Multigranular Alignment via Linguistic Decomposition and Reward Optimization for Text to Image Diffusion

Hengrui Liu
The University of Auckland
New Zealand
hliu825@aucklanduni.ac.nz

Luming Jin
The University of Auckland
New Zealand
ljin892@aucklanduni.ac.nz

Meng-Fen Chiang*
National Yang Ming Chiao Tung University
Taiwan
meng.chiang@nycu.edu.tw

ABSTRACT

While text-to-image (T2I) diffusion models have achieved remarkable fidelity, semantic alignment remains a critical bottleneck for complex, multi-entity prompts. Existing models frequently exhibit object omission, attribute misbinding, and spatial drift. To address these challenges, we propose **GranAligner**, a two-stage framework designed to enforce multi-granular correspondence through semantic decomposition, compositional synthesis, and reward-driven realignment. GranAligner comprises two key stages: (i) Structural Decomposition and Compositional Synthesis and (ii) Multi-Granular Semantic Realignment. In the first stage, complex prompts are factorized into semantically coherent sub-concepts to anchor structured image generation via cross-attention control. The second stage curates these outputs through reward-based optimization at both the global-scene and local-object granularities. This architecture establishes a bidirectional reinforcement cycle, where realignment signals are propagated back to the decomposition logic to iteratively narrow the text-visual semantic gap. Designed for architectural extensibility, GranAligner leverages noun-phrase factorization compatible with Diffusion Transformer (DiT) mechanisms and employs Low-Rank Adaptation (LoRA) for parameter-efficient tuning. Beyond specific backbones, our multi-level evaluation serves as an architecture-agnostic curation pipeline, enabling high-fidelity data selection and providing generalizable optimization insights for next-generation models like Stable Diffusion 3. Extensive evaluations on the MS-COCO, ABC-6K, and CC-500 benchmarks demonstrate that GranAligner significantly outperforms existing baselines in compositional fidelity and semantic adherence.

KEYWORDS

Compositional text to image generation; Multi granularity semantic alignment; Reward-guided diffusion training; Prompt decomposition and synthesis

ACM Reference Format:
Hengrui Liu, Luming Jin, and Meng-Fen Chiang. 2026. Multigranular Alignment via Linguistic Decomposition and Reward Optimization for Text to Image Diffusion. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/UEMK8185

*Corresponding author.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/UEMK8185

1 INTRODUCTION

Text-to-image (T2I) synthesis aims to generate visually realistic and semantically faithful imagery from natural language, enabling transformative applications in automated content generation and creative design [16, 28, 30]. While recent large-scale architectures, such as DALL·E 2 [20], Imagen [22], and Parti [34], have achieved remarkable fidelity, they often require immense computational overhead for fine-grained adaptation [29]. More critically, state-of-the-art models frequently falter on complex, multi-facet prompts: entities are often omitted, attributes leak across subjects, and spatial relations drift, eroding the semantic integrity of generated scenes.

While diffusion models such as Stable Diffusion [21] have advanced T2I through improved metrics and broad adoption, precise semantic alignment remains elusive. As illustrated in Figure 1, state-of-the-art models frequently swap attributes or distort spatial constraints: a prompt such as "*a brown bench in front of a white building*" may yield a white bench before a brown building, revealing a collapse in entity-attribute binding. Recent attempts to inject structural priors, such as training-free prompt splitting or structure-guided attention [3], narrow this gap but do not close it. Coarse strategies, such as splitting on conjunctions, fail on nuanced syntax, while global alignment signals (e.g., caption similarity) often prioritize scene-level coherence at the expense of local object-level fidelity [6, 31, 35].

We identify two fundamental gaps in current T2I alignment research that impede scalable knowledge discovery in generative modeling. First, a pervasive **hierarchical reasoning deficit** exists: standard architectures treat prompts as linear token sequences, failing to capture the recursive, tree-structured dependencies (e.g., nested attributes within entity scopes) inherent in natural language. Global embedding similarities effectively average these structural nuances, inducing relational drift. Second, we formalize the **initial quality bottleneck**. Frameworks utilizing coarse-to-fine schemes, such as Realign Diffusion [6], are *posterior-limited* by the semantic fidelity of Stage-1 denoising. If initial synthesis exhibits high semantic entropy, which results in entity omission or irrevocable attribute leakage, the subsequent optimization manifold becomes ill-posed. Broad optimization signals cannot perform semantic recovery on ungrounded pixel regions; without Stage-1 grounding, fine-tuning lacks the localized gradient signal required to induce entity emergence. Furthermore, existing mechanisms exacerbate this via **scalar reward sparsity**. Preference-based models like ImageReward [31] typically provide a singular global scalar lacking *spatial-semantic attribution*. Such rewards fail to disambiguate which specific sub-structure caused the alignment failure, diluting the learning signal during parameter-efficient adaptation. Finally,

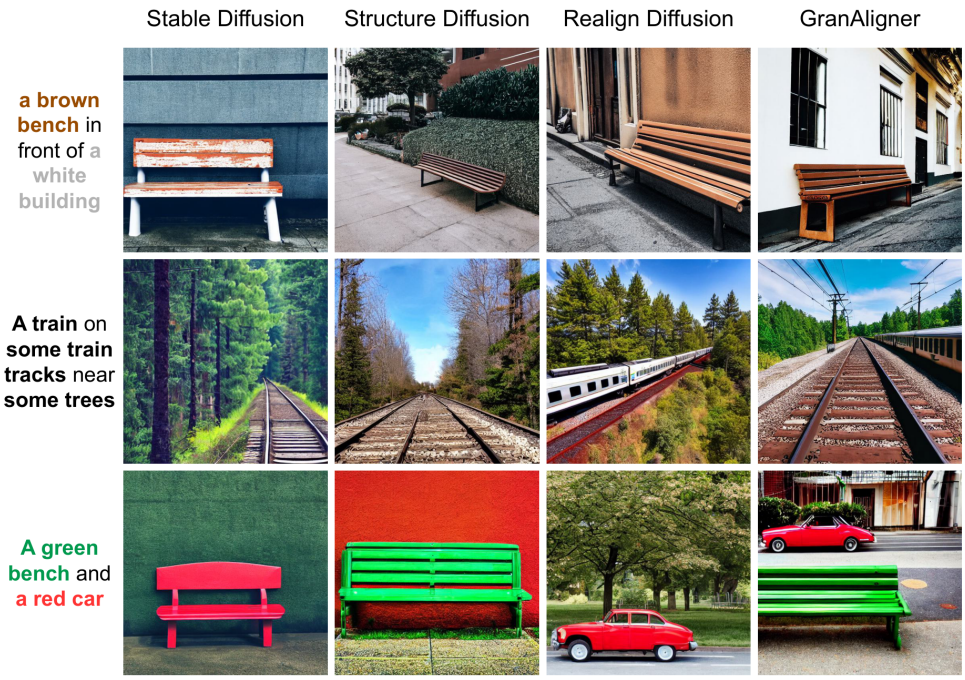

Figure 1: Visual comparisons with Stable Diffusion [20], Structure Diffusion [3], and Realign Diffusion [6] reveal common issues, such as attribute binding errors (e.g., mixed white and brown on the bench in the first image of the first row, and a red bench in the first image of the third row) and missing objects (e.g., the missing "train" in the first two images of the second row). In contrast, our method resolves these inconsistencies, improving text–image alignment. Bound parts are highlighted with colors for clarity (e.g., a brown bench).

while recent advances in test-time attention control [8] and multi-agent pipelines [13] mitigate mismatches during inference, they represent **non-persistent interventions**. They lack a systematic feedback loop to refine underlying model weights, requiring redundant per-sample computation rather than fundamentally regularizing the model's latent manifold.

To bridge this fine-to-coarse semantic discrepancy, we propose **GranAligner**, a two-stage framework that couples linguistic decomposition with multi-granular reward optimization. Our approach prioritizes the systematic discovery and curation of high-fidelity text-image pairs through two integrated stages: (i) **Structural Decomposition and Compositional Synthesis:** We employ a language parser to factorize complex prompts into constituent noun-phrase units. These units serve as structural anchors for discrete cross-attention channels, ensuring that entity-attribute bindings are preserved from the onset of the denoising process. (ii) **Multi-Granular Semantic Realignment:** We utilize a joint global-local reward mechanism to curate Stage-I outputs. A *global scene reward* evaluates contextual coherence and spatial layout, while a *local object reward* leverages RAM/LLM-assisted verification to isolate specific entities and assess fine-grained attribute fidelity. By filtering Stage-I samples through this dual-scale reward, we curate a high-precision dataset for parameter-efficient adaptation via Low-Rank Adaptation (LoRA). This self-supervised loop effectively synchronizes the model's latent manifold with the complex structural requirements of the prompts, progressively tightening alignment without exhaustive backbone retraining.

Our main contributions are summarized as follows:

- We introduce **GranAligner**, a two-stage, coarse-to-fine framework that improves T2I alignment by combining linguistic decomposition with multi-granularity reward realignment.
- We introduce a **compositional synthesis strategy** that factorizes prompts into noun-phrase anchors, mapping them to discrete cross-attention keys and values to preserve attribute-object integrity during initial synthesis.
- We develop a **dual-reward curation mechanism** that integrates global scene-level scoring with local entity-attribute verification via RAM/LLM-assisted analysis. This strategy solves the problem of scalar reward sparsity, identifying high-fidelity training pairs to regularize the optimization manifold during parameter-efficient adaptation.
- Comprehensive benchmarks on MS-COCO and ABC-6K show GranAligner outperforms state-of-the-art baselines, improving CLIP by up to +1.79% and reducing FID by 5.46%. On CC-500, it increases GLIP-based object count accuracy by 0.5%, confirming superior entity preservation.

2 RELATED WORK

Image Generation with Spatial Control. Text-to-image (T2I) models often exhibit spatial misalignment when precise control over object position, size, and layout is required. A growing line of work integrates spatial conditioning into diffusion frameworks [9, 18, 29, 35, 36]. ControlNet augments pre-trained diffusion models with cues such as edge maps and depth to enable fine control [35]. Dense Diffusion adjusts attention maps to improve layout without retraining [9]. BoxDiff introduces efficient box-level constraints for precise object placement [29], while FreeControl aligns spatial and appearance features via PCA for structured guidance [18]. Despite strong results, these methods depend on external spatial inputs or auxiliary predictors, which limits scalability and flexibility. We instead infer spatial structure from text alone by encoding multi-phrase prompt semantics and injecting them into the diffusion process, removing the need for external spatial supervision while retaining fine-grained control.

Image Generation with Scene Graph. Scene graphs provide structured, interpretable representations of objects and their relationships, offering compositional guidance for image generation. Recent approaches integrate them into diffusion models to improve text-image alignment. SGH [27] expands initial scene graphs from prompts to support controllable scene hallucination. SG-Adapter [23] injects relational priors into textual embeddings, enhancing semantic fidelity. SceneGenie [2] applies spatial constraints during sampling via bounding boxes and segmentation maps guided by CLIP features. Nonetheless, scene graphs often fail to capture abstract concepts and stylistic nuances in natural language.

Composable Image Generation. Text-only T2I models frequently suffer layout misalignment and weak attribute binding, causing incorrect object placement and relationships. Compositional methods address this by incorporating structured cues to boost spatial and semantic coherence [3, 15]. Structure Diffusion [3] segments prompts into concepts linked to encoded keys, guiding image regions without extra training, enabling plug-and-play improvements in multi-object and multi-attribute settings. Composable Diffusion [15] employs multiple specialized diffusion models, using logical operators (e.g., AND, NOT) for independent noise control across components, improving multitask generation. Despite these advances, challenges remain, such as object omissions, highlighting the need for more faithful representation of all entities beyond prompt segmentation alone.

Image Generation with Prompt Optimization. Integrating large language models (LLMs) into T2I pipelines has improved generation quality by optimizing prompt representations [1, 7, 17, 24]. CoOp [38] introduces learnable continuous prompts for vision-language tasks, while Prompt Engineering Diffusion [33] and in-context learning [10] further refine prompts to better handle complex scenes with multiple objects and spatial relations. Nonetheless, semantic ambiguity and weak attribute binding persist in complex descriptions. This work tackles these by extracting multi-granular features from captions and leveraging selected prompt-image pairs to isolate and strengthen object-attribute associations. Few-shot prompt optimization using LLMs (e.g., GPT-J) enables data-driven refinement without manual engineering [10, 32]. Fine-tuning prompts on a few examples improve the alignment between text and generated images through iterative gradient updates. Despite improved visual fidelity, object omission remains a key challenge. Furthermore, prior remedies, such as training-free composition and global caption-similarity fine-tuning, remain coarse, in which under-reason over full linguistic structure and often fail to preserve object–attribute bindings during synthesis. We address these gaps with a two-component design that injects structure before generation and maintains it throughout decoding.

3 METHODOLOGY

We propose GranAligner, a decomposable diffusion framework that couples semantic decomposition with compositional synthesis to narrow the text–image semantic gap.

3.1 Framework Overview

The GranAligner framework consists of two stages. (i) Decomposition and compositional synthesis: The prompt is grouped into structured concepts, objects, attributes and relations, and used to steer generation so that the object-attribute bindings are preserved in the initial image. In contrast to prompt-level embeddings and post-hoc attention reweighting, our decomposition-plus-composition pipeline reasons over the complete text structure and preserves bindings during synthesis, directly addressing the limitations that hinder fine-grained, scene- and object-level alignment. (ii) Semantic Realignment: residual discrepancies are corrected by joint alignment at the global (scene-level) and local (object-level) granularities. Iterating these stages progressively reduces omissions and attribute swaps, yielding high-quality images with strong semantic fidelity to the input prompt. Figure 2 illustrates the framework.

3.2 Decomposition and Compositional Synthesis

To optimize attribute binding in text-to-image generation through two key components: semantic decomposition and compositional image generation. First, the prompt instruction is decomposed into a structured concept graph, which represents the individual concept of the text. This concept graph is then utilized to synthesize the overall semantic meaning, guiding the generation of the image.

3.2.1 Semantic Decomposition. For complex prompts with multiple attributed objects and relationships, the CLIP text encoder's ability to process these intricacies is limited, which constrains the T2I generation quality. Semantic Decomposition addresses this by decomposing and independently encoding complex texts. To this end, we first extract the linguistic structure of the prompt using a constituency tree (or scene graph) [3], to represent syntactic components hierarchically. The root corresponds to the full prompt, branches denote phrases (e.g., noun phrases, verbs, relations), and leaf nodes represent individual noun phrases (NPs) that capture attribute-object pairs. Rather than encoding the entire prompt as a single vector, we encode each noun phrase (NP) together with its attributes to avoid attribute–object confusion, and feed these concept-specific embeddings to the cross-attention layers for precise control. These embeddings serve as inputs to the cross-attention layers for more precise generation. Formally, from the constituent tree, we extract noun phrases $\Gamma = \{\mathrm{NP}_1, \mathrm{NP}_2, \ldots, \mathrm{NP}_k\}$, each aligned with

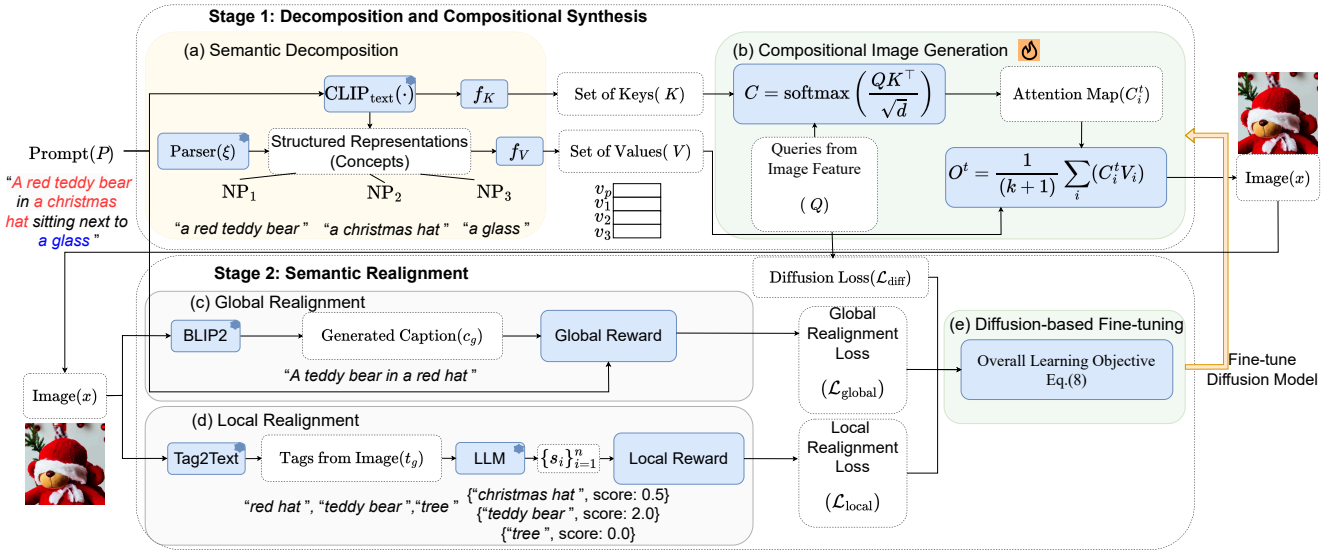

Figure 2: Overall Architecture of GranAligner. (a) *Semantic Decomposition* takes the prompt P as input to extract a set of keys and values to represent its semantic components. (b) *Compositional Image Generation* uses these representations to generate the image x. (c) *Global Realignment* takes the image x and the prompt P as input to compute the global realignment loss $\mathcal{L}_{\text{global}}$ by evaluating scene-level consistency. (d) *Local Realignment* takes the image x and the prompt P as input to compute the local realignment loss $\mathcal{L}_{\text{local}}$ by evaluating object-level consistency. (e) *Diffusion-based Fine-tuning* improves text-image alignment by optimizing the diffusion model with reward-based learning.

corresponding image regions in prompt order. Unlike traditional methods that process the entire prompt with the CLIP encoder, we independently encode each NP to capture its semantic structure more effectively as follows:

$$\Lambda = [\Lambda_w, \Lambda_1, \Lambda_2, \ldots, \Lambda_k], \quad \Lambda_i = \text{CLIP}_{\text{text}}(\text{NP}_i) \tag{1}$$

where Λ_w is the encoding of the entire prompt, and k is the number of NPs ($i = 1, \ldots, k$). The embedding sequence Λ_i is aligned with Λ_w, preserving the semantic order of the original prompt structure. Note that NP extraction is well-established for prompt decomposition. We use a tree-based structure extractor to isolate and encode noun phrases, although alternatives like scene graphs or LLM parsing are equally reliable and rarely error-prone. The semantic decomposition component manages this process, generating the keys ($\mathbb{K}$) and values ($\mathbb{V}$) required for subsequent stages:

$$\begin{aligned} \mathbb{K} &= \{f_K(\Lambda_i)\} = \{K_p, K_1, \ldots, K_k\} \\ \mathbb{V} &= \{f_V(\Lambda_i)\} = \{V_p, V_1, \ldots, V_k\} \end{aligned} \tag{2}$$

where f_K and f_V are functions that generate keys and values from NPs, with K_p and V_p representing the key and value derived from the entire prompt. These representations guide the generation of the attention map for $i = p, 1, 2, \ldots, k$.

3.2.2 Compositional Image Generation. We use Stable Diffusion as the backbone model to generate the initial image from the decomposed prompt. We preserve the model architecture while focusing on the cross-attention layer, which integrates semantic information from the input text. The set of keys and values derived from the prompt is applied, with the embedding sequence mapped to regions on the cross-attention map C_t, which guides the image generation process. Each noun phrase (NP) generates a unique attention map, and these maps are combined to independently control distinct image regions. The attention maps are computed as the product of queries and keys:

$$\mathbb{C}^t = \text{softmax}\left(\frac{Q^t K_i}{\sqrt{d}}\right), \quad i = p, 1, 2, \ldots, k. \tag{3}$$

where Q^t is the output of the previous layer containing image features, and K_i is the key of the i-th noun phrase. Each attention map is combined with the corresponding values to form the output of the cross-attention layer:

$$O^t = \frac{1}{(k+1)}\sum_i (C_i^t V_i), \quad i = p, 1, 2, \ldots, k. \tag{4}$$

where O^t is the output of the cross-attention layer, which is passed to the next layer. After processing through subsequent layers, the final image x is generated, with $\mathbb{C}$ representing the set of attention maps.

Example (Compositional Image Generation)**.** *The cross attention outputs are decoded into an initial image x, which then serves as input to the next stage. For the prompt "A red teddy bear in a Christmas hat sitting next to a glass," the result may still misalign with the text, e.g., omitting the glass as depicted in Figure 2. This discrepancy motivates diffusion-based fine-tuning to realign semantics between the prompt and the generated image.*

3.3 Semantic Realignment

While the semantic decomposition improves attribute binding, generated images may still suffer from missing objects or incomplete scenes. The Semantic Realignment stage refines alignment to enhance consistency, from fine-grained object details to overall scene structure. This stage consists of three components: global realignment for spatial and contextual coherence, local realignment for object positioning and attribute accuracy, and diffusion-based fine-tuning for iterative refinement. Unlike Realign Diffusion, GranAligner enhances semantic fidelity by selecting higher-quality training sets and utilizing a reward feedback framework with scene- and object-level losses to fine-tune Stable Diffusion, thus improving attribute binding and text-image alignment.

3.3.1 Global Realignment. Similarly to Realign Diffusion, the global realignment component improves spatial and contextual coherence by evaluating the alignment between the generated image x and the input prompt. To avoid losing critical image content in direct comparisons within a shared embedding space, we first generate a caption using the pre-trained BLIP-2 model [11] to extract image features and generate semantically relevant captions. Given an image x, BLIP-2 generates a caption c_g, which is encoded to obtain the text embedding v_g. Similarly, the input prompt is encoded as v_p. The global alignment is quantified by the cosine similarity between these embeddings, defining the global reward as: $\mathbf{R}_{\text{global}} = cosine(\mathbf{v}_g, \mathbf{v}_p)$. This reward measures scene-level consistency by evaluating the spatial and contextual coherence between the input prompt and the generated image, guiding diffusion-based fine-tuning for improved text-image alignment. The global reward optimization objective is formulated as follows:

$$\mathcal{L}_{\text{global}} = \text{ReLU}(\mathbf{R}_{\text{global}}(P, g_\theta(P)) \tag{5}$$

where P is the input prompt, $g_\theta(P)$ is the generated image, and the ReLU ensures non-negative local reward values.

3.3.2 Local Realignment. The local realignment component improves object-level consistency by refining the alignment between the generated image x and the input prompt. Local Realignment addresses the challenge of missing or misrepresented objects, ensuring that each object and its attributes are accurately positioned. Formally, we define a local reward that assesses the presence and accuracy of the object relative to the prompt by verifying the correct representation of each described object. We use the Recognize Anything Model (RAM) [37], a pre-trained image tagging model, to extract object tags from the generated image. These tags, along with CLIP encoded prompt embeddings, are inputted into a Large Language Model (LLM)[1] to compute likelihood scores s_i for each object o_i ($1 \le i \le n$) to reflect its presence in context as follows:

$$s_i = \begin{cases} 2, & \text{if } o_i \text{ is certain to appear in the scene.} \\ 0.5, & \text{if } o_i \text{ may appear in the scene.} \\ 0, & \text{if } o_i \text{ is unlikely to appear in the scene.} \end{cases} \tag{6}$$

The local reward is computed by normalizing the sum of these scores: $\mathbf{R}_{\text{local}} = \frac{1}{2 \cdot n} \sum_{i=1}^{n} (s_i - 2 \cdot n)$, where n is the total number of identified objects and s_i is the score for the i-th object. This reward quantifies the alignment between the image objects and the text description, with higher values indicating more substantial alignment. To optimize object-level alignment, the local reward loss is defined as:

$$\mathcal{L}_{\text{local}} = \text{ReLU}(\mathbf{R}_{\text{local}}(P, g_\theta(P)) \tag{7}$$

where P is the input prompt, $g_\theta(P)$ is the image generated by the diffusion model in the former stage, and the ReLU ensures non-negative local reward values. To **identify high-quality samples** from synthetic data, we use both coarse and fine semantic metrics. The coarse metric R_{global} (Eq. 5) evaluates the overall semantic consistency through $\mathcal{L}_{\text{global}}$. The fine metric, R_{local} (Eq. 7), measures object-level alignment as $\mathcal{L}_{\text{local}}$. Both scores rank images, and the top-K image-prompt pairs form the fine-tuning dataset, which consists entirely of synthetic samples.

Example (Local Reward Estimation)**.** *Given the prompt "a white sheep and a red car," the RAM identifies objects such as "sheep" and "car" An LLM computes likelihood scores based on these tags and the prompt. If both objects are present, the LLM assigns a score of 2 to both "sheep" and "cars", reflecting their high relevance. For less likely objects, such as "road," the LLM assigns a score of 0.5, while irrelevant tags such as "bike" and "dog" receive a score of 0. Given the relevant tags "sheep" and "car" (each assigned a score of 2), the local reward is −1, indicating that the objects in the image generated in stage one are highly consistent with the prompt. This object-level consistency is therefore maintained through the local realignment loss.*

3.3.3 Diffusion-based Fine-tuning. To enhance alignment efficiently, we employ diffusion-based parameter-efficient LoRA fine-tuning [] that integrates global and local realignment losses. A feedback learning framework optimizes the diffusion models by selecting high-quality samples using global and local reward estimators. This curated subset fine-tunes the pre-trained model, improving alignment at both scene and object levels. Specifically, image-text pairs are ranked based on their scores, with top-ranked pairs chosen to preserve alignment at both levels. Using this dataset, Reward Feedback Learning (ReFL) [31] optimizes text-image consistency. ReFL adapts the pre-trained model using a low-rank matrix, updating only a few parameters to avoid full retraining. The model is optimized by evaluating generated samples through a reward function and minimizing reward loss, which integrates global and local realignment terms to refine text-image consistency and object-level alignment. The overall loss function is defined as follows:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{global}} + \lambda_2 \mathcal{L}_{\text{local}} + \mathcal{L}_{\text{diff}} \tag{8}$$

where λ_1 and λ_2 are coefficients that balance the loss components for stable fine tuning. The diffusion loss is

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{x, \epsilon \sim \mathcal{N}(0,I), t} \left\| \epsilon - \varepsilon_\theta(x_t, t) \right\|^2.$$

Larger λ_1 and λ_2 prioritize reward-driven alignment for better text–image consistency, while smaller values emphasize diffusion stability.

4 EXPERIMENTS

4.1 Experimental Setup

4.1.1 Datasets. We utilize three commonly used benchmark datasets: MS-COCO [14], CC-500 [3], ABC-6K [3].

[1] Compatible with models-like LLaMA [26] or GPT-4 [25].

Dataset	Method	FID↓	CLIP↑
MS-COCO	Stable Diffusion	14.8545	0.3176
	Structure Diffusion	14.6278	0.3284
	Realign Diffusion	<u>13.8941</u>	<u>0.3341</u>
	GranAligner (Ours)	**13.1937** (↑5.0%)	**0.3394** (↑1.6%)
ABC-6K	Stable Diffusion	14.6836	0.3157
	Structure Diffusion	14.4564	0.3276
	Realign Diffusion	<u>13.9385</u>	<u>0.3352</u>
	GranAligner (Ours)	**13.1768** (↑5.5%)	**0.3412** (↑1.8%)

Table 1: Quantitative results on MS-COCO and ABC-6K. Best results in bold; second-best <u>underlined</u>. Blue denotes relative improvement over the strongest baseline.

Methods	Zero/One ↓	Two ↑
*Prompt: "a **[colorA] [objectA]** and a **[colorB] [objectB]**"*		
Stable Diffusion	69.0	31.0
Composable Diffusion	74.2	25.8
Structure Diffusion	68.8	31.2
Realign Diffusion	<u>65.1</u>	<u>34.9</u>
GranAligner (Ours)	**64.6**	**35.4** (+0.5%)

Table 2: Entity grounding performance on CC-500 via GLIP. Metrics Zero/One and Two quantify object omission failures (↓) and successful compositional generation (↑), respectively, for dual-entity prompts with distinct attributes.

Stage One	Stage Two	FID↓	CLIP↑
		14.8545 (↓ 12.58%)	0.3176 (↓ 6.42%)
✓		13.9103 (↓ 5.43%)	0.3346 (↓ 1.14%)
	✓	14.4641 (↓ 9.63%)	0.3214 (↓ 5.30%)
✓	✓	**13.1937**	**0.3394**

Table 3: Ablation study on MS-COCO. Stage One (Decomposition and Compositional Synthesis) and Stage Two (Semantic Alignment) improve text-to-image generation in GranAligner. Performance degradation relative to the full model (bold) is marked by ↓.

4.1.2 Baselines. We benchmark against two categories of strong T2I diffusion baselines: training-free methods, such as Stable Diffusion [21] and Structure Diffusion [3], and fine-tuned methods—Realign Diffusion [6] and Composable Diffusion [15].

4.1.3 Evaluation Metrics. We measure semantic consistency and image quality using two metrics on the the MS-COCO and ABC-6K datasets: **CLIP** [4, 19] and Fréchet Inception Distance (**FID**) [5]. For CLIP, we compute the cosine similarity between the feature vectors of the generated image X and its corresponding prompt P, encoded by the CLIP model. A higher CLIP score indicates better semantic alignment between the image and prompt. For FID, we compare feature distributions between generated and real images by computing activations from the Inception V3 model's penultimate layer. The Fréchet distance between the distributions of generated and real images quantifies image quality, with a lower FID score indicating better image quality and closer alignment with real images. Additionally, we utilize the phrase-based model **GLIP** [12] on the CC-500 dataset to evaluate whether the generated image contains the correct number of objects as specified by the prompt.

4.1.4 Implementation Details. We conduct all experiments on NVIDIA A100 GPUs (80 GB) using Stable Diffusion v1.5[2] as the base generative model, fine-tuned for our tasks at a fixed resolution of 512×512 [21]. To ensure fair comparison with baselines such as Structure Diffusion and Realign Diffusion, we standardize on v1.5; we are also extending evaluations to newer backbones (v2.0+). Complete hyperparameters and training schedules are reported in Appendix **??**. The detailed hyperparameter setup is provided in the Appendix **??**. The pipeline is compatible with diffusion transformers (DiT) with minor adaptations. In the Decomposition and Compositional Synthesis stage, prompts are parsed into noun-phrase units that steer cross-attention, a mechanism directly supported by DiT architectures. In the Semantic Realignment stage, we fine-tune a small subset of Stable Diffusion parameters using LoRA, preserving the pre-trained weights while enhancing semantic alignment.

[2]https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5

4.2 Overall Semantic Consistency

4.2.1 Setup. We compare our model against two categories of T2I diffusion models: *training-free* and *fine-tuned* methods, with the latter requiring training on a new dataset. For fairness, all fine-tuned models are trained on the same 3,000 randomly selected MS-COCO samples. To minimize randomness, we repeat the process three times and report the average results.

4.2.2 Results. Table 1 shows the quantitative comparison results of different methods on the MS-COCO and ABC-6K datasets. As shown in Table 1, our method outperforms other state-of-the-art (SOTA) methods on all evaluation metrics. Our method attains the highest scores (**bold**), with the second-best (Realign Diffusion) <u>underlined</u> for reference. Numbers in denote our relative improvement over the second-best baseline. Specifically, our method achieved FID scores of 13.1937 and 13.1768 on the MS-COCO and ABC-6k datasets, respectively, which are lower than those of competing methods, indicating the effectiveness of the generated image quality. Moreover, our method achieved CLIP scores of 0.3394 and 0.3412 on the MS-COCO and ABC-6k datasets, respectively. These scores indicate that the images generated by our model align more closely with the textual descriptions. Table 1 compares our method with three baselines on MS-COCO and ABC-6k.

4.3 Complex Semantic Comprehension

4.3.1 Setup. The CC-500 dataset features compositional prompts using conjunctions like "and," which often cause diffusion models

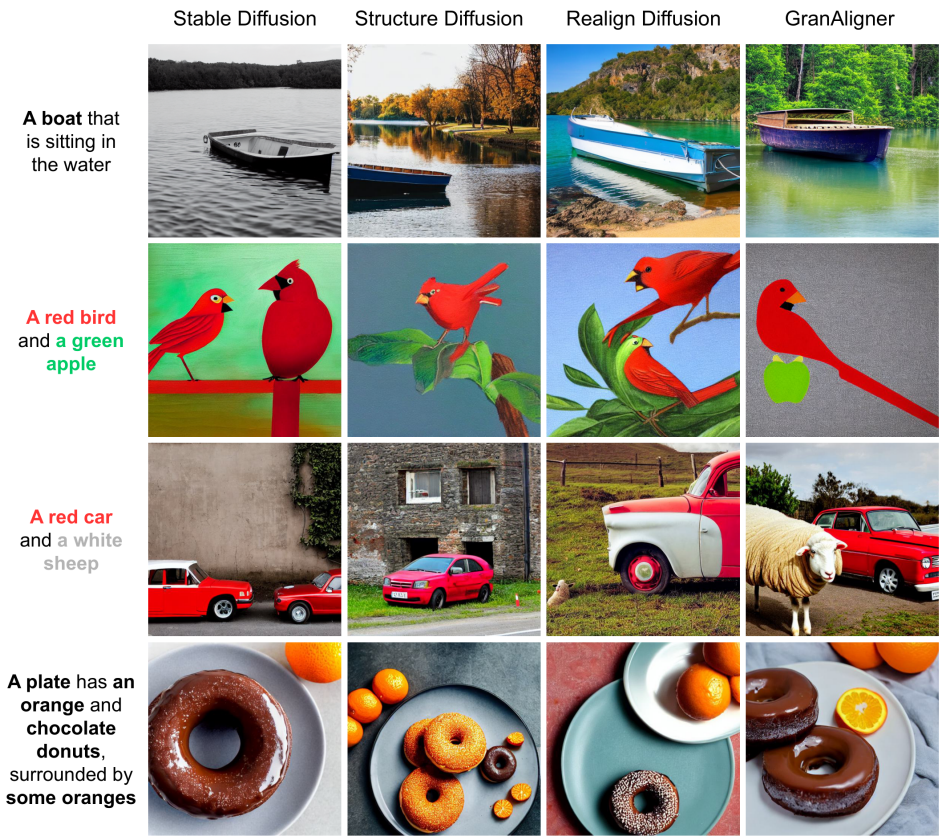

Figure 3: Visual Comparison of Existing Baselines. GranAligner demonstrates superior object count accuracy, attribute binding, and image quality across prompts of varying complexity.

to omit objects due to masking. We evaluate our method against baselines tailored for compositional prompts. Stable Diffusion and Structure Diffusion use default settings, while Composable Diffusion splits prompts at “and" and processes segments independently. Object count accuracy and missing objects are assessed using the GLIP detection model [12] via bounding box analysis.

4.3.2 Results. Table 2 reports object count accuracy on CC-500. Among the baselines, Composable Diffusion correctly counted objects in 25.8% of the prompts, while Structure Diffusion achieved a modest improvement to 31.2%, due to its approach of separately encoding objects, which enhances spatial layout and composition. Realign Diffusion further outperformed both, achieving 34.9%, by focusing on realigning text and image during generation for improved prompt alignment. GranAligner outperformed all baselines with a success rate of 35.4%, demonstrating that refining initial image generation produces a higher-quality training set and better text-image alignment.

4.4 Ablation Studies

We evaluate the effectiveness of each stage in GranAligner. Table 3 presents the ablation study results on the MS-COCO dataset, demonstrating that our two-stage approach significantly improves T2I generation performance. For image quality, the first-stage generation reduces the FID score from 14.8545 to 13.9103, while the second-stage re-alignment further lowers it to 14.4641. For text-image consistency, the first stage improves the CLIP score from 0.3176 to 0.3346, whereas the second stage adjusts it to 0.3214. The first-stage generation yields a more substantial improvement, aligning with expectations, as its output directly influences the effectiveness of the re-alignment process.

4.5 Qualitative Results

We qualitatively assess GranAligner against competitive T2I baselines across the MS-COCO and CC-500 benchmarks.

4.5.1 T2I with Typical Prompts. Figure 3 compares image generation for the prompt “A boat that is sitting in the water." While Stable Diffusion, Structure Diffusion, and Realign Diffusion generate aligned images, GranAligner produces higher-quality results. Specifically, boats in Stable and Structure Diffusion show shadowed areas at the waterline, and Realign Diffusion lacks realistic boat details. In contrast, GranAligner generates a more visually appealing boat with accurate water reflections and improved image quality.

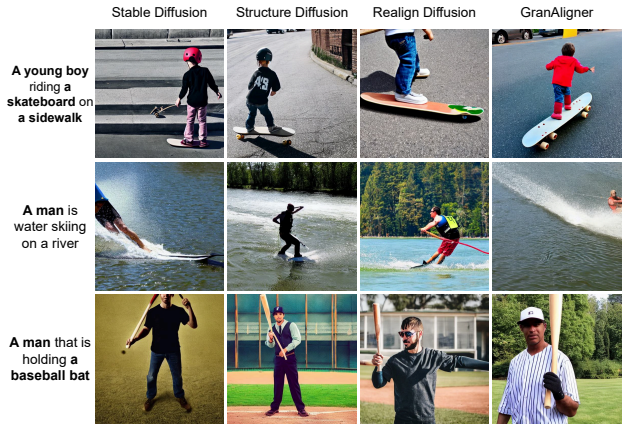

Figure 4: Baseline comparison on human-centric prompts. GranAligner yields anatomically realistic figures and natural poses, outperforming baselines in structural consistency.

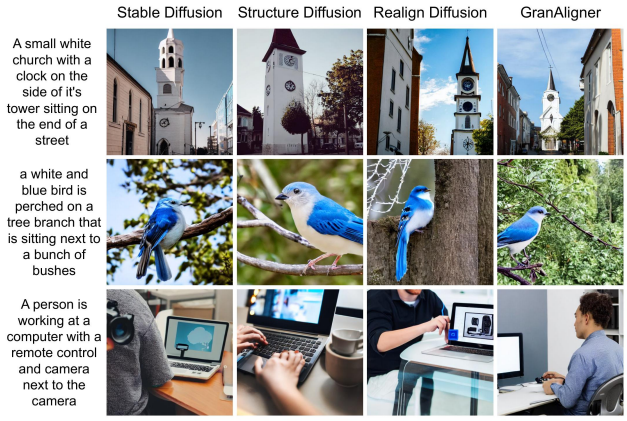

Figure 5: Failure case analysis. Highly complex prompts reveal common limitations across all evaluated models, including fidelity degradation and the systematic omission of fine-grained entities such as clocks, complex textures, and human anatomy.

4.5.2 T2I with Complex Prompts. Figure 3 illustrates compositional prompts like "A red bird and a green apple." While all methods generate simplistic images, Stable Diffusion, Structure Diffusion, and Realign Diffusion fail to render all objects and exhibit color leakage, assigning both "red" and "green" attributes to the "bird." In contrast, GranAligner correctly renders all objects with precise attribute binding. Similarly, for "A red car and a white sheep." baselines misrepresent object counts and proportions, while our method achieves accurate object presence and realistic sizing. For more complex prompts such as "A plate with an orange and chocolate donuts, surrounded by some oranges," involving multiple objects and specific quantities, other methods struggle to align objects. GranAligner generates correct object counts and captures relationships, outperforming baselines.

4.5.3 T2I with Human Objects. Figure 4 compares GranAligner with baselines on prompts involving human subjects (i.e., "human objects"). Baseline models frequently exhibit anatomical artifacts, distorted fingers, asymmetric limbs, and inconsistent face-body coherence, indicative of weak object-attribute binding at fine granularity. GranAligner mitigates many of these failures, producing more faithful layouts and attributes; nonetheless, high-quality human renderings remain challenging.

4.6 Error Analysis

Figure 5 shows the bad cases generated by all baseline models and our method, where there is a gap in the consistency between the image and the text. For example, in the prompt "*A small white church with a clock on the side of its tower sitting on the end of a street,*" the model, when generating based on the prompt, focuses on the overall object and environment, neglecting the hidden attributes of the "*clock*" object. The clock hands and scale were not correctly generated, which resulted in low image quality. Additionally, when processing "*a white and blue bird is perched on a tree branch that is sitting next to a bunch of bushes,*" the details of the "*a bunch of bushes*" object pose higher demands on the model. None of the methods paid attention to the object's detailed texture, making the bushes in the image appear similar to the tree leaves. The same issue occurred with "*A person is working on a computer with a remote control and a camera next to the camera*", where the model failed to correctly generate the human hand, resulting in poor image quality.

5 CONCLUSIONS

We propose **GranAligner**, a coarse-to-fine framework that mitigates attribute misbinding and object omission by optimizing self-supervised training fidelity. To bypass the signal dilution inherent in direct reward guidance, GranAligner comprises a two-stage reinforcement loop: (i) *structural decomposition and compositional synthesis*, which generate structure-aware image candidates, and (ii) *multi-granular semantic realignment*, which fuses global scene scoring with local object-attribute verification to curate high-fidelity pairs for parameter-efficient adaptation. This architecture progressively enforces semantic correspondence without requiring exhaustive backbone retraining. Empirical evaluations on MS-COCO and ABC-6K yield CLIP score improvements of +1.59% and +1.79%, alongside FID reductions of 5.04% and 5.46%, respectively. Furthermore, a 0.5% increase in GLIP-based object detection accuracy on CC500 confirms superior preservation of compositional entities.

ACKNOWLEDGMENTS

This work is partially supported by the National Science and Technology Council (NSTC), Taiwan (Grants: NSTC-114-2222-E-A49-004 and NSTC-114-2639-E-A49-001-ASP).

REFERENCES

[1] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. 2022. Plot: Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253* (2022).
[2] Azade Farshad, Yousef Yeganeh, Yu Chi, Chengzhi Shen, Böjrn Ommer, and Nassir Navab. 2023. Scenegenie: Scene graph guided diffusion models for image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 88–98.

[3] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2022. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032* (2022).

[4] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718* (2021).

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

[6] Zutao Jiang, Guian Fang, Jianhua Han, Guansong Lu, Hang Xu, Shengcai Liao, Xiaojun Chang, and Xiaodan Liang. 2023. RealignDiff: Boosting Text-to-Image Diffusion Model with Coarse-to-fine Semantic Re-alignment. *arXiv preprint arXiv:2305.19599* (2023).

[7] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics* 8 (2020), 423–438.

[8] Jeeyung Kim, Erfan Esmaeili, and Qiang Qiu. 2025. Text Embedding is Not All You Need: Attention Control for Text-to-Image Semantic Alignment with Text Self-Attention Maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8031–8040. https://openaccess.thecvf.com/content/CVPR2025/html/Kim_Text_Embedding_is_Not_All_You_Need_Attention_Control_for_CVPR_2025_paper.html

[9] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. 2023. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7701–7711.

[10] Seunghun Lee, Jihoon Lee, Chan Ho Bae, Myung-Seok Choi, Ryong Lee, and Sangtae Ahn. 2024. Optimizing prompts using in-context few-shot learning for text-to-image generative models. *IEEE Access* 12 (2024), 2660–2673.

[11] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[12] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10965–10975.

[13] Mingcheng Li, Xiaolu Hou, Ziyang Liu, Dingkang Yang, Ziyun Qian, Jiawei Chen, Jinjie Wei, Yue Jiang, Qingyao Xu, and Lihua Zhang. 2025. MCCD: Multi-Agent Collaboration-based Compositional Diffusion for Complex Text-to-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13263–13272. https://openaccess.thecvf.com/content/CVPR2025/html/Li_MCCD_Multi-Agent_Collaboration-based_Compositional_Diffusion_for_Complex_Text-to-Image_Generation_CVPR_2025_paper.html

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[15] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. 2022. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*. Springer, 423–439.

[16] Yan Luo, Zhichao Zuo, Zhao Zhang, Zhongqiu Zhao, Haijun Zhang, and Richang Hong. 2024. High-Fidelity Diffusion Editor for Zero-Shot Text-Guided Video Editing. In *2024 IEEE International Conference on Data Mining (ICDM)*. 291–300. https://doi.org/10.1109/ICDM59182.2024.00036

[17] K Mallikharjuna Rao and Tanu Patel. 2024. Enhancing Control in Stable Diffusion Through Example-based Fine-Tuning and Prompt Engineering. In *2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN)*. 887–894. https://doi.org/10.1109/ICIPCN63822.2024.00153

[18] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. 2024. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7465–7475.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[20] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv e-prints* (2022), arXiv–2204.

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

[22] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* 35 (2022), 36479–36494.

[23] Guibao Shen, Luozhou Wang, Jiantao Lin, Wenhang Ge, Chaozhe Zhang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, Guangyong Chen, et al. 2024. Sg-adapter: Enhancing text-to-image generation with scene graph guidance. *arXiv preprint arXiv:2405.15321* (2024).

[24] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980* (2020).

[25] OpenA Team. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[26] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[27] Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. 2023. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. *Advances in Neural Information Processing Systems* 36 (2023), 79240–79259.

[28] Xinyu Wu. 2024. Text-Driven Image Editing Based on Diffusion Models. In *2024 IEEE 7th International Conference on Information Systems and Computer Aided Education (ICISCAE)*. 608–612. https://doi.org/10.1109/ICISCAE62304.2024.10761137

[29] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. 2023. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7452–7461.

[30] Jiayu Xu, Changhong Liu, Juan Cai, Ji Ye, Zhenchun Lei, and Aiwen Jiang. 2024. Music-driven Character Dance Video Generation based on Pre-trained Diffusion Model. In *2024 International Joint Conference on Neural Networks (IJCNN)*. 1–8. https://doi.org/10.1109/IJCNN60899.2024.10650092

[31] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2024. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems* 36 (2024).

[32] Bingsheng Yao, Guiming Chen, Ruishi Zou, Yuxuan Lu, Jiachen Li, Shao Zhang, Yisi Sang, Sijia Liu, James Hendler, and Dakuo Wang. 2024. More samples or more prompts? exploring effective few-shot in-context learning for LLMs with in-context sampling. In *Findings of the Association for Computational Linguistics: NAACL 2024*. 1772–1790.

[33] Chang Yu, Junran Peng, Xiangyu Zhu, Zhaoxiang Zhang, Qi Tian, and Zhen Lei. 2024. Seek for Incantations: Towards Accurate Text-to-Image Diffusion Synthesis through Prompt Engineering. *arXiv preprint arXiv:2401.06345* (2024).

[34] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789* 2, 3 (2022), 5.

[35] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

[36] Tianyu Zhang and Haoran Xie. 2024. Sketch-Guided Text-to-Image Generation with Spatial Control. In *2024 2nd International Conference on Computer Graphics and Image Processing (CGIP)*. 153–159. https://doi.org/10.1109/CGIP62525.2024.00035

[37] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. 2024. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1724–1732.

[38] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348.