

Right Looks, Wrong Reasons: Compositional Fidelity in Text-to-Image Generation

Mayank Vatsa¹, Aparna Bharati², and Richa Singh¹

¹IIT Jodhpur, India,

²Lehigh University, USA

{mvatsa, richa}@iitj.ac.in, apb220@lehigh.edu

Abstract

The architectural blueprint of today’s leading text-to-image models contains a fundamental flaw: an inability to handle logical composition. This survey investigates this breakdown across three core primitives—negation, counting, and spatial relations. Our analysis reveals a dramatic performance collapse: models that are accurate on single primitives fail precipitously when these are combined, exposing severe interference. We trace this failure to three key factors. First, training data show a near-total absence of explicit negations. Second, continuous attention architectures are fundamentally unsuitable for discrete logic. Third, evaluation metrics reward visual plausibility over constraint satisfaction. By analyzing recent benchmarks and methods, we show that current solutions and simple scaling cannot bridge this gap. Achieving genuine compositionality, we conclude, will require fundamental advances in representation and reasoning rather than incremental adjustments to existing architectures.

Introduction

Text-to-image (T2I) generation systems such as DALL·E 3 (Betker et al. 2023), Stable Diffusion (Rombach et al. 2022), Imagen (Saharia et al. 2022), and Parti (Yu et al. 2022) can produce remarkably photorealistic images that capture style, aesthetics, and individual concepts with impressive fidelity. Despite their visual sophistication, these models systematically fail at compositional reasoning, i.e., the ability to satisfy multiple constraints such as counting, attribute binding, spatial relations, and negation, simultaneously. For example, the prompt “exactly three red apples to the left of a vase with no flowers” is trivial for humans but challenging for these models. As illustrated in Figure 1, even when a model can satisfy constraints individually, combining them causes the performance to drop significantly (Huang et al. 2023; Hsieh et al. 2023; Li et al. 2025a; Cong et al. 2025).

The failure modes of these models are now well documented. For negation, models often ignore cues and permit forbidden content (Alhamoud et al. 2025; Singh et al. 2025; Patel et al. 2024). For counting, errors grow with target cardinality, and recent work proposes targeted remedies within T2I pipelines (Paiss et al. 2023; Binyamin et al. 2025; Zafar

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

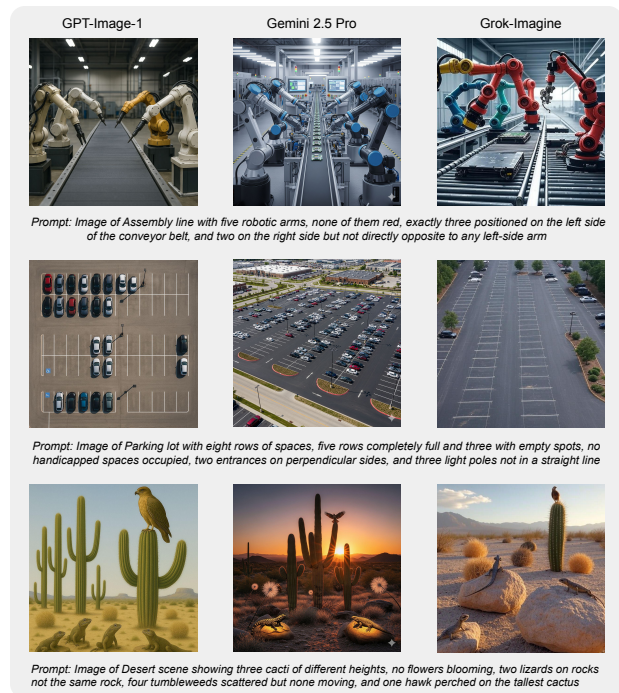


Figure 1: Examples of state-of-the-art Text-to-Image Generation models failing to correctly bind attributes and satisfy spatial relations in complex prompts.

et al. 2024). For spatial relations, systems may satisfy local or pairwise predicates while violating global layout consistency (Gokhale et al. 2023; Kamath et al. 2023; Cho et al. 2024). These failures stem from architectural biases such as continuous attention, which poorly handles discrete logic, and data sparsity, as web-scale captions rarely cover structured, negated, or high-count cases.

These limitations have real-world consequences. Educational platforms cannot reliably render illustrations with specific counts. Technical documentation tools misplace components in assembly diagrams. Under too many constraints, scientific and medical illustrations can show anatomically or physically impossible results. Although extensive iteration can lead to intended compositions, this undermines direct language control and hinders AI’s potential in high-stakes domains. Compositional fidelity is thus a prerequisite for re-

liable and trustworthy deployment (Mittal et al. 2024).

This survey analyzes *compositional fidelity* through the lens of *compositional primitives*, defined as the semantic mechanisms that combine to specify a scene. These include attribute binding; size and scale relations, actions and temporal order, part-whole structure, logical connectives (AND/OR/XOR), quantifier scope and coreference, symmetry and topology beyond pairwise relations, material/texture binding, occlusion reasoning, and 3D/viewpoint consistency (Thrush et al. 2022; Ma et al. 2023). We focus on three fundamental primitives: negation (absence), counting (cardinality), and spatial relations (geometric/topological layout).

This paper has three key contributions. (1) We provide a formal account of the intersectional failure, explaining why progress on isolated capabilities does not extend to joint prompts, and connect them to constrained optimization and combinatorial hardness. (2) We synthesize methods: data augmentation and contrastive training for negation; architectural strategies and mixture-of-experts for counting; layout and structural control for spatial relations; and hybrid neural-symbolic pipelines for joint composition. (3) We review 15 benchmarks, showcasing a shift from human studies to automated and adversarial evaluations, and discuss strengths, biases, and gaps. Finally, we propose a research direction that spans theory, architecture, training, and evaluation to bridge visual plausibility and logical faithfulness¹.

Problem Formulation

Let $\mathcal{X} = [0, 1]^{H \times W \times 3}$ denote the RGB image space where H and W are the image height and width, and \mathcal{Y} the space of natural-language prompts. A T2I model $G_\theta : \mathcal{Y} \rightarrow \Delta(\mathcal{X})$ induces the conditional distribution $p_\theta(x|y)$ for $y \in \mathcal{Y}$, where $\Delta(\mathcal{X})$ denotes the space of probability distributions over \mathcal{X} . Modern systems, including diffusion-based (e.g., Stable Diffusion (Rombach et al. 2022)), autoregressive (e.g., Parti (Yu et al. 2022)), and proprietary systems (e.g., DALL-E 3 (Betker et al. 2023)), map linguistic concepts to visual content using massive paired corpora. Despite impressive perceptual quality, early probes documented *compositional* failures on instructions that require logical structure and grounding (Saharia et al. 2022). Similar compositional challenges have been observed in vision-language models like CLIP (Radford et al. 2021), which struggle with relational semantics despite strong object-level understanding.

Definition & Significance of Primitives in T2I Gen.

Negation - The Challenge of Concept Absence: Negation requires reasoning about what should *not* appear (e.g., “a bowl of fruit *without* bananas”). Let y_{aff} be an affirmative prompt and y_{neg} its negated counterpart. With latent z , negated prompts plausibly span more admissible scenes (Vempati et al. 2023; Patel et al. 2024; Alhamoud et al. 2025), $\mathcal{H}(p(z | y_{\text{neg}})) > \mathcal{H}(p(z | y_{\text{aff}}))$ where $\mathcal{H}(\cdot)$

¹While prior surveys center on architectures, scaling trends, or benchmarks (Thrush et al. 2022; Ma et al. 2023; Vempati et al. 2023; Yuksekgonul et al. 2023a), we focus on compositional fidelity, introducing a primitive-based taxonomy that explains why scale alone falls short and points to promising solutions.

denotes the Shannon entropy. In T2I prompts, negation appears in a range of recurring expressions that operate at different semantic levels. **Morphological** negation uses affixes (e.g., *unstriped*, *nontoxic*, *sugarless*) to flip attributes. **Lexical/privative** adjectives (e.g., *empty*, *barefoot*, *devoid of*) express the absence of object or content. **Syntactic/clausal** cues (e.g., *no*, *not*, *without*, *neither...nor*) deny entity presence or relations. **Quantified** negation (e.g., *no N*, *fewer than n*) sets zero or upper-bound constraints on counts. **Relational** negation (e.g., *not left of*, *not touching*) denies specific spatial predicates.

Why this matters in T2I: Negation bridges explicit linguistic markers and implicit visual signals of absence—a mapping current models handle poorly. It is rare in web captions (Truong et al. 2023; Hossain et al. 2020) and competes with strong co-occurrence priors. In practice, we observe recurring failure patterns (Alhamoud et al. 2025). Models ignore negation markers, allowing forbidden objects to appear. Scope is misapplied, with “not” attaching to the wrong phrase. Models over-suppress content as a shortcut. UI-level “negative prompts” alter text guidance but do not guarantee semantic absence. Occlusion confounds evaluation, as hidden items can be mistaken for true absence or presence by automated judges. These issues amplify on joint prompts like “*exactly two red apples not touching*”, which require enforcing multiple types of absence while meeting precise positive constraints. In high-stakes domains (e.g., medical: “*no fracture*”), such failures yield misleading imagery.

Counting - Precise Cardinality Control: Counting, which demands precise control over the number of instances (e.g., “*exactly five roses*”), exposes a core architectural limitation of modern generators. Because Transformer attention operates in parallel, it lacks an explicit enumerator, forcing numerosity to be encoded implicitly within continuous weights rather than through discrete counting steps. This fundamental mismatch between the task and the architecture leads to a predictable pattern of failure. Empirically, accuracy degrades as the target number n increases, with errors compounding in cluttered or occluded scenes (Huang et al. 2023; Paiss et al. 2023; Zafar et al. 2024; Binyamin et al. 2025). Rather than a gradual decline, this degradation is super-linear: error growth consistently fits a power law with exponent $\beta \in [1.2, 1.5]$ across multiple benchmarks,

$$\text{Error}(n) \approx \Theta(n^\beta), \quad \beta \in [1.2, 1.5]. \quad (1)$$

Counting in T2I prompts manifests in a variety of complex linguistic forms. Models must handle *exact numerals* (“*exactly three red apples*”), interpret *bounded ranges* (“*between three and five*”), and resolve *vague quantifiers* (“*few*,” “*several*”). The task can also be relational, involving *comparatives* (“*more apples than pears*”), or compound, requiring *multi-type specifications* (“*three apples and two pears*”). Furthermore, counts are often distributed across the scene via *spatially partitioned tallies* (“*two vases on the left, one on the right*”). Fulfilling these prompts requires the model to successfully individuate instances, bind their attributes, and maintain count coherence across the entire composition.

Why this matters in T2I: Counting exposes the limits of discrete reasoning in architectures that operate on con-

tinuous representations. We observe consistent failure patterns (Kirstain et al. 2023; Binyamin et al. 2025) such as objects accidentally duplicated or merged under clutter. Attribute-binding leakage causes numbers to apply to wrong subsets. Comparative constraints fail when one category disappears rather than adjusting counts appropriately. These issues compound with other primitives. A prompt like “*exactly three red apples to the left of a vase with no flowers*” demands precise counting with correct attribute binding, specific spatial arrangement, and enforced absence of both extra apples and flowers. Any degradation in individuation, binding, or layout violates the instruction, making counting central to compositional fidelity.

Spatial Relations - Geometric & Topological Reasoning: Spatial relations dictate geometric object arrangements. A prompt like “a blue cube *on* a red sphere” requires parsing linguistic relations into geometrically consistent scenes (Gokhale et al. 2023; Kamath et al. 2023) and can be decomposed as:

$$\begin{aligned} \exists c, s \text{ s.t. } & \text{IsCube}(c) \wedge \text{IsSphere}(s) \wedge \text{Color}(c, \text{blue}) \\ & \wedge \text{Color}(s, \text{red}) \wedge \text{On}(c, s). \end{aligned} \quad (2)$$

While models often satisfy unary predicates (types, colors), enforcing the relational predicate $\text{On}(c, s)$ proves notably harder. In a viewer-centric frame, let object placements be axis-aligned rectangles $b_c = (x_c, y_c, w_c, h_c)$ and $b_s = (x_s, y_s, w_s, h_s)$. $\text{On}(c, s)$ requires the following:

$$\begin{aligned} |y_c + h_c - y_s| &< \epsilon, \\ [x_c, x_c + w_c] \cap [x_s, x_s + w_s] &\neq \emptyset, \end{aligned} \quad (3)$$

Here, (x, y, w, h) denote the top-left position and size of each axis-aligned bounding box, and ϵ is a proximity threshold. The first constraint ensures the bottom of c is near the top of s , while the second ensures horizontal overlap. Object- or world-centric frames yield different but analogous conditions.

To generate coherent scenes, models must deconstruct the complex language of *spatiality*, which manifests in multiple forms (Liu, Emerson, and Collier 2023). The process begins with parsing fundamental directional relations that specify relative positions, such as “left of” or “above.” Building on this, topological relations encode concepts of containment and overlap, like “in” or “on.” The challenge becomes more fine-grained with proximity relations that indicate distance (“near,” “touching”) and alignment patterns that create structured arrangements (“in a row”). Furthermore, models must grasp physical plausibility through *support* relations (“on top of”) and manage complex distributions with partitioned layouts (“two vases on the left, one on the right”). For clarity, we assume a viewer-centric interpretation for these relations unless a prompt specifies otherwise.

Why this matters in T2I: Spatial reasoning connects textual semantics with geometric feasibility. Current models exhibit characteristic failures (Vempati et al. 2023; Zhang et al. 2024). Pairwise predicates hold locally but conflict globally. Directional relations flip under viewpoint changes. Support relations are nominally satisfied while violating

physical plausibility. Contact constraints fail due to slight boundary overlaps. Structured patterns drift under clutter. These errors intensify on joint prompts like “*exactly three red apples to the left of a vase not touching it*”, which simultaneously require precise counting, correct attribute binding, specific layout, and absence of forbidden contacts. In high-stakes settings (e.g., medical: “lesion *left of* the hippocampus”), mis-grounded spatial relations yield visually credible but semantically incorrect images.

Joint Compositionality - The Compounding Challenge: The most significant failures occur when primitives are combined. Consider “two blue cubes to the left of a vase *with no flowers*”—a prompt joining counting, spatial relations, attribute binding, and negation. Failure rates on such prompts are higher than single primitives (Huang et al. 2023).

Let $F_\theta(y)$ denote compositional faithfulness: the probability that all constraints in prompt y are satisfied. Under independence of counting, spatial, and negation, joint success equals the product of individual rates.

$$F_\theta^{\text{ind}}(y) := F_\theta^{\text{cnt}}(y) F_\theta^{\text{spat}}(y) F_\theta^{\text{neg}}(y). \quad (4)$$

However, we consistently observe *submultiplicative* performance (Huang et al. 2023). We quantify interference via:

$$\rho(y) := \frac{F_\theta(y)}{F_\theta^{\text{ind}}(y)} \quad (5)$$

$\rho(y) = 1$ indicates independence and $\rho(y) < 1$ indicates intersectional failures. Equivalently, the additive gap $\Delta_{\text{joint}}(y) = F_\theta^{\text{ind}}(y) - F_\theta(y) > 0$ measures the deficit. We define interference as *pronounced* when $\rho(y) \leq \tau_\rho$ for threshold $\tau_\rho \in (0, 1)$ (e.g., 0.75). To illustrate the compounding nature, consider a scenario where individual primitives achieve moderate success rates (e.g., 70% each). Under independence, joint faithfulness would be $0.7^3 = 34.3\%$. However, due to interference ($\rho(y) \approx 0.58$), actual performance drops to $\approx 20\%$, highlighting the challenge.

Why this matters in T2I: Joint prompts expose hidden interactions between capabilities. Models exhibit constraint trading: enforcing layout breaks counting; honoring negation removes desired context; correct counts appear with incorrect attribute binding; and pairwise relations hold locally while global arrangement fails (Kamath et al. 2024). These behaviors reflect architectural limitations such as continuous attention without explicit enumeration and training biases such as sparse negation exposure and strong co-occurrence priors (Yuksekgonul et al. 2023b). The submultiplicative gap ($\rho(y) < 1$) separates surface realism from genuine compositional reasoning, motivating targeted methods like inference-time composition (Feng et al. 2023) and layout-guided image generation (Zheng et al. 2023).

Literature Review

Having established the challenges associated with the three compositional primitives, we now survey the benchmarks and methods proposed to measure and address these failures.

Methods for Addressing Negation: Early work traced negation blindness in T2I to data sparsity and architectural bias (Hossain et al. 2020; Kassner and Schütze 2020;

Benchmark	Size	Auto	Neg	Count	Spatial	Primary Focus
DrawBench (Saharia et al. 2022)	200	X	X	✓	✓	Human-eval comprehensiveness
PartiPrompts (Yu et al. 2022)	1,600+	X	✓	✓	✓	Multi-aspect evaluation
Winoground (Thrush et al. 2022)	400 ex. (1,600 pairs)	✓	X	X	✓	Visio-linguistic composition
CREPE (Ma et al. 2023)	370K+	✓	✓	X	✓	Systematic compositionality
VALSE (Vempati et al. 2023)	~8,782	✓	✓	✓	✓	Linguistic variations
ARO (Yuksekgonul et al. 2023a)	50K+	✓	X	X	✓	Attribution, relation, order
POPE (Li et al. 2023a)	variable	✓	✓	X	X	Object hallucination
T2I-CompBench (Huang et al. 2023)	6,000	✓	✓	✓	✓	Comprehensive automated eval
DSG (Cho et al. 2024)	1,060	✓	X	✓	✓	Scene-graph grounding
TIFA (Hu et al. 2023)	~4,000	✓	✓	✓	✓	VQA-based faithfulness
PhysiComp (Chen et al. 2024)	~1K	✓	X	✓	✓	Physical plausibility
SugarCrepe (Hsieh et al. 2023)	>1,000	✓	✓	✓	✓	Hard negative probes
CC-Neg (Singh et al. 2025)	228K	✓	✓	X	X	Negation training/eval
NegBench (Alhamoud et al. 2025)	79K	✓	✓	X	X	Comprehensive negation
CoCoCount (Binyamin et al. 2025)	200	✓	X	✓	X	Counting specialization

Table 1: Comparing key compositional benchmarks for T2I evaluation. **Auto**: supports automated evaluation. **Neg/Count/Spatial**: explicitly tests negation/counting/spatial relations. Sizes are from original papers or official repos; “variable” indicates no fixed canonical size.

Truong et al. 2023), and mainly proposed three mitigation avenues. Firstly, contrastive strategies such as TripletCLIP which builds triplets differing only by negation markers and NegCLIP which generates adversarial negatives automatically (Yuksekgonul et al. 2023a) show promise but struggle with complex negation. Secondly, data augmentation strategies address the scarcity directly (Brooks, Holynski, and Efros 2023). CC-Neg (Singh et al. 2025) dataset contains 228K image-caption pairs that cover morphological, syntactic, and semantic negation across diverse visual contexts and helps in improving negation understanding. Beyond data, architectural methods preserve negation by reducing word loss with asymmetric visual-semantic embeddings (Liu et al. 2025), modeling it as an energy constraint (Feng et al. 2023), or encoding absence with empty boxes (Li et al. 2023b). These structural approaches generally outperform data-driven contrastive methods, which struggle with complex negation and are prone to bias or overfitting from synthetic augmentations (Hsieh et al. 2023).

Approaches to Counting: Efforts to improve counting in T2I models follow two main paths. The first is data-centric: DALL-E 3, for instance, focuses on enhancing caption quality to ensure training data contains precise numerical information (Betker et al. 2023). The second path involves architectural innovations. These include modifying the attention mechanism, as seen in Stable Diffusion 3 and SDXL (Esser et al. 2024; Podell et al. 2023; Kang et al. 2025), using a mixture-of-experts to increase numeric capacity (Xue et al. 2023), or applying bounded attention to prevent the merging of distinct objects (Dahary et al. 2024). However, the success of these is often limited to a small number of objects.

To enhance counting in T2I generation, layout-based methods enforce precise counts via spatial conditioning, using bounding boxes (Li et al. 2023b), semantic regions (Yang et al. 2023b), or structural maps (Zhang, Rao, and Agrawala 2023). These, however, limit the flexibility of text-only prompts. Hybrid approaches, like LLM-grounded diffusion, restore this by parsing text into structured scene

representations, reducing constraint interference ($\rho(y) < 1$) (Lian et al. 2023). Dynamic methods, such as RPG’s iterative refinement with multimodal feedback (Yang et al. 2024), balance counting with other constraints, while Visual Programming decomposes tasks into modular steps (Gupta and Kembhavi 2023). Alternatively, CountGen embeds a differentiable counting loss in training (Binyamin et al. 2025), and detection-driven methods iteratively refine outputs to ensure accurate counts (Zafar et al. 2024).

Spatial Relation Methods: Spatial reasoning improvements ground language in geometric control. Many approaches use layout-based methods for direct spatial conditioning. Some methods integrate this control during training. GLIGEN, for instance, uses box conditioning to guide generation (Li et al. 2023b). ControlNet leverages structural cues like edges and keypoints (Zhang, Rao, and Agrawala 2023). LayoutDiffusion also employs semantic segmentation for object adjacencies (Zheng et al. 2023). Other methods offer training-free control by manipulating attention at inference time. ZestGuide, for example, achieves zero-shot control via attention manipulation (Couairon et al. 2023). Similarly, attention refocusing optimizes layout supervision during the sampling process (Phung, Ge, and Huang 2024).

Approaches to spatial grounding can be broadly categorized into sampling-based, attention-based, architectural, and 3D-aware strategies. Sampling-based methods include Composable Diffusion, which decomposes spatial relations into separate energy functions that combine during generation (Feng et al. 2023), and MultiDiffusion, which enforces scene coherence by generating multiple aligned views that are fused into a consistent whole (Bar-Tal et al. 2023). Attention-based methods such as Attend-and-Excite refine attention maps iteratively until spatial constraints are satisfied (Chefer et al. 2023; Rassin et al. 2023). Architectural approaches embed spatial reasoning more directly: Set-of-Mark prompting uses visual markers to disambiguate references (Yang et al. 2023a), while CoMPaSS curates accurate training pairs with spatial constraints and strength-

ens text encoders through ordered token encoding (Zhang et al. 2024). Finally, 3D-aware methods such as Zero123 learn viewpoint-consistent representations, ensuring coherence across multiple perspectives (Liu et al. 2023). However, layout-based methods often require additional structural inputs that may not scale to complex scenes.

Joint Compositional Methods: To address the performance collapse when multiple primitives are combined, specialized approaches have been developed to mitigate interference. One strategy is inference-time composition, which processes different constraints separately to prevent them from conflicting with one another during generation (Feng et al. 2023; Lian et al. 2023; Xue et al. 2023). Alternatively, training-based approaches target compositional generalization. For instance, compositional augmentation integrates primitives systematically in training data to enhance joint handling (Li et al. 2025a; Cong et al. 2025). On the other hand, EvoGen leverages curriculum learning with language models and visual checkers to progressively refine compositional skills (Han et al. 2025). Despite their promise, these training strategies often face challenges. Catastrophic forgetting and distribution shift often arise, where enhancing one primitive via augmentation can degrade others due to overfitting (Hsieh et al. 2023).

Evaluation Benchmarks: Progress in compositional generation has been enabled by increasingly sophisticated evaluation frameworks. Table 1 summarizes fifteen major benchmarks spanning 2022-2025. Early human-evaluated benchmarks like DrawBench (Saharia et al. 2022) and PartiPrompts (Yu et al. 2022) established the problem space but faced scalability limitations. The shift to automated evaluation enabled systematic assessment at scale, with T2I-CompBench evaluating prompts across all three primitives (Huang et al. 2023) and CREPE providing 370K+ variations (Ma et al. 2023). More recent benchmarks emphasize adversarial testing through hard negatives and VQA-based probing. While existing frameworks cover negation, counting, and spatial relations, critical gaps remain. Key missing areas include temporal reasoning and complex multi-object interactions. Furthermore, assessing physical plausibility requires moving beyond current specialized efforts.

Synthesis and Open Challenges

Our analysis of negation, counting, and spatial relations highlights fundamental limitations in current text-to-image generation systems. The submultiplicative performance degradation quantified in Equation 5 emerges from deeper architectural and methodological constraints. We summarize these findings into actionable insights and present future directions toward compositionally faithful generation.

The Data-Architecture Mismatch: The compositional failures we observe stem from a fundamental misalignment between training data distributions and architectural inductive biases. Consider the joint distribution of compositional primitives in training data:

$$p_{\text{train}}(\text{neg}, \text{count}, \text{spat.}) \approx p(\text{neg}) \cdot p(\text{count}) \cdot p(\text{spat.}), \quad (6)$$

an independence approximation that highlights the extreme sparsity of joint cases. In reality, not only are the marginal

probabilities of each primitive low, but their co-occurrence is even rarer, with some pairs exhibiting negative correlations. For example, in popular datasets, captions with explicit negation are rare: MS COCO ($\sim 0.4\%$), CC3M (1.63%), CC12M ($\sim 2.5\%$), and LAION-400M ($\sim 0.6\%$); larger corpora such as LAION-2B and DataComp-1B show similar low sub-percent to low single-digit rates depending on cues like *no/not/without* (Bui et al. 2025; Park et al. 2025). High-count scenes ($n > 5$) appear in under 2% of samples, and complex spatial arrangements with multiple relations constitute less than 5%. Thus, the effective probability of encountering prompts that combine negation, counting, and spatial constraints during training is vanishingly small.

The architectural response to this sparsity is predictable but problematic. Continuous attention mechanisms learn to approximate the majority modes:

$$x^*(y) = \arg \max_{x \in X} S_\theta(x | y),$$

$$S_\theta(x | y) = \sum_i \alpha_i f_i(x, y) + \lambda R(x). \quad (7)$$

where $S_\theta(x|y)$ denotes the composite score over constraint-specific functions $f_i(x, y)$ with learned weights α_i , and $R(x)$ encodes regularization or visual (realism) priors. The regularization term dominates when compositional requirements conflict with learned priors, explaining why models generate plausible yet unfaithful images. This mismatch manifests differently across primitives. Models lack explicit mechanisms to represent negation:

$$\mathcal{L}_{\text{neg}} = \min_{\theta} \mathbb{E}_{y \sim p(y)} [D(G_\theta(y), \emptyset_{\text{target}})] \quad (8)$$

where $\emptyset_{\text{target}}$ denotes the absence of specific content and D is a distance/divergence measure. Current architectures cannot differentiate this from general content suppression, leading to over-deletion or complete ignorance.

Emergent Complexity in Joint Composition: The submultiplicative performance reflects genuine computational complexity rather than simple capability multiplication. When primitives combine, the constraint satisfaction problem becomes NP-hard in the general case². Consider a prompt with n objects, m spatial relations, and k negation constraints. The search space grows as: $|\mathcal{S}| = O(n! \cdot 2^m \cdot \binom{n}{k})$. Current models use greedy local search, leading to characteristic failures: objects meet local constraints but break global consistency, counts hold until spatial layout is enforced, and negation applies to the wrong scope when combined with relations. We can formalize the joint generation problem as a constrained optimization: $x^* = \arg \max_{x \in \mathcal{X}} p_\theta(x|y)$ such that $C_{\text{count}}(x, y) = 1$, $C_{\text{spatial}}(x, y) = 1$, and $C_{\text{neg}}(x, y) = 1$, where C_i are binary constraint satisfaction functions. Current methods only partially satisfy these constraints, leading to compositional failures.

²Satisfying sets of qualitative spatial constraints is NP-complete in general (RCC-8); the analogous temporal algebra is NP-hard/NP-complete; and related layout problems such as cartographic label placement are NP-hard, highlighting the combinatorial nature of global spatial composition (Renz and Nebel 1999; Nebel and Bürckert 1995; Formann and Wagner 1991).

Inter-Primitive Interference: The submultiplicative performance ($\rho(y) < 1$) arises from interference between primitives, exacerbating issues like hallucinations and physical implausibility. For instance, enforcing negation (e.g., “no flowers”) may trigger hallucinations of unrelated objects due to strong co-occurrence priors in training data (Li et al. 2023a). Similarly, combining counting and spatial constraints can produce physically implausible scenes, such as objects defying gravity when satisfying local relations (Chen et al. 2024). These interactions highlight the need for models to disentangle primitives explicitly, through structured representations or constraint prioritization, to mitigate unintended outputs while preserving compositional fidelity.

Critical Research Gaps

Theoretical Foundations: The theory for compositional generation remains largely unwritten and critical questions unanswered. For instance, the computational lower bounds for faithful composition are unknown, and we cannot yet formally characterize which compositional patterns are learnable from the data. Mechanistic interpretability has shown that attention heads specialize in different primitives (negation and counting) (Rassin et al. 2023; Yuksekogonul et al. 2023b). However, the interaction mechanisms between these specialized components remain largely opaque to current analytical methods. Empirical evidence already suggests that merely increasing model scale is not a viable solution; larger models show marginal gains on isolated primitives but suffer the same performance collapse on joint tasks (Esser et al. 2024; Betker et al. 2023). This points to a deeper architectural limitation. Exploring connections to classical constraint satisfaction, such as SAT solving or neuro-symbolic methods (Mao et al. 2019), could provide a path forward. However, these parallels are currently underexplored, and formalizing them could inspire the architectural innovations needed to overcome today’s compositional failures.

Evaluation Methodology: Current evaluation paradigms reveal significant limitations across the literature. Human evaluation provides nuanced judgment but suffers from inconsistency and lacks scalability (Saharia et al. 2022). Automated metrics using object detectors and VQA models introduce systematic biases (Hu et al. 2023). The correlation between existing metrics and human compositional assessment remains weak, particularly for complex multi-constraint prompts (Ghosh, Hajishirzi, and Schmidt 2023). Natural language ambiguity makes ground truth specification difficult—“a few cats near some boxes” admits multiple valid interpretations. Occlusion and viewpoint variation further complicate accurate counting and spatial assessment. The trade-off between aesthetic quality and compositional fidelity is poorly characterized, with humans often preferring visually appealing but compositionally incorrect images. Evaluation frameworks that measure compositional generalization on novel combinations are an open challenge.

Architectural Innovations: Transformers excel at semantic understanding but lack explicit mechanisms for logical operations. Recent neuro-symbolic approaches show promise (Lian et al. 2023), yet integration remains empirically driven rather than principled. Architectural directions

such as modular architectures that decouple reasoning from generation, memory-augmented networks for explicit state, and hierarchical models separating “what,” “where,” and “how many”, merit deeper investigations. Intermediate representations such as scene graphs and layouts improve fidelity (Li et al. 2023b) but depend on costly annotations, leaving the challenge of learning structured representations from text alone unresolved (Feng et al. 2023).

Training Paradigms: Current training objectives, such as reconstruction for autoregressive models and denoising for diffusion models, optimize average-case performance on common patterns. These objectives misalign with compositional requirements that demand worst-case guarantees on logical constraints. The literature lacks frameworks explicitly designed for compositional fidelity. Curriculum learning with progressive complexity helps specific primitives but requires careful engineering (Mao et al. 2019; Srinivasan, Ren, and Thomason 2023). Compositional data augmentation improves targeted capabilities but struggles with diverse or out-of-distribution prompts (Hsieh et al. 2023). Reinforcement learning from human feedback could align models with compositional objectives; however, reward specification for complex constraints is an open problem (Li et al. 2025b). In addition, data efficiency questions persist in the literature. Synthetic data risk distribution shift and massive augmentation yields only marginal gains, suggesting fundamental inefficiency in learning compositional patterns.

Future Directions: The path forward requires coordinated advances across multiple fronts. Better benchmarks must measure compositional generalization rather than memorization. Architectural innovations should balance continuous and discrete reasoning (Thakral et al. 2025; Dosi et al. 2025). Training paradigms must align logical objectives with statistical likelihood, and theoretical frameworks should guide development beyond empirical trial-and-error. The submultiplicative performance gap remains a critical metric of progress, and closing it demands fundamental advances. As the field matures, compositional fidelity becomes not just a technical challenge but a prerequisite for practical deployment. Techniques developed for negation, counting, and spatial relations will extend to temporal reasoning, causal relations, and abstract concepts, paving the way for controllable, reliable, and truly intelligent image generation.

Conclusion

This survey examined compositional fidelity in text-to-image generation through negation, counting, and spatial relations. The submultiplicative performance degradation highlights systemic limitations, showing that current architectures—optimized for visual plausibility—fundamentally mismatch the discrete logic required for compositional reasoning. Progress demands architectural innovations that integrate explicit symbolic mechanisms, training objectives that prioritize constraint satisfaction over perceptual quality, and evaluation metrics that measure true compositional understanding. Solving these core challenges is a prerequisite for trustworthy deployment in high-stakes applications and will enable models to reason about more abstract concepts like time and causality.

Acknowledgments

Vatsa and Singh are supported through Srijan: Center of Excellence on Generative AI.

References

- Alhamoud, K.; Alshammari, S.; Tian, Y.; Li, G.; Torr, P. H. S.; Kim, Y.; and Ghassemi, M. 2025. Vision-Language Models Do Not Understand Negation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 29612–29622.
- Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 1737–1752.
- Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Lee, J.; Guo, Y.; Manassra, W.; Dhariwal, P.; Chu, C.; Jiao, Y.; and Ramesh, A. 2023. Improving Image Generation with Better Captions. Technical report, OpenAI.
- Binyamin, L.; Tewel, Y.; Segev, H.; Hirsch, E.; Rassin, R.; and Chechik, G. 2025. Make It Count: Text-to-Image Generation with an Accurate Number of Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13242–13251.
- Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instruct-Pix2Pix: Learning to Follow Image Editing Instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18392–18402.
- Bui, N.-T.; Hoang, D.-H.; Trinh, Q.-H.; Tran, M.-T.; Nguyen, T.; and Gauch, S. 2025. NeIn: Telling What You Don't Want. arXiv:2409.06481.
- Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. In *ACM Transactions on Graphics (TOG)*, volume 42, 1–10.
- Chen, D.; Singh, K.; Ivanova, A.; and Rohrbach, A. 2024. PhysiComp: A Benchmark for Physical and Causal Compositionality in Text-to-Image Generation. In *Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Cho, J.; Hu, Y.; Gao, R.; Gu, M.; Gebreab, J.; Dixit, A.; and Kwon, Y. 2024. Davidsonian Scene Graph: Improving Reliability in Fine-grained Evaluation for Text-to-Image Generation. In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Cong, Y.; Min, M. R.; Li, L. E.; Rosenhahn, B.; and Yang, M. Y. 2025. Attribute-Centric Compositional Text-to-Image Generation. *International Journal of Computer Vision*, 133(7): 4555–4570.
- Couairon, G.; Careil, M.; Cord, M.; Lathuilière, S.; and Verbeek, J. 2023. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2174–2183.
- Dahary, O.; Patashnik, O.; Aberman, K.; and Cohen-Or, D. 2024. Be Yourself: Bounded Attention for Multi-Subject Text-to-Image Generation. In *European Conference on Computer Vision (ECCV)*, 432–448.
- Dosi, M.; Chiranjeev, C.; Thakral, K.; Vatsa, M.; and Singh, R. 2025. Harmonizing Geometry and Uncertainty: Diffusion with Hyperspheres. In *International Conference on Machine Learning*.
- Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; Podell, D.; Dockhorn, T.; English, Z.; Lacey, K.; Goodwin, A.; Marek, Y.; and Rombach, R. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. *arXiv preprint arXiv:2403.03206*.
- Feng, W.; He, X.; Fu, T.-J.; Jampani, V.; Akula, A.; Narayanan, P.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2023. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. In *The Eleventh International Conference on Learning Representations (ICLR)*.
- Formann, M.; and Wagner, F. 1991. A Packing Problem with Applications to Lettering of Maps. In *Proceedings of the 7th Annual ACM Symposium on Computational Geometry (SoCG)*, 281–288. ACM.
- Ghosh, D.; Hajishirzi, H.; and Schmidt, L. 2023. GENEVAL: An Object-focused Framework for Evaluating Text-to-Image Alignment. In *Advances in Neural Information Processing Systems*, volume 36, 32267–32283.
- Gokhale, T.; Palangi, H.; Nushi, B.; Vineet, V.; Horvitz, E.; Kamar, E.; Baral, C.; and Yang, Y. 2023. Benchmarking Spatial Relationships in Text-to-Image Generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 6566–6575.
- Gupta, T.; and Kembhavi, A. 2023. Visual Programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14953–14962.
- Han, X.; Jin, L.; Liu, X.; and Liang, P. P. 2025. Progressive Compositionality in Text-to-Image Generative Models. In *The Thirteenth International Conference on Learning Representations*.
- Hossain, M. M.; Anastasopoulos, A.; Blanco, E.; and Palmer, A. 2020. It's Not a Non-Issue: Negation as a Source of Error in Machine Translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3869–3885.
- Hsieh, C.-Y.; Zhang, J.; Ma, Z.; Kembhavi, A.; and Krishna, R. 2023. SugarCrepe: Fixing Hackable Benchmarks for Vision-Language Compositionality. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Hu, Y.; Liu, B.; Kasai, J.; Wang, Y.; Ostendorf, M.; Krishna, R.; and Smith, N. A. 2023. TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 20349–20360.
- Huang, K.; Sun, K.; Xie, E.; Li, Z.; and Liu, X. 2023. T2I-CompBench: A Comprehensive Benchmark for Compositional Text-to-Image Generation. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

- Kamath, A.; Clark, J.; Gupta, T.; Kolve, E.; Hoiem, D.; and Kembhavi, A. 2023. What’s “up” with Vision-Language Models? Investigating their Struggle with Spatial Reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8645–8661.
- Kamath, A.; Hsieh, C.-Y.; Chang, K.-W.; and Krishna, R. 2024. The Hard Positive Truth About Vision–Language Compositionality. In *Computer Vision – ECCV 2024*, 37–54. Springer.
- Kang, W.; Galin, K.; Koo, H. I.; and Cho, N. I. 2025. Counting Guidance for High Fidelity Text-to-Image Synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 4221–4231.
- Kassner, N.; and Schütze, H. 2020. Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 7811–7818.
- Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Li, S.; Le, H.; Xu, J.; and Salzmänn, M. 2025a. Enhancing Compositional Text-to-Image Generation with Reliable Random Seeds. In *The Thirteenth International Conference on Learning Representations*.
- Li, W.; Kumar, S.; Feng, W.; He, X.; and Chandra, V. 2025b. Align and Bind: Enhancing Structural Composition in Text-to-Image Models. In *The Thirteenth International Conference on Learning Representations (ICLR)*.
- Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023a. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 292–305.
- Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023b. GLIGEN: Open-Set Grounded Text-to-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22511–22521.
- Lian, L.; Li, B.; Yala, A.; and Darrell, T. 2023. LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models. *Transactions on Machine Learning Research*.
- Liu, F.; Emerson, G.; and Collier, N. 2023. Visual Spatial Reasoning. *Transactions of the Association for Computational Linguistics*, 11: 635–651.
- Liu, R.; Renzo, R.; Wu, C.; Shen, W. T.; Reid, C.; Gao, C.; and Vondrick, C. 2023. Zero-1-to-3: Zero-shot One Image to 3D Object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9298–9309.
- Liu, Y.; Liu, M.; Huang, S.; and Lv, J. 2025. Asymmetric Visual Semantic Embedding Framework for Efficient Vision-Language Alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5676–5684.
- Ma, Z.; Hong, J.; Gul, M. O.; Gandhi, M.; Gao, I.; and Krishna, R. 2023. CREPE: Can Vision-Language Foundation Models Reason Compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10910–10921.
- Mao, J.; Gan, C.; Kohli, P.; Tenenbaum, J. B.; and Wu, J. 2019. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *International Conference on Learning Representations*.
- Mittal, S.; Thakral, K.; Singh, R.; Vatsa, M.; Glaser, T.; Canton Ferrer, C.; and Hassner, T. 2024. On responsible machine learning datasets emphasizing fairness, privacy and regulatory norms with examples in biometrics and healthcare. *Nature Machine Intelligence*, 6: 936–949.
- Nebel, B.; and Bürckert, H.-J. 1995. Reasoning about Temporal Relations: A Maximal Tractable Subclass of Allen’s Interval Algebra. *Journal of the ACM*, 42(1): 43–66.
- Paiss, R.; Ephrat, A.; Tewel, O.; Mosseri, I.; Irani, M.; and Dekel, T. 2023. Teaching CLIP to Count to Ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3170–3180.
- Park, J.; Lee, J.; Song, J.; Yu, S.; Jung, D.; and Yoon, S. 2025. Know “No” Better: A Data-Driven Approach for Enhancing Negation Awareness in CLIP. *arXiv preprint arXiv:2501.10913*.
- Patel, M.; Gokhale, T.; Baral, C.; and Yang, Y. 2024. Triplet-CLIP: Improving Compositional Reasoning of CLIP via Synthetic Vision-Language Negatives. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2348–2357.
- Phung, Q.; Ge, S.; and Huang, J.-B. 2024. Grounded Text-to-Image Synthesis with Attention Refocusing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7932–7942.
- Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv preprint arXiv:2307.01952*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 8748–8763.
- Rassin, R.; Hirsch, E.; Glickman, D.; Ravfogel, S.; Goldberg, Y.; and Chechik, G. 2023. Linguistic Binding in Diffusion Models: Enhancing Attribute Correspondence through Attention Map Alignment. In *Advances in Neural Information Processing Systems*, volume 36, 64928–64940.
- Renz, J.; and Nebel, B. 1999. On the Complexity of Qualitative Spatial Reasoning: A Maximal Tractable Fragment of the Region Connection Calculus. *Artificial Intelligence*, 108(1–2): 69–123.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent

- Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.
- Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems*, volume 35, 33979–34001.
- Singh, J.; Shrivastava, I.; Vatsa, M.; Singh, R.; and Bharati, A. 2025. Learning the Power of “No”: Foundation Models with Negations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 7991–8001.
- Srinivasan, T.; Ren, X.; and Thomason, J. 2023. Curriculum Learning for Data-Efficient Vision-Language Alignment. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 5619–5624.
- Thakral, K.; Glaser, T.; Hassner, T.; Vatsa, M.; and Singh, R. 2025. Fine-Grained Erasure in Text-to-Image Diffusion-based Foundation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9121–9130.
- Thrush, T.; Jiang, R.; Bartolo, M.; Singh, A.; Williams, A.; Kiela, D.; and Ross, C. 2022. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5238–5248.
- Truong, T. H.; Baldwin, T.; Verspoor, K.; and Cohn, T. 2023. Language Models Are Not Naysayers: An Analysis of Language Models on Negation Benchmarks. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM)*, 101–114.
- Vempati, S.; Tucker, A.; de Lassence, G.; Goyal, A.; Bengio, Y.; Hsieh, T.-I.; and Hjelm, C. D. 2023. VALSE: A Task-Independent Benchmark for Vision and Language Models through Visual Sentence Composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 15609–15620.
- Xue, Z.; Song, G.; Guo, Q.; Liu, B.; Zong, Z.; Liu, Y.; and Luo, P. 2023. RAPHAEL: Text-to-Image Generation via Large Mixture of Diffusion Experts. In *Advances in Neural Information Processing Systems*, volume 36.
- Yang, J.; Zhang, H.; Li, F.; Zou, X.; Li, C.; and Gao, J. 2023a. Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V. arXiv:2310.11441.
- Yang, M.; Song, F.; Alonso-Martinez, D.; Wu, T.; Lu, L.; Zhang, S.; Biswas, S.; and Jin, H. 2024. RPG: Mastering Text-to-Image Diffusion: Recaptioning, Planning, and Generating with Multimodal LLMs. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.
- Yang, Z.; Wang, J.; Gan, Z.; Li, L.; Lin, K.; Wu, C.; Duan, N.; Liu, Z.; Liu, C.; Zeng, M.; and Wang, L. 2023b. ReCo: Region-Controlled Text-to-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14246–14255.
- Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; Hutchinson, B.; Han, W.; Parekh, Z.; Li, X.; Zhang, H.; Baldrige, J.; and Wu, Y. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *arXiv preprint arXiv:2206.10789*.
- Yuksekgonul, M.; Bianchi, F.; Kalluri, P.; Jurafsky, D.; and Zou, J. 2023a. And the Winner is... Towards Compositional Evaluation of Vision-Language Models. In *Advances in Neural Information Processing Systems*, volume 36.
- Yuksekgonul, M.; Bianchi, F.; Kalluri, P.; Jurafsky, D.; and Zou, J. 2023b. When and Why Vision-Language Models Behave like Bags-of-Words, and What to Do About It. In *The Eleventh International Conference on Learning Representations (ICLR)*.
- Zafar, O.; Cohen, Y.; Wolf, L.; and Schwartz, I. 2024. Detection-Driven Object Count Optimization for Text-to-Image Diffusion Models. *arXiv preprint arXiv:2408.11721*.
- Zhang, G.; Fu, B.; Fan, Q.; Zhang, Q.; Liu, R.; Gu, H.; Zhang, H.; and Liu, X. 2024. CoMPaSS: Enhancing Spatial Understanding in Text-to-Image Diffusion Models. arXiv:2412.13195.
- Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3836–3847.
- Zheng, G.; Zhou, X.; Li, X.; Qi, Z.; Shan, Y.; and Li, X. 2023. LayoutDiffusion: Controllable Diffusion Model for Layout-to-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22490–22499.