SRA-SD: A LIGHTWEIGHT FRAMEWORK FOR STRUCTURE-GUIDED COMPOSITIONAL IMAGE SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models have demonstrated remarkable capabilities in text-to-image generation. However, they often fail to faithfully reflect the details specified in the text, missing objects or exhibiting objects with unmatched attributes and wrong spatial locations. To address this problem, we propose SRA-SD, a lightweight structure-aware framework that enhances generation fidelity by explicitly modeling both spatial relations and attribute bindings. Our method introduces two complementary modules: (1) a spatial relation enhancement module that extracts relational triples via a large language model and encodes them into heterogeneous semantic graphs, enriching the text representation with structural layout knowledge through graph neural networks; and (2) an attribute enhancement module that enforces fine-grained object-attribute alignment via contrastive cross-attention learning, using syntactically derived positive pairs and semantically plausible negative samples. To better evaluate both capabilities, we introduce SRA-Bench, a new benchmark that jointly assesses spatial reasoning and attribute binding. Experiments on three datasets show that SRA-SD significantly improves generation accuracy with minimal parameter overhead, outperforming existing methods in complex, compositional scenarios.

1 INTRODUCTION

Recent advances in diffusion models for text-to-image generation have demonstrated remarkable capabilities in producing highly realistic images (Rombach et al., 2022; Saharia et al., 2022). These models allow users to control the generated images through natural-language text prompts, which can be rich and complex. However, despite their impressive performance, a significant problem remains: the generated images often fail to faithfully reflect the details specified in the text prompts (Conwell & Ullman, 2022; Rassin et al., 2022). Specifically, stable diffusion models commonly exhibit three types of errors: (1) missing objects that are mentioned; (2) generating objects with unmatched attributes like color and size; and (3) placing objects with incorrect spatial relations. For example, in Fig.1(a), the model fails to generate the keyboard mentioned in the prompt; in Fig.1(b), it swaps the colors of the "*pink clock*" and "*brown chair*"; and in Fig.1(c), it places the bottle outside the bowl instead of "*inside the bowl*".

Some efforts have been made to address the aforementioned issues. For example, (Wu et al., 2023a) proposed a method to tackle mislocated objects by employing layout predictors to estimate object regions and applying spatial attention control. Similarly, (Rassin et al., 2023) enhanced entity-attribute alignment by parsing syntactic dependency graphs and introducing a specialized loss function for cross-attention map alignment. While these methods focus on one isolated task, which is either spatial misalignment or attribute mismatches, neglecting their inherent interdependencies in visual scene construction. In real-world scenarios, textual prompts often include intricate combinations of object attributes and spatial relations as shown in Fig 1(d). The prompt "*a purple camera on a red chair*" requires the model to not only accurately assign attributes (i.e., the colors) to the objects but also correctly position them relative to one another. This dual requirement highlights a significant gap in existing approaches.

Figure 1: Examples of text-to-image generation by Stable Diffusion 1.4. Erroneous parts are highlighted in red.

To bridge this gap, we propose SRA-SD, a lightweight **S**tructure-aware framework for enhancing **S**patial **R**elations and **A**ttributes in diffusion models. Inspired by findings that the `[EOT]` token plays a central role in controlling global semantics and layout (Wu et al., 2024), our method injects structured knowledge into the denoising process through two complementary modules. In the spatial relation enhancement module, we employ a large language model to extract relational triples from the input text and construct heterogeneous semantic graphs, which are processed by a graph neural network to refine the `[EOT]` token embedding with explicit spatial layout information, thereby reducing object omissions and improving positional accuracy. Concurrently, the attribute enhancement module leverages syntactic dependency parsing to identify object-attribute pairs as positive instances and generates fine-grained negative samples through attribute swapping and inter-object sampling. A contrastive cross-attention loss is then applied during denoising to enforce precise alignment between objects and their attributes, enhancing fine-grained semantic fidelity.

To enable a comprehensive evaluation of our proposed framework, we introduce the SRA-Bench, a new benchmark tailored to assess joint spatial and attribute reasoning in text-to-image generation. We further validate our method on GPT-synthetic (Wu et al., 2023a) and ABC-6K (Feng et al., 2023), demonstrating strong generalization in complex compositional settings. Experimental results show that SRA-SD achieves significant improvements over state-of-the-art baselines, particularly in challenging scenarios, with minimal parameter overhead, underscoring its effectiveness, efficiency, and robustness.

2 Related Work

2.1 Text-to-Image Generation

Text-to-image generation has advanced significantly with generative models like GANs (Bau et al., 2021), VAEs (Ramesh et al., 2021), and diffusion models (Gu et al., 2022). Among these, diffusion models have become state-of-the-art generating high-quality, semantically consistent images. Built on principles like heat diffusion (Perona & Malik, 1990), frameworks such as DDPM (Ho et al., 2020) and SBGM (Song et al., 2021) leverage cross-attention mechanisms (Vaswani et al., 2017) to align text with visual content, making them highly effective for text-to-image synthesis.

2.2 Compositional Synthesis with Diffusion Models

Existing text-to-image (T2I) models often struggle with compositional complexity in textual prompts. Recent advances address specific issues but lack a comprehensive solution. For object presence, Attend-and-Excite (Chefer et al., 2023) updates latent representations to ensure token incorporation, improving both object presence and attribute alignment. However, it relies on predefined token sets and does not dynamically adapt to complex spatial relations. For attribute-object alignment, Structured Diffusion (Feng et al., 2023) enhances the alignment of semantics in the cross-attention maps by extracting and separately embedding noun phrases from the prompt. (Rassin et al.,

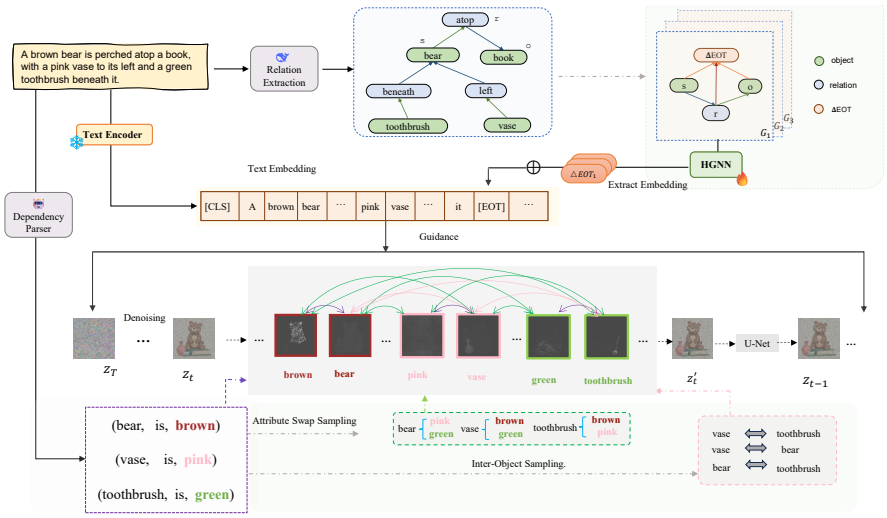

Figure 2: **Overview of the SRA-SD framework.** It consists of two modules. In the *spatial relation enhancement* module, a LLM extracts relational triples from the input text, which are structured into heterogeneous semantic graphs. A graph neural network processes these graphs to refine the `[EOT]` token embedding . In the *attribute enhancement* module, syntactic dependency parsing identifies object-attribute pairs as positive instances, and fine-grained negative samples are generated via attribute swapping and inter-object sampling. Then a contrastive cross-attention loss is then applied during denoising.

2023) propose SynGen, which leverages syntactic dependency parsing to identify entity-modifier pairs and applies a contrastive loss on cross-attention maps to enforce alignment. However, it treats all non-matching tokens as negative samples, leading to an overly large and noisy negative set that leads to an inflated loss value. For spatial relations, Diffusion-SpaceTime-Attn (Wu et al., 2023a) uses a layout predictor and dynamic attention control but it still suffers from object omission and attribute mismatches.

Unlike existing methods that address isolated challenges, our approach integrates structural information from text to simultaneously handle object presence, attribute alignment and spatial relations. This unified framework enables more effective generation of images from complex compositional prompts.

3 Methodology

3.1 Problem Formulation

We investigate the standard text-to-image generation task, aiming to produce images faithful to a given description D. For example, given: "*A brown bear is perched atop a book, with a pink vase to its left and a green toothbrush beneath it.* ". Our research focuses on enhancing the alignment between the generated image and the textual input, emphasizing three key aspects of fidelity:

- **Object Consistency:** The generated image must include all objects specified in D. In our example, the generated image should contain bear, book, vase and toorhbrush.
- **Attribute Accuracy:** The visual attributes of each object in the image must correspond to those described in D. In our example, the bear should be brown, and the vase should be pink.
- **Spatial Relationship Preservation:** The relative spatial arrangement of objects in the image must match the relations described in D. For example, the bear should be positioned atop the book.

3.2 METHOD OVERVIEW

To improve the ability of diffusion models in generating complex scenes with accurate object attributes and spatial relationships, we propose a structure-enhanced text-to-image generation framework that introduces structural priors at different stages of the denoising process. Our method consists of two complementary modules: spatial relation enhancement and attribute enhancement.

In the spatial relation enhancement module, we first employ a large language model to extract the relation on the input text D, constructing heterogeneous semantic graphs. Subsequently, a graph neural network is utilized to learn structured representations, thereby improving the ability of the text embedding $c(D)$ to represent spatial layout semantics.

In the attribute enhancement module, we extract object-attribute pairs via dependency parsing to form the positive set R_d, and propose a fine-grained negative set R_d^- with semantically similar but attribute-mismatched instances. A fine-grained contrastive cross-attention loss is then applied to enforce accurate attribute-object alignment during denoising.

By integrating both modules into the diffusion process, our approach enhances text representations with structured knowledge, leading to more accurate and semantically consistent image generation.

3.3 SPATIAL RELATION MODELING WITH HETEROGENEOUS GRAPHS

To improve the modeling of spatial relations in text-to-image generation, we model relational structures from input text as heterogeneous graphs and enhance the input text embedding through graph representation learning. Given the input text description D, we first employ a pre-trained LLM (DeepSeek-Chat[1]) to extract relation triples, yielding a relation set: R_d = {(bear, atop, book), (toothbrush, beneath, bear), (vase, left, bear)}. For each distinct spatial relation type (e.g., *atop*, *beneath*, *left*), we construct a separate heterogeneous graph to encode the corresponding relational semantics.

Heterogeneous graphs construction. Each graph is built upon one relation type from R_d. Within a graph, objects and the specific relation are represented as two node types: object nodes V_o and relation nodes V_r. To facilitate structured knowledge injection, we introduce a dedicated adjustment node $v_{\Delta\text{EOT}}^{(r)}$ for each relation r, which learns a relation-specific adjustment vector $\mathbf{h}_{\Delta\text{EOT}}^{(r)}$. This design allows the model to capture directional and type-specific spatial dependencies.

Graph Neural Network Learning. For each constructed graph, we initialize the embeddings of object nodes $v_o \in V_o$, relation node $v_r \in V_r$, and the adjustment node $v_{\Delta\text{EOT}}^{(r)}$ using CLIP word embeddings for rich semantic initialization. We then apply GraphSAGE (Hamilton et al., 2017) to perform message passing and learn contextualized representations.

GraphSAGE updates $v_{\Delta\text{EOT}}^{(r)}$ by aggregating features from its local neighborhood through sampling and a learnable aggregation function (e.g., mean or max-pooling). The update rule at layer $l+1$ is:

$$\mathbf{h}_{\Delta\text{EOT}}^{(r,l+1)} = \sigma\left(\mathbf{W}^{(l)} \cdot \text{AGGR}\left(\left\{\mathbf{h}_j^{(l)} \mid j \in \mathcal{N}(i)\right\}\right) + \mathbf{b}^{(l)}\right), \tag{1}$$

where $\mathcal{N}(i)$ denotes the sampled neighbors of node i, and σ is a non-linear activation function.

After L layers of graph propagation, each relation-specific adjustment vector $\mathbf{h}_{\Delta\text{EOT}}^{(r)}$ is obtained. We then aggregate these vectors across all relation types and fuse them with the original [EOT] token embedding:

$$\mathbf{v}_{\text{EOT}}^{*} = \mathbf{v}_{\text{EOT}} + \lambda \cdot \sum_{r \in \mathcal{R}_d} \mathbf{h}_{\Delta\text{EOT}}^{(r)}, \tag{2}$$

where $\mathcal{R}_d$ is the set of relation types in R_d, and $\lambda \in [0, 1]$ controls the overall strength of structural modulation. This formulation explicitly integrates spatial semantics from multiple relational graphs into the final text representation.

Training Objective. To train the GraphSAGE module end-to-end within the diffusion framework, we adopt the standard denoising objective:

$$\mathcal{L}_{\text{denoise}} = \mathcal{E}_{z\sim\mathcal{N}(0,I),y,t}\left[\|\varepsilon - \varepsilon_\theta(x_t, t, \phi(c(y)))\|_2^2\right] \tag{3}$$

[1] https://chat.deepseek.com/

where ϕ denotes the GraphSAGE, x_t is the noised image at timestep t, and y is the corresponding caption. Through $\mathcal{L}_{\text{denoise}}$, GraphSAGE learns to generate adjustment vectors that align the text embeddings of D with the relational semantics present in real images.

3.4 Fine-Grained Attribute Alignment via Contrastive Learning (FCL)

To improve the fidelity of object-attribute bindings in generated images, we enhance the cross-attention mechanisms in diffusion models through fine-grained contrastive learning. Our approach encourages alignment between an object and its correct attribute by minimizing the distance between their attention maps, while maximizing divergence from non-corresponding attributes. The effectiveness of this strategy hinges on high-quality negative sampling, particularly semantically plausible but incorrect pairings that provide strong supervisory signals.

Given an input text description D, we follow (Rassin et al., 2023) to extract object-attribute dependencies using spaCy's transformer-based dependency parser, forming the attribute triple set:

$$A_d = \{(o_i, is, c_i) \mid o_i \in O,\ c_i \in C\} \tag{4}$$

where $O = \{o_1, o_2, \dots\}$ denotes the set of detected objects (e.g., *bear*, *vase*), and $C = \{c_1, c_2, \dots\}$ denotes the set of attribute values (e.g., *brown*, *pink*). For example, from D, we obtain $A_d = \{(\text{bear}, \text{is}, \text{brown}), (\text{vase}, \text{is}, \text{pink}), (\text{toothbrush}, \text{is}, \text{green})\}$.

From A_d, we derive the positive object-attribute pairs:

$$N_{\text{pos}} = \{(o_i, c_i) \mid (o_i, is, c_i) \in A_d\} \tag{5}$$

such as $(\text{bear}, \text{brown})$ and $(\text{vase}, \text{pink})$. We then construct a comprehensive negative sample set A_d^- using two strategies:

- **Attribute Swap Sampling.** We generate negative samples by pairing objects with non-corresponding attributes through attribute permutation. The resulting negative sample set N_{swap} is defined as:
$$N_{\text{swap}} = \{(o_1, c_2), (o_2, c_1) \mid o_1 \neq o_2\} \tag{6}$$
where o_1 and o_2 denote different objects, c_1 and c_2 represent their respective attributes.
- **Inter-Object Sampling.** We construct negative samples by contrasting different objects. The negative sample set N_{inter} is defined as:
$$N_{\text{inter}} = \{(o_i, o_j) \mid j \neq i\} \tag{7}$$
where o_i and o_j represent distinct objects.

Then comprehensive negative sample set $A_d^- = N_{\text{swap}} \cup N_{\text{inter}}$.

Contrastive Cross-Attention Loss. Cross-attention maps in diffusion models serve as a semantic bridge between text tokens and spatial features. To enforce fine-grained alignment, we define a contrastive loss over these maps. For each token pair, we compute its attention map over the latent space and normalize it into a probability distribution.

We measure the dissimilarity between two attention maps M_i and M_j using symmetric Kullback-Leibler (KL) divergence:

$$\text{dist}(M_i, M_j) = \frac{1}{2} D_{\text{KL}}(M_i \,\|\, M_j) + \frac{1}{2} D_{\text{KL}}(M_j \,\|\, M_i), \tag{8}$$

where $D_{\text{KL}}(M_i \,\|\, M_j) = \sum_{\text{pixels}} M_i \log\left(\frac{M_i}{M_j}\right)$.

The positive loss minimizes the distance between matched object-attribute pairs:

$$\mathcal{L}_{\text{pos}} = \sum_{(m,n) \in N_{\text{pos}}} \text{dist}(M_m, M_n). \tag{9}$$

The negative loss maximizes the distance between mismatched pairs:

$$\mathcal{L}_{\text{neg}} = -\sum_{(m,n) \in R_d^-} \text{dist}(M_m, M_n). \tag{10}$$

The overall objective is:

$$\mathcal{L}_{\text{FCL}} = \mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{neg}}. \tag{11}$$

This loss is integrated into the denoising training process, guiding the model to produce images with accurate and semantically consistent attribute bindings.

4 EXPERIMENTS

4.1 BASELINES

Given that model performance generally improves with increasing parameter size, we selected models with approximately 1B for our experiments. We used stable diffusion 1.4 (SD1.4) and Stable Diffusion 2.1 (SD2.1) (Rombach et al., 2022) as our base models for architecture implementation and freeze them in all experiments. In addition, we evaluate several state-of-the-art approaches that address attribute fidelity or spatial relation modeling: (1) Structured Diffusion (Feng et al., 2023). (2) Attend-and-Excite (A&E) (Chefer et al., 2023). (3) SynGen (Rassin et al., 2023). (4) Diffusion-SpaceTime-Attn (Wu et al., 2023a). Detailed descriptions of baselines are provided in Related work, and full experimental settings are given in the appendix.

4.2 DATASETS

Training Dataset. MS-COCO (Lin et al., 2014) is a large-scale dataset consisting of images paired with corresponding captions. We employed DeepSeek-Chat to analyze the entities and their spatial relationships within the caption sentences, selecting only those that explicitly describe spatial relations between entities. After manual filtering, we constructed a final dataset containing 4,717 image-caption pairs.

Evaluation Dataset. We evaluate our approach on two existing datasets and a newly introduced benchmark, **SRA-Bench**, designed to jointly assess object-attribute binding and spatial relations in text-to-image generation.

(1) GPT-synthetic. This dataset contains 500 manually verified prompts generated by GPT-3 (Brown et al., 2020), featuring complex spatial descriptions such as *left of*, *right of*, and *below*. Example: *"The elephant was standing in the center of the room, with the bird to its left"*. It provides linguistically diverse scenarios for evaluating spatial understanding.

(2) ABC-6K. A compositional benchmark with 3.2K human-written prompts from MS-COCO, each containing at least two color-modified objects (e.g., *"A blue cat is wearing a yellow plastic baseball hat"*). We randomly sample 500 instances for evaluation.

(3) SRA-Bench. It is built upon SR_{2D} (Gokhale et al., 2022), which contains two-dimensional spatial relations (e.g., left/right/above/below) between object pairs commonly found in MS-COCO. The original sentences follow the format: "*a* {*objectA*} {*relation*} *a* {*objectB*}". We extend this by enriching each object with descriptive attributes using the DeepSeek-Chat, primarily focusing on color attributes due to their visual saliency. This results in an updated sentence structure: "*a* {*colorA*} {*objectA*} {*relation*} *a* {*colorB*} {*objectB*}".

To increase spatial diversity, we introduce four additional spatial relations: *inside*, *on*, *behind*, and *in front of*, incorporated via relation replacement with DeepSeek-Chat. Only syntactically and semantically valid outputs are retained after manual filtering. The final benchmark consists of 100 verified instances per relation type (8 types total). Additional details are provided in appendix.

4.3 AUTOMATIC EVALUATION

To quantitatively assess the alignment between generated images and textual descriptions, we adopt two complementary evaluation paradigms: (1) a **fine-grained composite metric** that explicitly evaluates object presence, spatial relations, and attribute fidelity; and (2) **holistic quality metrics**, including CLIPScore (Radford et al., 2021) and HPSv2 (Wu et al., 2023b), which assess text-image consistency from representation-based and human-preference perspectives, respectively.

Model	Params ∼(B)	Object Recall A	Object Recall B	Object Recall AB	VISOR (AB)_a	VISOR (AB)_c	Attribute Fidelity A	Attribute Fidelity B	Attribute Fidelity (AB)_a	Attribute Fidelity (AB)_c	Final	Human
SD 1.4	1	70.88	57.62	30.62	17.63	57.55	61.13	38.00	20.13	65.71	11.63	12.92
SD 2.1	1.2	76.25	67.88	45.50	26.38	57.97	64.25	50.00	31.38	68.96	18.50	23.33
Structured Diffusion	1	58.75	61.50	23.88	13.25	55.50	45.63	44.00	15.63	65.45	9.13	5.42
Diffusion-SpaceTime-Attn	1+125M	64.12	65.38	34.75	20.63	59.35	55.88	51.75	24.50	70.50	14.75	17.92
SynGen	1	80.88	65.13	50.13	26.00	51.87	77.88	62.00	46.50	**92.77**	24.88	39.58
SRA_SD (1.4)	1+2.95M	85.88	69.38	57.38	34.00	59.25	80.38	66.00	52.38	91.27	31.13 ↑ 167.7%	42.08
SRA_SD (2.1)	1.2+5.25M	**94.12**	**76.00**	**70.75**	**46.50**	**65.72**	**87.38**	**71.75**	**63.50**	89.75	**41.00** ↑ 121.6%	**57.92**

Table 1: Results (%) of our method and other baselines on SRA-Bench. The **bold** represents the best result. (AB)_a and (AB)_c represent overall and conditional accuracy, respectively.

4.3.1 Composite Metric: Object Recall, VISOR, and Attribute Fidelity

The composite metric consists of three components: (1) **Object Recall** (OR), (2) **VISOR**, and (3) **Attribute Fidelity** (AF). These metrics evaluate different aspects of the model's ability to generate images consistent with input text. We also introduce a combined metric, denoted as **Final**, to jointly assess the accuracy of the spatial and attribute.

1. Object Recall (OR) (Gokhale et al., 2023) measures the percentage of objects mentioned in the text that are successfully detected in the generated image. For an image x generated by a sentence T of the form "*a* {*colorA*} {*objectA*} {*relation*} *a* {*colorB*} {*objectB*}", OR is defined as:

$$OR(x, O_A, O_B) = \begin{cases} 1, & \text{if } F(\exists O_A \cap \exists O_B) \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

where F denotes an object detection model. In this work, we use OWL-ViT (Minderer et al., 2022), a CLIP-based open-vocabulary detector built on the ViT-B/32 architecture, with a confidence threshold of 0.1.

2. VISOR (Gokhale et al., 2023) evaluates the correctness of the spatial relation R between two detected objects A and B. Specifically, for an image x, if both objects are present, their highest-confidence bounding boxes are extracted and used as input to a vision-language model to verify whether the spatial relation matches the description. The metric is defined as:

$$VISOR(x, O_A, O_B) = \begin{cases} 1, & \text{if } M(x', O_A, O_B) \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

Here, M is a vision-language model and we employ Qwen2-VL-7b[2] in this work. We report two variants of VISOR: $\textbf{VISOR}_a$, the overall accuracy across all samples, and $\textbf{VISOR}_c$, the conditional accuracy given both objects are correctly generated.

3. Attribute Fidelity (AF) measures how well the attributes (e.g., color) of the generated objects match the textual description. For each detectable object, we crop its region and use a vision-language model to verify attribute consistency. Formally:

$$AF(x, C_A, O_A, C_B, O_B) = \begin{cases} 1, & \text{if } M(x', C_A, O_A) \cap M(x', C_B, O_B) \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

where C_A and C_B represent the expected colors of objects O_A and O_B, respectively. Similar to VISOR, we define $\textbf{AF}_a$ and $\textbf{AF}_c$ to evaluate overall and conditional attribute accuracy.

4. Final combines both spatial and attribute evaluations into a unified score. It is defined as:

$$Final(T) = \begin{cases} 1, & \text{if } VISOR(x, O_A, O_B) \cap AF(x, C_A, O_A, C_B, O_B) \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

This metric reflects the model's ability to simultaneously satisfy both relational and attribute constraints in image generation.

4.3.2 CLIPScore and HPSv2

To complement our fine-grained evaluation, we report two global metrics for text-image alignment: CLIPScore (Radford et al., 2021) and HPSv2 (Wu et al., 2023b). CLIPScore computes the cosine

[2]https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct

Model	ABC_6K		GPT-synthetic		SRA-Bench	
	CLIPscore	HPSv2	CLIPscore	HPSv2	CLIPscore	HPSv2
SD 1.4	0.322	28.23	0.311	26.45	0.318	28.50
SD 2.1	0.330	28.83	0.322	26.92	0.328	29.66
Structured Diffusion	0.324	26.90	0.291	26.91	0.318	29.22
Diffu-SpaceTime-Attn	0.329	28.42	0.294	27.11	0.329	28.90
SynGen	0.325	28.72	0.311	26.75	0.330	30.06
SRA_SD (1.4)	0.337	29.08	0.321	26.81	0.343	30.22
SRA_SD (2.1)	**0.339**	**29.44**	**0.324**	**27.66**	**0.347**	**30.81**

Table 2: The results(%) of CLIPScore and HPSv2 on three datasets. The **bold** represents the best result.

similarity between CLIP embeddings of the input text and generated image, serving as a proxy for semantic consistency in the representation space. HPSv2, by contrast, is trained on large-scale human preference data to predict which image better aligns with a caption from a human perspective, thereby approximating judgments of realism, relevance, and overall generation quality.

Results: As shown in Table 1, our proposed method achieves state-of-the-art performance on the SRA-Bench, significantly outperforming both standard diffusion models and specialized baselines. On the overall **Final** score, SRA_SD (1.4) and SRA_SD (2.1) achieve a remarkable improvement of **+167.7%** and **+121.6%** over their respective base models (SD 1.4 and SD 2.1), with only **2.95M** and **5.25M** additional parameters. This demonstrates that our framework delivers substantial gains in semantic alignment at minimal parameter cost. A closer look reveals strengths in both spatial reasoning and attribute fidelity. In terms of $\textbf{VISOR}_c$, our method achieves 66.00% and 71.75%, surpassing all baselines including Diffusion-SpaceTime-Attn, despite being specifically designed for spatial modeling. While SynGen achieves the highest AF_c, indicating strong attribute generation capability, it underperforms in spatial relation handling.

Our method also achieves the highest CLIPScore and HPSv2 across all three datasets (Table 2), with the most significant gains observed on SRA-Bench. This pattern validates our design: SRA-Bench is explicitly constructed to evaluate the joint understanding of spatial and attribute semantics, which aligns perfectly with our model's strengths. The smaller gains on GPT-synthetic and ABC-6K are expected, as these datasets either contain semantically complex but less spatially structured phrases (e.g., "walking", "placed" in GPT-synthetic) or focus primarily on color attributes (ABC-6K), where our spatially-focused HGNN provides less relative advantage. Notably, improved semantic alignment also enhances visual quality (HPSv2), suggesting that accurate relational reasoning reduces hallucination and improves generation coherence.

In summary, our method achieves superior performance across all datasets while maintaining high parameter efficiency, demonstrating that lightweight, semantics-driven design enables effective joint modeling of spatial and attribute relationships—leading to more accurate and robust structured image generation.

4.4 Subjective Evaluation

To further assess the fidelity and quality of generated images, we conduct a subjective evaluation on SRA-Bench. For each of the eight spatial relation categories, we randomly select 10 samples from 100 text descriptions, resulting in 80 annotated samples. Annotation was conducted by three annotators with master's-level education. Each prompt follows the format: "*a* {*colorA*} {*objectA*} {*relation*} *a* {*colorB*} {*objectB*}". Annotators are asked to verify:

1. "Is there a {*colorA*} {*objectA*} in the image?"
2. "Is there a {*colorB*} {*objectB*} in the image?"
3. "Is the {*relation*} between {*objectA*} and {*objectB*}?"

An image is considered as correct only if all the three answers are "YES". The final score, denoted as **Human**, is computed as the overall accuracy.

The results, presented in the rightmost column of Table 1, show that our method significantly outperforms the base models. Notably, even when using SD 1.4 as the backbone, our framework surpasses all baseline methods, further demonstrating its effectiveness. In particular, human evaluation scores exhibit strong consistency with the **Final** metric, further validating the effectiveness and rationality of our fine-grained evaluation approach in addressing complex compositional text prompts.

We also perform qualitative comparisons on three datasets. We observe that existing methods (e.g., Diffusion-SpaceTime-Attn and SynGen) often fail to generate accurate object configurations—either missing objects, mismatching attributes, or misplacing them spatially. In contrast, SRA-SD consistently generates images that accurately align with the prompt semantics. For visual examples and further analysis, please refer to the appendix.

Model	OR(AB)	VISOR(_a)	AF(_a)	Final
RGAT+FCL	57.12	32.50	51.88	29.38
RGCN+FCL	53.12	31.63	51.50	29.13
GraphSAGE+FCL	57.38	34.00	52.38	31.13

(a) Results of different HGNN models

Model	OR(AB)	VISOR(_a)	AF(_a)	Final
w/o GraphSAGE	51.00	26.38	51.88	29.38
w/o FCL	33.62	20.50	24.00	15.00
GraphSAGE+SynGen	55.25	33.13	51.75	30.75
Our	57.38	34.00	52.38	31.13

(b) Ablation study on module components

Table 3: Results of ablation study

Model	Params(B)	ABC_6K		GPT-synthetic		SRA-Bench	
		CLIPscore	HPSv2	CLIPscore	HPSv2	CLIPscore	HPSv2
FLUX	12	0.330	30.61	0.324	28.75	0.343	31.31
SD_XL	3.3	0.341	29.84	0.329	26.45	0.318	28.50
SRA_SD (2.1)	1.2+5.25M	0.339	29.44	0.324	27.66	0.347	30.81

Table 4: Results of comparison with large-scale models on three datasets

4.5 Ablation Study

Effect of two Components. To validate the necessity and effectiveness of each module, we conduct ablation studies by systematically removing or replacing key components. Results are summarized in Table 3b.Removing GraphSAGE (w/o HGNN) leads to a significant drop in VISOR($\_a$) with only minor degradation in OR(AB), indicating that GraphSAGE is crucial for modeling spatial relationships, particularly for preserving structural coherence. In contrast, removing FCL degrades both OR(AB) and AF($\_a$), demonstrating that FCL module plays a vital role in preventing object omission and ensuring attribute-level alignment (e.g., color, texture) during denoising. To isolate the effect of our attention map constraint within FCL, we replace it with SynGen's design while retaining GraphSAGE. The variant scores 30.75 above w/o HGNN but below our model demonstrating that our FCL's attention map regularization yields superior fine-grained feature alignment during denoising.

Effect of differernt HGNN models. We evaluate three types of HGNNs: GraphSAGE, RGAT (Busbridge et al., 2019), and RGCN (Schlichtkrull et al., 2018). As shown in table 3a, GraphSAGE achieves the best performance, especially in OR(AB) and VISOR($\_a$), indicating robust object preservation and structural coherence. RGAT matches GraphSAGE in OR(AB) but lags in VISOR($\_a$), suggesting weaker spatial modeling. RGCN shows the lowest OR(AB), implying it struggles to retain object presence during denoising — possibly due to less effective neighborhood aggregation under noisy inputs. This highlights that sampling-based aggregation (GraphSAGE) better preserves semantic completeness than message-passing variants in our task.

4.6 Impact of LLM Choice and Modulation Strength (λ)

We evaluate DeepSeek, GPT-3.5[3], and Qwen_2.5_72b[4] as relation extractors with varying λ on GPT-synthetic, a dataset with complex spatial descriptions. Results are shown in Figure 4. We can find that small performance variation across LLMs, indicating that modern models provide comparable

[3] https://openai.com/gpt-3.5

[4] https://huggingface.co/Qwen/Qwen2.5-72B-Instruct

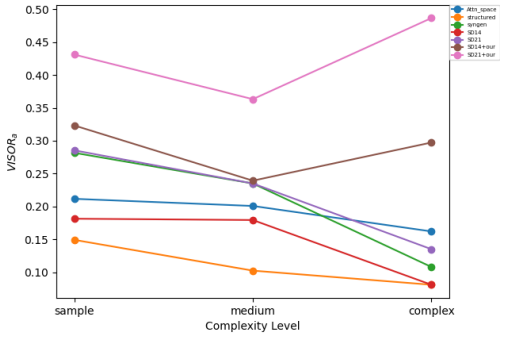

Figure 3: Impact of Text Complexity on $VISOR_a$.

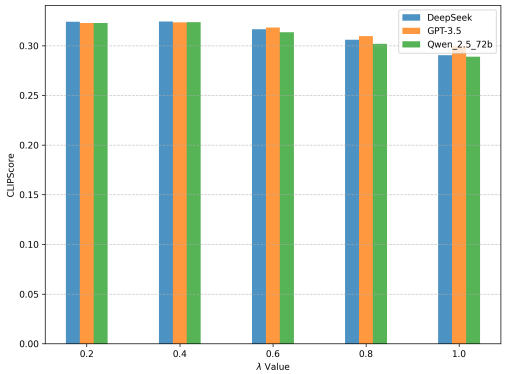

Figure 4: CLIPScore under different LLMs and modulation weights λ on GPT-synthetic

relational priors. However, performance drops as λ increases, suggesting weaker structural modulation is more effective in multi-relation scenes to avoid over-correction.

4.7 Robustness to Prompt Complexity

Impact of Prompt Complexity on Spatial Relation Understanding. To assess how scene complexity affects spatial relation modeling, we classify SRA-Bench prompts into three levels (simple, medium, difficult) based on commonsense object and relation properties (see appendix for criteria). We then conducted a stratified analysis using the VISOR metric, with results shown in Figure 3.

As shown in figure, most models exhibit a decline in performance as prompt complexity increases. In contrast, our method consistently achieves the best results across all complexity levels, regardless of whether SD 1.4 or SD 2.1 is used as the backbone. Notably, it shows improved performance on the most complex cases, while other methods deteriorate. Given that spatial relations constitute only 37.5% of the training data, this robustness highlights the effectiveness of our HGNN framework in capturing compositional semantics.

4.8 Comparison with Large-Scale Models

We compare our method with state-of-the-art large-scale models on three datasets, as summarized in Table 4. In terms of CLIPScore, our approach underperforms on ABC-6K and GPT-synthetic but achieves the best performance on SRA-Bench. We attribute this to the fact that SRA-Bench explicitly contains object-object relationships and object-attribute bindings, semantic structures that our framework is designed to model, while the other two datasets do not consistently emphasize such semantics. Notably, FLUX performs best on HPSv2 due to its superior visual quality from larger capacity.

Despite underperforming on certain datasets, our framework remains valuable: it introduces only a negligible number of additional parameters (5.25M), making it highly lightweight and easy to integrate. More importantly, in scenarios where both object attributes and relational semantics are present, such as SRA-Bench, it even surpasses much larger models, demonstrating its effectiveness in structured semantic alignment.

5 Conclusion

In the work, we propose SRA-SD, a lightweight structure-aware framework that enhances spatial relations and object-attribute alignment via two complementary modules. To support joint evaluation, we introduce SRA-Bench. Experiments show SRA-SD outperforms existing methods in complex scenarios with minimal parameter overhead, demonstrating its effectiveness and robustness.

REFERENCES

David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. *CoRR*, abs/2103.10951, 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.

Dan Busbridge, Dane Sherburn, Pietro Cavallo, and Nils Y. Hammerla. Relational graph attention networks. *CoRR*, abs/1904.05811, 2019.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Trans. Graph.*, 42 (4):148:1–148:10, 2023.

Colin Conwell and Tomer D. Ullman. Testing relational understanding in text-guided image generation. *CoRR*, abs/2208.00005, 2022.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun R. Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *ICLR*. OpenReview.net, 2023.

Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*, 2022.

Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation, 2023. URL https://arxiv.org/abs/2212.10015.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, pp. 10686–10696. IEEE, 2022.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NIPS*, pp. 1024–1034, 2017.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*. OpenReview.net, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV (5)*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014.

Matthias Minderer, Alexey A. Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. *CoRR*, abs/2205.06230, 2022.

Pietro Perona and Jitendra Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(7):629–639, 1990.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831. PMLR, 2021.

Royi Rassin, Shauli Ravfogel, and Yoav Goldberg. DALLE-2 is seeing double: Flaws in word-to-concept mapping in text2image models. In *BlackboxNLP@EMNLP*, pp. 335–345. Association for Computational Linguistics, 2022.

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. In *NeurIPS*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10674–10685. IEEE, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *ESWC*, volume 10843 of *Lecture Notes in Computer Science*, pp. 593–607. Springer, 2018.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*. OpenReview.net, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pp. 5998–6008, 2017.

Qiucheng Wu, Yujian Liu, Handong Zhao, Trung Bui, Zhe Lin, Yang Zhang, and Shiyu Chang. Harnessing the spatial-temporal attention of diffusion models for high-fidelity text-to-image synthesis. In *ICCV*, pp. 7732–7742. IEEE, 2023a.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 2096–2105. IEEE, 2023b. doi: 10.1109/ICCV51070.2023.00200. URL https://doi.org/10.1109/ICCV51070.2023.00200.

Yinwei Wu, Xingyi Yang, and Xinchao Wang. Relation rectification in diffusion model. In *CVPR*, pp. 7685–7694. IEEE, 2024.

A Appendix

B The use of LLM

We used large language models (e.g., GPT-4, DeepSeek) only for language editing and proofreading. They were not involved in idea generation, experimental design, data analysis, or technical writing. All content is authored by the researchers.

C Implementation details

C.1 Hyperparameters

For GraphSAGE, we adopt a two-layer architecture and train it for 50 epochs with a batch size of 40. The FCL phase follows the configuration used in SynGen (Rassin et al., 2023). We set λ from 0.1 1.0.

C.2 Optimization Strategy

We apply the loss functions from Eqs. (8)-(10) during the first 25 out of 50 denoising steps. In each of these steps, a pretrained U-Net first denoises the latent variable z_t. We then compute the cross-attention maps as described in (Hertz et al., 2023). Next, we update the latent representation z_t using the loss $\mathcal{L}$ via a gradient step: $z'_t = z_t - \alpha \cdot \nabla_{z_t}\mathcal{L}$. Finally, the U-Net denoises the updated latent variable z'_t for the next timestep.

C.3 The Workflow

During the training phase, we train GraphSAGE using the training dataset. In the inference phase, the trained GraphSAGE is integrated into the generation pipeline for the evaluation dataset.

C.4 Efficiency.

While our framework introduces additional computational overhead compared to the base models, it remains more efficient than the state-of-the-art baseline, SynGen. Specifically, image generation with SRA_SD (2.1) and SRA_SD (1.4) takes 7.0 and 7.6 seconds per image on average, respectively, compared to 1.7 and 1.4 seconds for the original SD 2.1 and SD 1.4 models. Despite this increase in latency due to spatial relation conditioning, our method achieves a 22.2% speedup over SynGen , which requires approximately 9.0 seconds per image. To ensure a fair comparison, we measured inference time by randomly sampling 20 images from both the SRA_Bench and GPT-synthetic datasets and averaging the generation times across all methods.

C.5 Computing resources.

All experiments are conducted on an NVIDIA A800 GPU with 80GB memory, leveraging CUDA 12.2 and PyTorch 2.4.0 for efficient computation.

D SRA-Bench Details

D.1 Object Categories

The SRA-Bench includes a diverse set of object categories commonly found in the MS-COCO dataset. The complete list of object categories used in the benchmark is as follows:

- **Objects**: bed, frisbee, cup, oven, keyboard, donut, cake, microwave, train, flower, cat, camera, umbrella, backpack, tv, airplane, mouse, banana, case, key, snowboard, bus, rabbit, skateboard, bowl, zebra, dog, hat, box, notebook, motorcycle, scissors, bench, bottle, vase, parrot, laptop, refrigerator, suitcase, horse, remote, skis, bag, carrot, giraffe, sandwich, surfboard, cow, durian, toilet, broccoli, car, couch, toy, handbag, pizza, boat, book, tie, truck, apple, sink, sheep, chair, kite, bird, toothbrush, bicycle, orange, clock.

D.2 Attributes

To enhance the complexity of the benchmark, we assign color attributes to each object. The colors used in SRA-Bench are selected from a diverse palette to ensure variability and realism. The complete list of color attributes is as follows:

- **Colors**: bronze, yellow, white, copper, red, peach, maroon, black, lime, brown, cobalt, pink, indigo, gray, emerald, purple, sage, ivory, teal, blue, silver, coral, amber, plum, tan, lavender, jade, green, navy, charcoal, turquoise, olive, beige, orange.

D.3 Spatial Relations

The benchmark evaluates the following spatial relations:

- **Basic Relations**: left, right, above, below.

Figure 5: Qualitative comparisons for our model and baselines. The images synthesized by our method have the best compositional alignment compared with other baselines.

- **Extended Relations**: inside, on, behind, in front of.

These relations are applied to pairs of objects with assigned color attributes, resulting in sentences of the format: *"a {colorA} {objectA} {relation} a {colorB} {objectB}."*

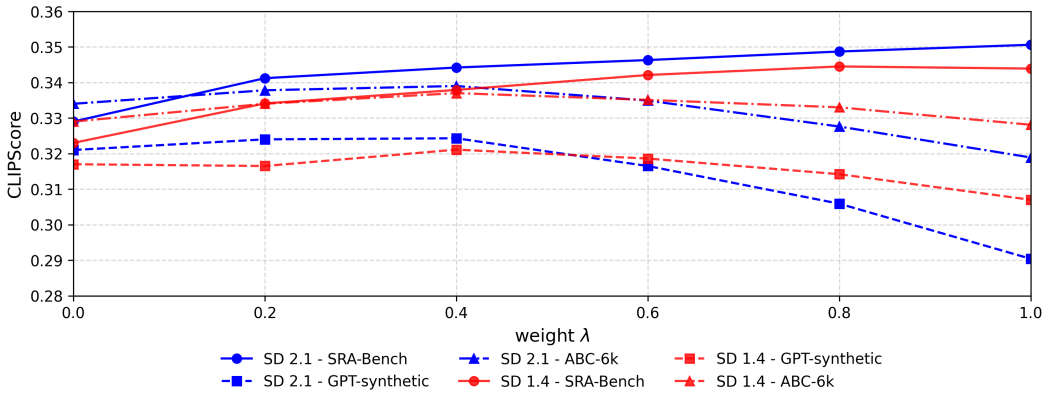

Figure 6: Impact of λ on CLIPscore

E Qualitative Comparisons

In Figure 5, we present the generated results for various prompts across three datasets: SRA-Bench, ABC-6K, and GPT-synthetic. Prompts (a), (b), (c), and (d) are from SRA-Bench, which focuses on spatial relations and attribute binding; prompts (e), (f), and (g) are from ABC-6K, a compositional benchmark with human-written prompts containing color-modified objects; and prompts (h) and (i) are derived from GPT-synthetic, featuring complex and linguistically diverse scenarios involving multiple entities or detailed attribute descriptions (e.g., object actions and spatial relations). For each prompt, we compare seven models: SD 1.4, SD 2.1, StructuredDiffusion, SynGen, Diffusion-SpaceTime-Attn, SRA_SD (1.4), and SRA_SD (2.1).

Existing methods show significant limitations in handling complex prompts. For example:

- **SD 1.4** and **SD 2.1**: The base diffusion models often fail to preserve compositional structure, generating missing objects (e.g., "black remote" in (c)), incorrect spatial arrangements (e.g., "left of" in (a), "inside" in (b)), and attribute mismatches (e.g., "black flowers" in (e)). While SD 2.1 shows slight improvements over SD 1.4, both struggle with precise semantic grounding.
- **StructuredDiffusion** misses objects (e.g., "black TV" in (a), "fire hydrant" in (g)), mismatches attributes (e.g., "rabbit" and "box" colors in (b)), and misplaces objects spatially.
- **Diffusion-SpaceTime-Attn** captures spatial relations but misses objects (e.g., "black TV" in (a), "screen" in (f)) and mismatches attributes (e.g., "rabbit" and "box" colors in (b)).
- **SynGen** aligns colors well but fails in object accuracy (e.g., "boat" in (a), "keyboard" in (f)) and spatial relations (e.g., "inside" relationship in (c))

In contrast, our method, **SRA-SD**, effectively addresses these limitations by simultaneously enhancing spatial relation understanding and object-attribute alignment through structured graph learning and fine-grained contrastive supervision. As shown in Figure 5, SRA-SD generates images that more accurately reflect the prompt semantics, with correct object placement, faithful attribute binding, and minimal hallucinations. This demonstrates its superior capability in handling complex, compositionally rich text-to-image generation tasks.

F The analysis of Adjustment Strength λ

Impact of Adjustment Strength λ. As shown in Figure 6, performance first improves and then declines with λ, indicating that moderate structural enhancement is beneficial, while excessive injection introduces noise. The gain is largest on SRA-Bench, moderate on ABC-6K, and smallest on GPT-synthetic. We attribute this to parsing reliability: SRA-Bench has clean, unambiguous prompts, yielding accurate relational graphs. In contrast, GPT-synthetic uses free-form text, where relation

parsing is more error-prone. As λ increases, the model over-trusts these noisy structures, harming generation. Thus, the benefit of structural enhancement depends on the fidelity of the parsed structure.

G TEXT COMPLEXITY CLASSIFICATION CRITERIA

To evaluate the impact of scene complexity on spatial relation understanding, we classified the text prompts in the SRA-Bench dataset into three levels of difficulty: **simple**, **medium**, and **difficult**. The classification was based on common sense factors, such as object size and the typicality of object relations. The criteria for each category are defined as follows:

G.1 SIMPLE

Texts describe scenes with common objects and relations that are easily observable in daily life. Key features include:

- Objects have typical sizes and proportions.
- Relations between objects are frequently encountered.

Example: "A red apple on a table."

G.2 MEDIUM

Texts describe plausible but moderately complex scenes. Key features include:

- Object combinations or relations are less common or require some abstraction.
- Scenarios are observed less frequently in real life.

Example: "A green book on the ground."

G.3 DIFFICULT

Texts describe highly complex scenes involving uncommon or unlikely object combinations and relations. Key features include:

- Relations between objects are rare or require abstract conceptual understanding.
- Scenarios are improbable or rarely observed.

Example: "A blue balloon floating in the air, next to a transparent crystal."

This classification framework provides a systematic basis for analyzing model performance across varying levels of scene complexity.