as independent evidence of improved generation quality.

Table 2: Compute-matched diagnostic on 60 prompt–instance evaluations. All methods select from the identical $K = 3$ candidate pool. Values are the mean and standard deviation of the CLIP counterfactual margin. These are selection diagnostics, not independent compositional-fidelity results.

Method	Candidate budget	Margin ↑
First-K	3	0.057 ± 0.567
Random-K	3	0.094 ± 0.558
CCD reranking	3	$\mathbf{0.303 \pm 0.571}$

4.3 Prompt-Level Counterfactual Augmentation

Table 1 shows that prompt-level counterfactual augmentation lowers the mean counterfactual margin by 0.0023. Its bootstrap 95% confidence interval excludes zero, and it wins on 36.1% of matched instances. Within this pilot, textual exclusions do not provide an effective way to express the intended structured contrast and may interfere with the original prompt semantics.

4.4 Compute-Matched Reranking Diagnostic

We additionally construct a fixed diagnostic suite containing 60 prompt–instance evaluations across four families: attribute binding, left–right relations, above–below relations, and containment. The suite contains 20 unique prompt forms, each evaluated with three independently seeded candidate pools. For every prompt–instance, all methods use the *same* pool of $K = 3$ images generated by SDXL-Turbo. We compare: (i) First-K, which returns the first candidate; (ii) Random-K, which uniformly selects one candidate; and (iii) CCD reranking, which selects the candidate with maximum Eq. 7.

As expected, CCD reranking selects images with a larger value of the score it optimizes (Table 2). Importantly, this comparison removes the candidate-budget confound in the original single-sample-versus-K-sample pilot: all three methods receive three generated candidates. It does not, however, establish that the selected images have better compositional fidelity, because the CLIP margin remains both the selection signal and the reported diagnostic.

4.5 Blinded Evaluation Protocol

To obtain an evaluation independent of the selection score, we construct 120 randomized pairwise comparison sheets. For each prompt–instance, CCD reranking is paired separately with First-K and Random-K. The method identities and image order are hidden from raters. Raters are shown the target prompt and two images, and choose *A*, *B*, *Tie*, or *Invalid* according to which image better satisfies the specified object identities, attribute bindings, and spatial or containment relations. They are instructed not to judge aesthetics or image sharpness unless these prevent interpretation of the composition.

The released evaluation package contains the candidate images, randomized pair sheets, frozen generation and selection settings, and one rating template per annotator. We will report blinded human results only after collecting ratings from at least three independent raters, including CCD win rate, paired bootstrap confidence intervals, a paired permutation test, and inter-rater agreement. We do not treat the unfilled rating templates as experimental evidence.

4.6 Summary

The pilot provides a robust negative result for prompt-level counterfactual augmentation. The compute-matched diagnostic further confirms that explicit counterfactual comparison is an effective candidate-selection signal under its own frozen margin. However, it remains an optimization-aligned diagnostic. The blinded evaluation protocol is designed to determine

whether this signal translates into independently judged compositional fidelity. The latent-level CCD formulation is not evaluated in the present experiments.

5 Limitations

This work evaluates prompt-level augmentation and post-hoc candidate reranking, rather than the latent-level CCD formulation. The reranker does not modify the denoising trajectory, and we do not instantiate a differentiable latent verifier f_ϕ or evaluate the guided update in Eq. (6).

The empirical evidence remains preliminary. The prompt-level augmentation study uses 12 prompts, while the compute-matched reranking diagnostic contains 20 unique prompt forms evaluated across three independently seeded candidate pools each. Although the latter controls the candidate-generation budget, it uses the same frozen CLIP counterfactual margin for selection and diagnostic reporting. Its higher margin therefore does not independently establish improved compositional fidelity.

We construct randomized, method-blind pairwise evaluation sheets for independent human assessment, but do not report uncollected ratings as evidence. A complete evaluation should include ratings from multiple independent annotators, inter-rater agreement, paired statistical tests, substantially broader prompt coverage, and established compositional benchmarks. Finally, the present experiments use SDXL-Turbo only; conclusions may not generalize to larger diffusion models, alternative samplers, or models with different prompt following behavior.

6 Discussion

Prompt-level counterfactual augmentation lowers the counterfactual margin in our pilot. One plausible explanation is that textual exclusions such as "not $\tilde{p}$" modify or ambiguously condition the original prompt without requiring the model to explicitly compare the intended composition against a nearby structural alternative. In this setting, natural-language negation is therefore not a reliable surrogate for structured contrast.

By contrast, reranking makes the comparison explicit. In the compute-matched diagnostic, CCD reranking, First-K, and Random-K all select from the identical pool of three generated candidates. CCD selects candidates with higher counterfactual margins than either budget-matched baseline. This result removes the candidate-budget confound in the original single-sample comparison, but it has a deliberately narrow interpretation: the counterfactual margin is useful for selecting images with higher values of that same margin.

The central open question is whether this selection criterion improves actual attribute binding and relational fidelity under an evaluation independent of the selector. The randomized blind evaluation protocol introduced here is intended to answer that question. More generally, the results motivate structured counterfactual comparison as a candidate-selection mechanism and as a possible future signal for latent-level diffusion guidance, rather than establishing a finished guidance method.

7 Conclusion

We formulate compositional text-to-image generation using structured counterfactual alternatives constructed through attribute swaps, relation inversions, and argument-role perturbations. Our pilot finds that directly appending counterfactual exclusions to a prompt reduces the frozen counterfactual margin. In contrast, post-hoc CCD reranking selects candidates with larger values of that margin.

A compute-matched diagnostic confirms this selection behavior when all methods receive the same $K = 3$ candidate pool. Because the reported margin is also the selection objective, we do not interpret this result as independent evidence of improved compositional generation. The

work instead establishes a constrained empirical observation: explicit structured comparison is more promising than textual negation as an inference-time use of counterfactual information.

Future work should complete independent blinded evaluation, test broader compositional benchmarks and diffusion backbones, and instantiate the proposed differentiable latent verifier for counterfactual-guided denoising.

Broader Impact. Better compositional fidelity may improve the reliability of generative tools for design, visualization, and education. Greater controllability can also be misused to create more convincing synthetic content, so deployment should include appropriate safeguards.

AI Use Disclosure

We used large language models as writing and programming assistants during manuscript preparation. Specifically, they assisted with language editing, LaTeX formatting, and drafting experimental code. The authors designed the research question, specified the experimental protocol, executed the experiments, inspected the generated outputs, verified the reported results, and take full responsibility for the scientific claims and the final manuscript. No AI-generated text, code, or result was included without author review and verification.

Reproducibility Statement

We provide the complete generation and selection protocol for the compute-matched reranking diagnostic, including the random seed, diffusion backbone, image resolution, sampling steps, candidate budget, prompt construction rules, counterfactual transformations, and frozen CLIP selection model. All methods select from the same generated candidate pools, and the released materials include the raw candidate images, per-instance selection diagnostics, randomized method-blind comparison sheets, and human-rating templates. The prompt-level augmentation experiment reports matched prompt–seed instances and bootstrap confidence intervals. These materials are intended to permit reproduction of the reported diagnostic results and independent completion of the blinded human evaluation.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[2] C. Saharia et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

[3] A. Ramesh et al. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

[4] A. Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[5] J. Ho and T. Salimans. Classifier-free diffusion guidance. *NeurIPS Workshop*, 2022.

[6] P. Dhariwal and A. Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021.

[7] A. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021.

[8] A. Hertz et al. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023.

[9] H. Chefer et al. Attend-and-excite: Attention-based semantic guidance. In *SIGGRAPH*, 2023.

[10] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

[11] Y. Li et al. GLIGEN: Open-set grounded text-to-image generation. In *CVPR*, 2023.

[12] N. Liu et al. Composable diffusion models for compositional generation. In *ICML*, 2023.

[13] S. Goyal et al. NegPrompt: Improving compositionality with negative prompts. *arXiv*, 2023.

[14] Y. Balaji et al. eDiffi: Text-to-image diffusion models with ensemble of experts. *arXiv*, 2022.

[15] A. Nichol et al. GLIDE: Towards photorealistic image generation. In *ICML*, 2022.

[16] O. Avrahami et al. Blended diffusion for text-driven editing. In *CVPR*, 2023.

[17] B. Poole et al. DreamFusion: Text-to-3D using 2D diffusion. In *ICLR*, 2023.

[18] A. Brock et al. CLIPDraw: Exploring text-to-drawing synthesis. *arXiv*, 2021.

[19] K. Crowson et al. VQ-diffusion: Vector quantized diffusion models. *arXiv*, 2023.

[20] W. Chen et al. Layout-to-image generation with diffusion models. In *CVPR*, 2023.

[21] J. Xu et al. Structured compositional generation with diffusion. In *ICCV*, 2023.

[22] G. Parmar et al. Causal diffusion for compositional generation. *arXiv*, 2024.

[23] R. Tang et al. Improving compositional reasoning in diffusion models. *arXiv*, 2024.

[24] Y. Li et al. Improving attention for compositional diffusion. *arXiv*, 2024.

[25] Z. Wang et al. Relation-aware diffusion for structured generation. *arXiv*, 2024.

[26] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[27] A. Ramesh et al. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

[28] C. Saharia et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

[29] A. Nichol et al. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022.

[30] R. Rombach et al. Stable diffusion: High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2022.

[31] J. Wang et al. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *NeurIPS*, 2023.

[32] P. Ghosh et al. Improving text-to-image generation with compositional prompts. In *ICCV*, 2023.

[33] B. Li et al. Object-aware text-to-image generation using diffusion models. In *CVPR*, 2023.

[34] W. Feng et al. Training-free structured diffusion guidance for compositional text-to-image generation. In *ICCV*, 2023.

[35] A. Hertz et al. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023.

[36] H. Chefer et al. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. In *SIGGRAPH*, 2023.

[37] N. Liu et al. Composable diffusion models for image synthesis. In *ICML*, 2023.

[38] Y. Li et al. GLIGEN: Open-set grounded text-to-image generation. In *CVPR*, 2023.

[39] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

[40] H. Huang et al. Direct inversion: Boosting diffusion-based image editing. In *ICCV*, 2023.

[41] J. Chen et al. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024.

[42] B. Esser et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

[43] L. Yang et al. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*, 2023.

[44] Y. Hu et al. Winoground: Probing vision and language models for visual compositionality. In *CVPR*, 2023.

[45] J. Wu et al. Generative visual instruction tuning for compositional reasoning. *arXiv preprint arXiv:2305.*, 2023.