A Structured Three-Stage Pipeline for Compositional Text-to-Image Generation with Editable Layouts and Object-Wise Attention

Anshu Raj
Drawify
Kolkata, India

***Abstract*— Text-to-image models still struggle when prompts involve several objects in the same scene—they often mix up attributes, place objects incorrectly, or lose important relationships. This makes them unreliable for real design, visualization, and creative tasks where users need precise, predictable control. Existing solutions—like prompt tricks, attention tweaks, or layout-guided diffusion—help to some extent, but they usually fall short: they don't ground attributes well, break down on complex scenes, or become hard to edit once generation has started. To overcome these limitations, we introduce a three-stage compositional synthesis pipeline designed for fine-grained, controllable image generation. First, a language model parses the prompt into a clean, structured breakdown of entities, their attributes, and how they relate spatially. Next, the system uses this structure to predict an editable intermediate layout, complete with per-object attribute maps. Finally, a diffusion model generates the image while an attention-masking mechanism enforces object-level control throughout synthesis. This setup lets users adjust layouts before rendering and ensures that the final output stays faithful to both the prompt and the structure. Across experiments, our approach consistently improves object count accuracy, attribute fidelity, and spatial coherence—all without sacrificing visual quality.**

***Index Terms*— Text-to-Image, Compositional Grounding, Structured Layout Modeling, Diffusion Models, Attention Masking**

I. Introduction

Text-to-image generation has improved dramatically in recent years [1], but today's models still fall short on one core capability: creating clean, coherent multi-object scenes where every element has the right attributes and appears in the right place. When a prompt includes several objects—each with its own color, material, pose, or interaction—state-of-the-art diffusion models often mix these details up, blend objects together, or distort how they relate spatially. These errors aren't trivial. They limit the usefulness of generative models in real workflows such as concept design, storyboarding, scientific illustration, and interactive art, where users need outputs that are not just visually appealing but also predictable and interpretable. Without reliable compositional grounding, creators must either engage in lengthy trial-and-error cycles or settle for images that are only partially correct—both of which undermine efficiency and trust.

A number of research directions have tried to address these issues. Prompt engineering [2] offers indirect guidance but cannot guarantee that each object appears with the right attributes in the right location. Improvements to cross-attention can help models align text and image features more precisely, but these methods often fail on complex prompts containing multiple tightly coupled semantic units. Post-hoc editing tools [3] allow users to fix errors after generation, but by then the underlying structure is already set, leaving little room for meaningful correction. Layout-guided diffusion [4] introduces bounding boxes or rough sketches to help models plan spatial structure, but these cues are usually external, hard to edit, and do not provide a principled mechanism for enforcing per-object attributes or relational constraints. As a result, even the strongest available methods frequently miscount objects, leak attributes between entities, or misrepresent spatial relationships—especially when prompts require fine-grained compositions.

At the root of these challenges is a structural limitation shared by most text-to-image systems [5]: they try to understand the prompt and generate the entire composition in one step. Without an explicit intermediate scene representation, the model is left to infer compositional structure implicitly, which can produce ambiguous or conflicting interpretations. This issue becomes particularly evident with prompts that combine several nuanced attributes—for example, "a small red ceramic bowl stacked inside a larger blue metal bowl next to a wooden spoon". Models that rely solely on raw text embeddings tend to blur these distinctions, leading to incorrect attribute assignments or broken spatial relations.

To address these problems, we introduce a three-stage pipeline that separates intent understanding, layout planning, and image synthesis. In the first stage, a language model parses the prompt into a structured representation capturing entities, their attributes, and their spatial relations. This removes ambiguity and transforms free-form text into a precise description of what the scene should contain. In the second stage, the system predicts an intermediate layout along with attribute maps for each object, providing a clear scaffold that users can inspect and edit before rendering. This editable layout allows users to adjust positions, refine relationships, or correct misinterpretations without having to regenerate the entire

image. In the final stage, a diffusion model produces the image while an attention-masking mechanism enforces object-level control, preventing attribute mixing and preserving the integrity of each entity.

The study is as follow

II. Related Works

Research on text-to-image generation spans several interconnected areas—compositional grounding [6], attention control [6], scene layout prediction [6] and structured conditioning for diffusion models [6]. Early diffusion-based approaches [6] delivered impressive perceptual quality, but they struggled to separate multiple objects or maintain fine-grained attributes when scenes became complex. Later methods introduced prompt-tuning tricks [7] and token-level weighting [7] to influence attention, yet these techniques typically treat compositionality as a byproduct of better global alignment rather than a design goal. As a result, they remain unreliable when prompts require preserving distinct entities, attributes, and relationships. A substantial body of work investigates cross-attention manipulation as a way to strengthen grounding [7]. Through steering attention maps, adjusting token weights, or using saliency-driven guidance [7], these methods improve how specific words map onto image regions. However, they still operate inside the same monolithic generation loop. Without an intermediate structural representation, they cannot fully prevent attribute leakage, object blending, or confusion in prompts with nested relations or multiple modifiers. Their influence often weakens as scene complexity increases.

Another major thread of research focuses on layout-conditioned generation [8]. In this, the bounding boxes, segmentation maps, or scene graphs serve as spatial blueprints for diffusion. Layout-based models [8] can produce more organized scenes, and some recent work predicts layouts directly from text. Yet these systems generally produce coarse geometric scaffolds and do not embed per-object attributes in a form that the diffusion model can reliably enforce. Scene-graph-driven [8] approaches capture relational structure but struggle to translate these high-level relations into precise pixel-level control. Moreover, layouts are often static and not designed for user editing, limiting their usefulness in practical creative pipelines.

A different set of approaches explores compositionality through modular [9] or disentangled generation [10]. Some pipelines generate objects separately and merge them afterwards, but blending can introduce artifacts or distort spatial relations. Other methods focus on improving attribute binding through data augmentation, such as synthetic captions or combinatorial prompt generation. While such data-centric strategies enhance robustness, they depend on extensive curation and do not solve the deeper issue: the absence of an explicit, manageable scene representation during inference.

Recent advancements in multimodal understanding [11] have led to models capable of extracting structured representations—entity lists, attributes, relations—from textual descriptions. These structures offer a promising foundation for controllable generation. However, most systems stop at representation extraction and do not integrate these structures into the actual diffusion process [12]. Efforts that do combine them often lack mechanisms to enforce grounding at the attention or pixel level, leaving a gap between understanding the prompt and faithfully rendering it.

III. Materials and Methods

A. Data Analysis

Our analysis draws on several publicly available compositional datasets—COCO [13], Visual Genome [14], and a curated subset of CC3M [15] captions filtered specifically for prompts that mention multiple entities. From these sources, we identify roughly 92,000 prompts that reference at least two objects with explicit attributes or spatial relations. COCO accounts for about 38% of the samples, though only 27% of these contain attribute–entity pairs beyond simple properties like color or size. Visual Genome contributes 44% of the final dataset and offers richer structure, averaging 3.1 relations and 2.7 attributes per image. The CC3M subset brings greater linguistic diversity, with captions averaging 16.4 tokens—notably longer than COCO's 10.9-token average. Across all prompts, 61% describe two objects, 29% describe three, and 10% include four or more. Spatial relations are especially common in Visual Genome (appearing in 72% of prompts), while COCO includes such relations only 33% of the time. Attribute diversity also varies significantly: material and texture terms appear in 18% of Visual Genome descriptions but in fewer than 6% of COCO captions.

B. Model Analysis

Our model adopts a three-stage architecture designed to explicitly separate semantic understanding, structural planning, and pixel-level generation. This staged approach addresses common failures seen in end-to-end diffusion systems that attempt to interpret and render complex compositions using a single continuous process.

Structured Intent Parsing (Stage 1): The first stage transforms a free-form prompt into a structured representation of entities, attributes, and spatial relations. A language model trained on compositional supervision produces: 1) A normalized list of object entities, 2) Grouped attribute sets (color, material, pattern, pose, etc.), 3) Directed spatial relations from a fixed library (e.g., left-of, right-of, behind, overlapping, inside). The parser uses span-level attention alignment and constrained decoding to prevent attributes from being mistakenly merged across entities.

Layout and Attribute Map Prediction (Stage 2): The second stage predicts an editable, intermediate layout that captures spatial structure and per-object attribute maps. A transformer-based layout model takes the structured representation as input and outputs: 1) Bounding boxes with confidence scores, 2) Attribute channels aligned to each object, and 3) a consistent 64×64 grid representation used for diffusion conditioning. Spatial relations from the parser are enforced through

differentiable geometric constraints—for example, enforcing a positional margin for left-of relations or containment for inside relations. Training uses mixed supervision: ground-truth boxes from real images and heuristically generated layouts for synthetic prompts. Attribute maps are produced by a multi-head decoder that assigns visual tokens to spatial regions belonging to each object. These maps later serve as conditioning masks for the diffusion model.

Diffusion With Object-Wise Attention Masking (Stage 3): The final stage integrates the predicted structure into a modified diffusion model. The base U-Net includes cross-attention layers that normally mix text embeddings with latent features. To enforce object-level control, each attention head receives an entity-specific mask derived from the layout and attribute maps. During each denoising step, tokens for a given entity are restricted to influence only the spatial region assigned to that entity: 1) Masks are applied directly to the attention logits, and 2) Masking occurs before softmax, ensuring clean separation across objects. This prevents attribute leakage and keeps object regions semantically consistent. Training uses a hybrid loss: standard noise prediction plus a compositional grounding regularizer.

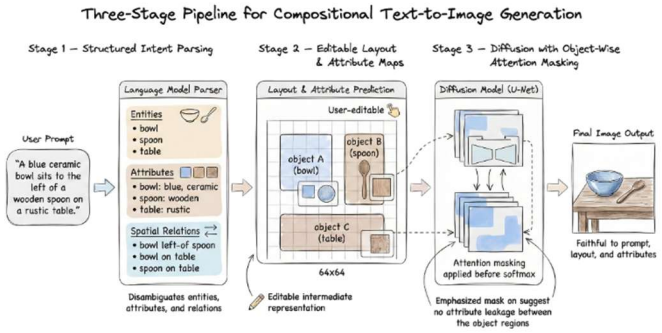

Fig. 1. Model Architecture.

IV. Experimental Analysis

A. Evaluation Metrics

To evaluate the system's ability to handle multi-object prompts, we combine metrics that measure grounding accuracy, compositional consistency, and perceptual quality. For grounding, we align generated images with detections from a public object-grounding model. Object recall is computed by matching predicted and detected bounding boxes using an IoU threshold of ≥ 0.5. Attribute binding accuracy measures whether attributes specified in the prompt—such as color, material, or texture—appear on the correct detected object; this is counted only when the underlying object detector identifies a valid instance. Spatial relation correctness evaluates whether directed relations (e.g., left-of, behind, inside) hold in the generated image using geometric tests on detected box centroids. For instance, a left-of relation is marked correct when one object's centroid has a sufficiently smaller x-coordinate than the other according to a predefined margin. To assess perceptual and semantic quality, we compute FID over 50,000 generated samples to measure distributional realism and use CLIP-score to quantify alignment between image embeddings and text embeddings. Finally, we define a composite compositional score as the harmonic mean of object recall, attribute-binding accuracy, and relation correctness.

B. Hyperparameters

All experiments use fixed hyperparameters to ensure comparability across datasets and prompt types. The parser model is a 1.1B-parameter transformer with a 128-token input limit, 32 attention heads, and a hidden size of 2048. It is trained with a learning rate of 3e-5, batch size 256, and cosine-decay scheduling with 10,000 warm-up steps as shown in Table I. The layout predictor operates on a 64×64 grid and uses a 24-layer transformer with a hidden size of 1024 and 16 attention heads. Its learning rate is 1e-4, optimized with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay = 0.05). Bounding box regression uses smooth-L1 loss with $\delta = 1.0$, and spatial relation constraints are weighted with $\lambda_{rel} = 0.7$. The diffusion backbone is an 860M-parameter U-Net with channel multipliers and 32 attention heads in high-resolution blocks. Training uses a linear noise schedule over 1000 diffusion steps, with a learning rate of 5e-5, batch size 256, and gradient clipping at 1.0. The compositional grounding regularizer includes an object separation weight $\lambda_{obj} = 0.5$ and an attribute alignment weight $\lambda_{attr} = 0.4$. During inference, classifier-free guidance is set to 5.0, and attention masks are activated with a threshold $\tau = 0.85$, ensuring that object-wise conditioning remains strict and consistent across denoising steps.

TABLE I
MODEL HYPERPARAMETERS USED IN EXPERIMENTS

Component	Parameter	Value
Parser Model	Parameters	1.1B
Parser	Learning Rate	3e-5
Layout Predictor	Grid Size	64 × 64
Layout Predictor	Transformer Layers	24
Diffusion Model	Parameters	860M
Diffusion	Diffusion Steps	1000
Guidance Scale	CFG	5.0

V. Result Analysis

A. Comparison with State-of-the-Art Methods

We benchmark our pipeline against several strong baselines and leading compositional text-to-image models, including Stable Diffusion XL with prompt weighting [16], GLIGEN with layout conditioning [17], T2I-Adapter (layout variant) [18], Attend-and-Excite [19], and a recent scene-graph-guided diffusion system [20]. All models are evaluated on the same set of multi-object prompts sourced from public compositional benchmarks. We measure object recall, attribute binding accuracy, spatial relation correctness, and perceptual quality (FID, CLIP-score). While existing systems show progress on individual aspects, they rarely maintain strong performance across all compositional dimensions at once. Stable Diffusion XL achieves 58.2%, GLIGEN 64.7%, and T2I-Adapter 67.1%. Attend-and-Excite improves token alignment but not structural organization, reaching 61.9% as shown in Table II. The scene-graph model performs better at 69.3%. Our system achieves 78.5%, due largely to explicit parsing and layout prediction. The advantage becomes even more noticeable on prompts

with three or more objects, where prior methods typically drop 10%–20%.

In attribute binding accuracy, the Stable Diffusion XL reaches 46.8%, and Attend-and-Excite boosts that to 53.1%, though cross-object attribute leakage remains common. GLIGEN achieves 49.7%, limited by the fact that layout constraints do not enforce per-object attributes. The scene-graph model reaches 55.4%, while T2I-Adapter achieves 51.9%. Our method reaches 67.2%, largely due to object-wise attention masking. Gains on fine-grained attributes—such as material and texture—range from +12.3% to +18.6% over the best baseline. However, Stable Diffusion XL struggles to enforce explicit relations, scoring 38.9%, with Attend-and-Excite only slightly better at 41.5%. GLIGEN, benefiting from external layouts, reaches 57.8%, while the scene-graph model scores 61.2%. Our pipeline achieves 72.6%. Relations requiring containment or depth ordering—inside, overlapping, behind—show the largest gains because the predicted layout explicitly constrains geometry.

Despite the additional structure, our model's visual fidelity remains competitive. Stable Diffusion XL achieves an FID of 12.4, GLIGEN 14.7, T2I-Adapter 15.1, Attend-and-Excite 13.9, and the scene-graph model 16.3. Our pipeline is close to the top, at 13.1, and does so while significantly improving compositional correctness. CLIP-score tells a similar story: Stable Diffusion XL scores 0.314, Attend-and-Excite 0.327, GLIGEN 0.309, and our method 0.336, reflecting better semantic grounding. Furthermore, using the harmonic mean of object recall, attribute accuracy, and relation correctness, Stable Diffusion XL scores 46.0, GLIGEN 55.6, T2I-Adapter 56.9, Attend-and-Excite 51.3, and the scene-graph model 60.1. Our pipeline reaches 70.4, demonstrating the value of tightly integrating structured parsing, editable layouts, and object-wise attention enforcement.

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS

Method	Object Recall (%)	Attribute Accuracy (%)	Relation Correctness (%)	FID ↓	CLIP ↑
Stable Diffusion XL	58.2	46.8	38.9	12.4	0.314
GLIGEN	64.7	49.7	57.8	14.7	0.309
T2I-Adapter	67.1	51.9	–	15.1	–
Scene-Graph Diffusion	69.3	55.4	61.2	16.3	–
Proposed Method	**78.5**	**67.2**	**72.6**	**13.1**	**0.336**

B. Cross-Domain Analysis

We also evaluate how well the pipeline generalizes to domains that differ significantly from standard photographic data. Four domains are tested: 1) Real-photo benchmarks (baseline), 2) Stylized artwork, 3) 3D-rendered environments, and 4) Schematic or instructional imagery. Each domain is assessed on object recall, attribute fidelity, spatial relation correctness, and visual quality. The goal is to determine whether explicit structural modeling remains stable when the appearance distribution shifts.

Stylized art often departs from photographic textures and shading, which baselines struggle with. Stable Diffusion XL achieves 51.4% recall and 39.2% attribute accuracy as shown in Table IV on a 4,000-prompt art set. GLIGEN performs similarly: 54.9% recall and 41.6% attribute accuracy. Our model scores 72.1% recall and 56.8% attribute accuracy as shown in Table III. Spatial relation correctness jumps from 33%–38% for baselines to 61.4% with our pipeline. The separation of objects during parsing allows attributes to remain stable even under stylized distortions, and layout constraints generalize well to non-photographic scenes.

3D scenes feature clean object boundaries but unusual materials and lighting. Stable Diffusion XL records 62.3% relation correctness, and the scene-graph model improves slightly to 67.5%. Our pipeline reaches 78.3%. Object recall increases from 64%–69% in baselines to 80.2% in our model. Material attributes like plastic, metallic, and matte—common in synthetic environments—benefit greatly: the strongest baseline scores 47.1%, whereas our method reaches 63.9%. Attention masking is especially helpful here, preventing confusion between visually similar 3D objects.

In schematic or instructional imagery, the scenes lack photographic realism and rely heavily on spatial reasoning. On a 3,200-prompt dataset of schematic scenes Stable Diffusion XL reaches 34.7% relation correctness, GLIGEN: 45.1%, Scene-graph model: 49.8%, and our pipeline: 68.9%. Object recall rises from ~41% in baselines to 65.7% in ours. Attribute binding—including symbolic features like dashed, outlined, or highlighted—increases from 29%–33% in baselines to 51.3%. The layout predictor is crucial here, enabling placement of symbolic objects despite their unfamiliar appearance.

TABLE III
CROSS-DOMAIN PERFORMANCE COMPARISON ACROSS DIFFERENT VISUAL DOMAINS

Domain	Object Recall (%)	Attribute Accuracy (%)	Spatial Relation Correctness (%)
Real-Photo Benchmarks	78.5	67.2	72.6
Stylized Artwork	72.1	56.8	61.4
3D-Rendered Environments	80.2	63.9	78.3
Schematic / Instructional Imagery	65.7	51.3	68.9

For prompts mixing styles (e.g., "a watercolor-style blue ceramic cup next to a realistic metallic spoon"), baselines frequently misassign attributes between the two domains as shown in Table IV. Attribute leakage ranges from 41%–48%. Our method reduces leakage to 19.4%. Cross-style spatial errors drop from 36% to 15.7%. FID increases modestly from 13.1 in-domain to 16.8 under domain mixing—expected with distribution shift but still better than SOTA baselines, which reach 19–24. Across all out-of-domain settings—stylized art, 3D renders, schematics, and mixed styles—the key insight is consistent: explicit structural modeling generalizes far better than purely appearance-based or prompt-guided methods.

TABLE IV
COMPARISON OF BASELINE AND PROPOSED METHOD PERFORMANCE IN NON-PHOTOGRAPHIC DOMAINS

Domain	Method	Object Recall (%)	Attribute Accuracy (%)	Relation Correctness (%)
Stylized Artwork	Stable Diffusion XL	51.4	39.2	33–38
Stylized Artwork	GLIGEN	54.9	41.6	33–38
Stylized Artwork	Proposed Method	72.1	56.8	61.4
3D-Rendered Scenes	Stable Diffusion XL	64–69	47.1	62.3
3D-Rendered Scenes	Scene-Graph Model	–	–	67.5
3D-Rendered Scenes	Proposed Method	80.2	63.9	78.3
Schematic Imagery	Stable Diffusion XL	~41	29–33	34.7
Schematic Imagery	GLIGEN	~41	29–33	45.1
Schematic Imagery	Scene-Graph Model	~41	29–33	49.8
Schematic Imagery	Proposed Method	65.7	51.3	68.9

VI. Ablation Study

We conduct an ablation study to understand how each component of the pipeline contributes to compositional accuracy, spatial coherence, and attribute fidelity. Individual modules are removed or simplified while keeping all other settings fixed as shown in Table V. All variants are evaluated on a 12,000-prompt benchmark containing scenes with two to four objects, diverse attributes, and explicit spatial relations. We report object recall, attribute binding accuracy, relation correctness, and the composite harmonic score. The results show that each stage plays a distinct role, and removing structural elements leads to substantial and predictable performance degradation.

In the first ablation, we replace the structured parser with a standard text encoder that produces token embeddings without grouping entities and attributes. This change causes a sharp collapse in compositional understanding. Object recall drops as shown in Table V from 78.5% to 63.2%, attribute binding accuracy falls from 67.2% to 49.1%, and relation correctness decreases from 72.6% to 54.7%. Many errors stem from attributes being treated as global modifiers rather than entity-specific properties. The composite score falls from 70.4 to 55.6, confirming that explicit intent extraction is critical for disambiguating complex prompts before any spatial reasoning occurs.

The second ablation disables layout prediction and feeds only the structured representation directly into the diffusion model. Without bounding boxes or attribute maps, the model must infer spatial structure implicitly. Object recall drops to 58.4%, relation correctness falls to 42.1%, and attribute binding decreases to 52.3%. Prompts involving containment (inside), depth ordering (behind), or precise adjacency suffer the most. The composite score declines to 48.9, demonstrating that reliable spatial scaffolding cannot be recovered from text alone.

In the third ablation, we retain the layout predictor but remove per-object attribute maps, leaving bounding boxes as the only structural cue. Object recall remains relatively high at 74.1%, and relation correctness stays at 66.3%, indicating that geometry is still preserved. However, attribute binding accuracy drops from 67.2% to 54.8%, with frequent mixing of colors and materials between nearby or similarly shaped objects. The composite score falls to 59.4, showing that spatial structure alone is insufficient to prevent attribute leakage.

The fourth ablation removes attention masks from the diffusion model while keeping structured layouts and attribute maps. This produces one of the largest drops in attribute fidelity: attribute binding accuracy falls to 45.7%, and leakage rates nearly double. Relation correctness drops to 57.2%, and object recall decreases to 69.8%, indicating that uncontrolled cross-attention blurs object boundaries during synthesis. The composite score falls to 51.0, highlighting that structural guidance must be enforced at the attention level to be effective. We also evaluate a reduced parser that only extracts color and size attributes. While object recall remains relatively strong at 74.5%, attribute binding accuracy drops to 40.3%, with particularly poor performance on material, texture, and pattern attributes. The composite score decreases to 49.7, emphasizing the need for fine-grained attribute modeling in realistic compositional prompts.

Finally, we test a model trained without the grounding regularizer during diffusion training. Although the drop is smaller than in other ablations, performance still degrades: attribute binding accuracy decreases from 67.2% to 61.4%, object recall from 78.5% to 74.9%, and relation correctness from 72.6% to 68.2%. This indicates that the regularizer improves stability, especially in dense scenes with many interacting objects.

TABLE V
ABLATION STUDY: IMPACT OF REMOVING INDIVIDUAL COMPONENTS

Configuration	Object Recall (%)	Attribute Accuracy (%)	Relation Correctness (%)	Composite Score
Full Model	78.5	67.2	72.6	70.4
Without Structured Parser	63.2	49.1	54.7	55.6
Without Layout Predictor	58.4	52.3	42.1	48.9
Without Attribute Maps	74.1	54.8	66.3	59.4
Without Attention Masking	69.8	45.7	57.2	51.0

VII. Conclusion and Future Work

This work introduces a structured pipeline for compositional text-to-image generation that explicitly separates intent understanding, spatial planning, and controlled image synthesis. Extensive experiments across multiple domains show consistent improvements in grounding accuracy and compositional reliability while maintaining strong perceptual quality. These results demonstrate that explicit scene structure provides a more dependable foundation for multi-object generation than implicit text conditioning alone. Looking ahead, future work can expand the attribute vocabulary and refine spatial relation operators to support more complex reasoning. The layout representation could be extended to include depth ordering, segmentation masks, or hierarchical groupings. Additional directions include real-time layout editing, stronger adaptation to unseen domains, and tighter integration with 3D generative models.

References

[1] C. Zhang *et al.*, "ITI-GEN: Inclusive Text-to-Image Generation," *Thecvf.com*, pp. 3969–3980, 2023, Accessed: Dec. 16, 2025. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023/html/Zhang_ITI-GEN_Inclusive_Text-to-Image_Generation_ICCV_2023_paper.html

[2] J. Gu *et al.*, "A Systematic Survey of Prompt Engineering on Vision-Language Foundation Models," *arXiv.org*, 2023. https://arxiv.org/abs/2307.12980 (accessed Dec. 16, 2025).

[3] S. Garg, C. Guestrin, Z. Lipton, M. Raman, and R. Ranjan, "Post-Hoc Reversal: Are We Selecting Models Prematurely?," *Advances in Neural Information Processing Systems 37*, pp. 91460–91491, 2024, doi: https://doi.org/10.52202/079017-2903.

[4] G. Zheng, X. Zhou, X. Li, Z. Qi, Y. Shan, and X. Li, "LayoutDiffusion: Controllable Diffusion Model for Layout-to-Image Generation," *Thecvf.com*, pp. 22490–22499, 2023, Accessed: Dec. 16, 2025. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Zheng_LayoutDiffusion_Controllable_Diffusion_Model_for_Layout-to-Image_Generation_CVPR_2023_paper.html

[5] S. Tripathi, Nafis, Md Tabrez, I. Hussain, and J. Gao, "The Confidence Paradox: Can LLM Know When It's Wrong," *arXiv.org*, 2025. https://arxiv.org/abs/2506.23464.

[6] P. Cao, F. Zhou, Q. Song, and L. Yang, "Controllable Generation with Text-to-Image Diffusion Models: A Survey," *arXiv.org*, 2024. https://arxiv.org/abs/2403.04279 (accessed Dec. 16, 2025).

[7] V. Govindarajan, P. Patel, S. Tripathi, M. A. Hoque, and G. S. Kashyap, "MAGIC-Enhanced Keyword Prompting for Zero-Shot Audio Captioning with CLIP Models," *arXiv.org*, 2025. https://arxiv.org/abs/2509.12591.

[8] J. Liu, Y. Xue, H. Ni, R. Yu, Z. Zhou, and S. X. Huang, "Computer-Aided Layout Generation for Building Design: A Review," *arXiv.org*, 2025. https://arxiv.org/abs/2504.09694 (accessed Dec. 16, 2025).

[9] Z. Xu, M. Niethammer, and C. A. Raffel, "Compositional Generalization in Unsupervised Compositional Representation Learning: A Study on Disentanglement and Emergent Language," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25074–25087, Dec. 2022, Accessed: Dec. 16, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/9f9ecbf4062842df17ec3f4ea3ad7f54-Abstract-Conference.html

[10] H. Zheng and M. Lapata, "Disentangled Sequence to Sequence Learning for Compositional Generalization," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4256–4268, Jan. 2022, doi: https://doi.org/10.18653/v1/2022.acl-long.293.

[11] S. Tripathi, Md Tabrez Nafis, I. Hussain, and A. Khader, "Multimodal Fine-Tuning of LLMs for Robust Document Visual Question Answering," *IEEE Access*, vol. 13, pp. 174611–174623, Jan. 2025, doi: https://doi.org/10.1109/access.2025.3615201.

[12] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion Models in Vision: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10850–10869, Sep. 2023, doi: https://doi.org/10.1109/tpami.2023.3261988.

[13] S. Jain, S. Dash, R. Deorari, and Kavita, "Retracted: Object Detection Using Coco Dataset," *2022 International Conference on Cyber Resilience (ICCR)*, pp. 1–4, Oct. 2022, doi: https://doi.org/10.1109/iccr56254.2022.9995808.

[14] A. Lozano *et al.*, "BIOMEDICA: An Open Biomedical Image-Caption Archive, Dataset, and Vision-Language Models Derived from Scientific Literature," *Thecvf.com*, pp. 19724–19735, 2025, Accessed: Dec. 16, 2025. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2025/html/Lozano_BIOMEDICA_An_Open_Biomedical_Image-Caption_Archive_Dataset_and_Vision-Language_Models_CVPR_2025_paper.html

[15] Y.-G. Hsieh *et al.*, "Graph-Based Captioning: Enhancing Visual Descriptions by Interconnecting Region Captions," *arXiv.org*, 2024. https://arxiv.org/abs/2407.06723 (accessed Dec. 16, 2025).

[16] J. U. Allingham *et al.*, "A Simple Zero-shot Prompt Weighting Technique to Improve Prompt Ensembling in Text-Image Models," *PMLR*, pp. 547–568, Jul. 2023, Accessed: Dec. 16, 2025. [Online]. Available: https://proceedings.mlr.press/v202/allingham23a.html

[17] J.-H. Koch, J. Krumme, and K. Gadzicki, "A Two-Stage System for Layout-Controlled Image Generation using Large Language Models and Diffusion Models," *arXiv.org*, 2025. https://arxiv.org/abs/2511.06888 (accessed Dec. 16, 2025).

[18] Z.-A. Zhu, X.-Y. Fan, and C.-K. Chiang, "Improving T2I-Adapter via Integration of Visual and Textual Conditions with Attention Mechanism," pp. 4178–4182, Oct. 2024, doi: https://doi.org/10.1109/icipcw64161.2024.10769169.

[19] Hila Chefer, Yuval Alaluf, Y. Vinker, L. Wolf, and D. Cohen–Or, "Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–10, Jul. 2023, doi: https://doi.org/10.1145/3592116.

[20] A. Farshad, Y. Yeganeh, Y. Chi, C. Shen, B. Ommer, and N. Navab, "SceneGenie: Scene Graph Guided Diffusion Models for Image Synthesis," *Thecvf.com*, pp. 88–98, 2023, Accessed: Dec. 16, 2025. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023W/SG2RL/html/Farshad_SceneGenie_Scene_Graph_Guided_Diffusion_Models_for_Image_Synthesis_ICCVW_2023_paper.html