Local Prompt Adaptation for Style-Consistent Multi-Object Generation in Diffusion Models

Ankit Sanjyal[1]

[1]Fordham University, as505@fordham.edu

Abstract

Diffusion models have become a powerful backbone for text-to-image generation, enabling users to synthesize high-quality visuals from natural language prompts. However, they often struggle with complex prompts involving multiple objects and global or local style specifications. In such cases, the generated scenes tend to lack style uniformity and spatial coherence, limiting their utility in creative and controllable content generation. In this paper, we propose a simple, training-free architectural method called Local Prompt Adaptation (LPA). Our method decomposes the prompt into content and style tokens, and injects them selectively into the U-Net's attention layers at different stages. By conditioning object tokens early and style tokens later in the generation process, LPA enhances both layout control and stylistic consistency. We evaluate our method on a custom benchmark of 50 style-rich prompts across five categories and compare against strong baselines including Composer, MultiDiffusion, Attend-and-Excite, LoRA, and SDXL. Our approach outperforms prior work on both CLIP score and style consistency metrics, offering a new direction for controllable, expressive diffusion-based generation.

1 Introduction

Text-to-image diffusion models have witnessed a significant leap in recent years, propelled by breakthroughs such as Denoising Diffusion Probabilistic Models (DDPM) [1], improved latent architectures [2], and large-scale language-vision training paradigms [3]. These models learn to synthesize images by progressively denoising a latent or pixel space guided by textual prompts, resulting in photorealistic and often imaginative outputs. With the advent of large-scale pre-trained models such as Stable Diffusion XL (SDXL), these capabilities have reached a broader user base through accessible interfaces and creative applications.

Despite these advancements, existing models still face considerable limitations when prompted with complex compositions—particularly prompts that include multiple entities, intricate spatial arrangements, or globally applied artistic styles. For instance, a prompt such as "a cat on a flying car in vaporwave style" often results in images where the style is inconsistently applied—only one object may reflect the desired aesthetic, or the spatial layout between elements becomes incoherent. Such inconsistencies reveal a fundamental bottleneck in the current prompt conditioning pipeline: all tokens are treated uniformly, regardless of their semantic role.

Our hypothesis is grounded in the temporal nature of the diffusion process itself. We argue that spatial structure and stylistic refinement emerge at different stages in the denoising trajectory. Motivated by this, we propose a simple yet effective architectural strategy called **Local Prompt Adaptation (LPA)**. Rather than encoding the entire prompt uniformly across all layers, we segment the prompt into functionally distinct sub-components—*object tokens* and *style tokens*—and inject them into separate stages of the U-Net.

Specifically, LPA introduces targeted cross-attention injection wherein object tokens are routed to early downsampling blocks to guide spatial layout, while style tokens are injected into middle and late blocks to control high-level texture and appearance. This enables a more coherent compositional understanding, wherein each object is both spatially grounded and stylistically aligned to the global prompt specification. Our method operates *without any fine-tuning or retraining*, making it a practical solution for large models such as SDXL.

To validate our hypothesis, we construct a benchmark of 50 prompts spanning five categories of increasing compositional complexity. We compare LPA against baselines including vanilla SDXL, high-classifier-free-guidance SDXL, Attend-and-Excite [4], Composer [5], ControlNet [6], LoRA [7], and MultiDiffusion [8]. Quantitative results using CLIP score, style consistency metrics, and LPIPS demonstrate that LPA consistently produces more faithful and globally styled generations. Qualitative comparisons further highlight improved semantic grounding and visual harmony.

All code, prompts, and evaluation scripts are available open-source at: https://github.com/ANKITSANJYAL/local-style-diffusion.

2 Related Work

2.1 Text-to-Image Diffusion Models

Diffusion models have rapidly become the backbone of modern generative image synthesis. Starting with Denoising Diffusion Probabilistic Models (DDPM) [1], the field has progressed toward scalable and efficient variants such as Latent Diffusion Models (LDM) [2], which operate in com-

pressed latent spaces to drastically reduce memory and sampling costs. More recently, Stable Diffusion XL (SDXL) introduced larger architectures and refined conditioning techniques, pushing the boundaries of text-to-image realism and accessibility.

2.2 Prompt Compositionality and Attention Guidance

The challenge of generating coherent images from compositionally complex prompts has led to methods that enhance spatial attention and token specificity. Attend-and-Excite [4] guides the diffusion process to ensure all prompt elements are faithfully attended to during generation. MultiDiffusion [8] explores combining outputs from multiple diffusion paths to enhance compositional integrity, while Composer [5] proposes a modular decoding system for assembling semantic regions under prompt control. Although effective, many of these require modified sampling strategies or retraining, making them harder to deploy in real-world setups.

2.3 Style and Structure Control in Diffusion

Parallel efforts have sought greater control over style and structure through conditioning or fine-tuning. ControlNet [6] introduces trainable adapters that align generation with structural priors like pose or depth maps. LoRA [7], initially proposed for language models, has been adapted to image diffusion to enable lightweight, targeted fine-tuning. However, these methods either rely on paired data or require training-specific modules, limiting their zero-shot applicability to style consistency across multiple objects.

2.4 Positioning Our Work

Our approach differs by focusing on inference-time control without architectural changes or retraining. By selectively routing object and style tokens to different stages of the U-Net's attention mechanism, we address both compositional grounding and stylistic alignment in a unified framework. Unlike existing solutions that require auxiliary modules or data-specific training, our method is compatible with off-the-shelf models and enhances control using only prompt structure and cross-attention manipulation.

3 Method

3.1 Prompt Token Segmentation

The first step in our approach is to disentangle the prompt into semantically distinct components that serve different roles in the image generation process. Natural language prompts often blend content descriptors (e.g., "a cat", "a flying car") with stylistic cues (e.g., "vaporwave style", "in ukiyo-e style"). Treating all tokens uniformly, as in standard diffusion pipelines, ignores this distinction and can result in outputs where either spatial coherence or stylistic consistency is compromised.

To address this, we adopt a linguistic parsing-based strategy to segment prompts into two disjoint token sets:

- **Object Tokens** (T_{obj}): Noun phrases or entities that define spatial elements to be grounded in the image.
- **Style Tokens** (T_{style}): Adjectives, stylistic descriptors, or artistic genres that globally influence appearance.

We implement this via spaCy's dependency parser, which efficiently extracts noun chunks and modifier dependencies from the prompt. For example, given the prompt:

"A cat on a flying car in vaporwave style", our parser identifies:

$$T_{obj} = \{\texttt{cat}, \texttt{flying car}\}, \quad T_{style} = \{\texttt{vaporwave}\}$$

This decomposition is fully automated, robust to prompt length, and does not require any additional labeling. By isolating T_{obj} and T_{style}, we can inject them independently at different stages of the diffusion process (detailed next). Importantly, this modular token routing scales linearly with prompt length, and can generalize to prompts containing multiple styles or hierarchies of objects.

Moreover, this token-level decomposition aligns well with transformer-based U-Nets used in diffusion models, where attention layers are explicitly conditioned on token embeddings. It enables fine-grained control over which linguistic attributes influence which spatial and visual layers of the network.

3.2 Controlled Cross-Attention Injection

In diffusion models such as Stable Diffusion XL, the U-Net architecture is heavily influenced by cross-attention mechanisms that align image features with the input text prompt. These cross-attention layers appear across all resolution stages of the U-Net—from early downsampling blocks that shape coarse layout, to middle bottleneck layers, and finally to late upsampling blocks responsible for detail refinement.

In standard pipelines, the full prompt embedding is broadcast uniformly to all cross-attention layers and all timesteps during denoising. This ignores the fact that different semantic components of the prompt—objects and style—may be most relevant at different stages of the generation process. We hypothesize that:

- **Object tokens** primarily influence the spatial layout and should dominate early attention.
- **Style tokens** shape texture, palette, and mood, and are best applied later in the generation.

To realize this, we implement **selective token injection** by modifying the cross-attention layers in SDXL's U-Net. Specifically, for each block B_i and denoising timestep t, we inject only a subset of prompt tokens into the cross-attention layer:

$$\text{CrossAttn}^{t}_{B_i}(\cdot) \leftarrow \begin{cases} T_{obj} & \text{if } B_i \in \text{Down Blocks and } t < 35 \\ T_{style} & \text{if } B_i \in \text{Mid/Up Blocks and } t \geq 35 \end{cases}$$

This routing scheme is simple yet effective: we only need to mask out irrelevant token embeddings in the attention module without modifying the network architecture or training procedure. In practice, we overwrite the attention mask or the key/value matrices in HuggingFace's 'CrossAttention' implementation to control token visibility at runtime.

We empirically set $t = 35$ as the boundary between coarse and fine stages, but this can be adapted. Our design ensures that object tokens dominate early generation (spatial structure), while style tokens reinforce coherence in the later stages (visual refinement).

This approach is highly scalable: it requires no additional parameters, is batch-invariant, and supports per-layer customization. More importantly, it aligns closely with how human users intuitively expect compositional prompts to be interpreted—first placing objects, then applying stylistic rules.

3.3 Spatial Attention Localization

Understanding how specific prompt tokens influence different regions of the generated image is essential not only for interpretability but also for validating whether the token injection strategy is functioning as intended. To this end, we record and analyze the spatial attention maps generated by the U-Net's cross-attention layers during inference.

Let $A_{ij}^{(t)}$ represent the cross-attention weight at denoising timestep t, where i indexes spatial locations in the feature map and j indexes the prompt tokens. These attention maps effectively encode the influence of each token on each spatial location, as computed by the transformer-style attention mechanism:

$$A_{ij}^{(t)} = \text{softmax}\left(\frac{Q_i K_j^\top}{sqrtd_k}\right)$$

where Q and K are the query and key matrices, and d_k is the key dimension.

By extracting $A_{ij}^{(t)}$ from multiple attention layers (especially from early, mid, and late stages), we obtain token-specific activation heatmaps over the image grid. These heatmaps serve several purposes:

- They confirm that object tokens activate expected spatial regions (e.g., "cat" $\rightarrow$ cat's location).
- They verify that style tokens primarily activate background and texture-related regions.
- They enable users to debug, visualize, or interact with generation in interpretable ways.

In our implementation, we register forward hooks on the U-Net's cross-attention modules to capture these maps at runtime, with minimal overhead. Since the attention weights are normalized over tokens, we normalize and upsample the attention grids to match the output image resolution for visualization.

This attention traceability enhances the transparency of our method. Compared to opaque diffusion pipelines where prompt effects are difficult to localize, LPA offers explainable generation by linking linguistic elements to spatial activations. This is particularly valuable in creative or educational tools, where understanding "why" an image looks a certain way is as important as the image itself.

3.4 Optional: Style Consistency Loss

While our Local Prompt Adaptation (LPA) method enhances style control via architectural injection, we optionally apply a post-generation re-ranking step to select the most stylistically consistent samples from a batch. This is particularly useful when sampling multiple outputs per prompt and desiring the one most aligned with the intended artistic style.

Let R_i denote the visual region corresponding to object i in the generated image, and let S represent the prompt's target style token. We embed both using a pretrained vision-language model such as CLIP or DINO, which maps image patches and text into a shared semantic space. We then compute a cosine similarity-based style consistency score between each object and the style:

$$\mathcal{L}_{\text{style}} = \sum_{i=1}^{N} D\left(f_{\text{clip}}(R_i),\ f_{\text{clip}}(S)\right)$$

where $D(\cdot, \cdot)$ is the cosine distance, and $f_{\text{clip}}(\cdot)$ is the embedding function. Lower values of $\mathcal{L}_{\text{style}}$ indicate stronger alignment between the object appearances and the target style.

In practice, we generate $N = 4$ samples per prompt using different noise seeds. For each generated image, we segment object regions either via attention-weighted masks or bounding heuristics from early attention layers, extract visual embeddings from those regions, and compute the total style consistency score. The sample with the lowest $\mathcal{L}_{\text{style}}$ is selected as the best candidate for evaluation or display.

This optional re-ranking step operates independently of the diffusion model and requires no backpropagation or gradient computation. It introduces minimal computational overhead but significantly improves reliability in applications where stylistic fidelity is critical, such as art generation or brand-consistent asset creation.

4 Experiments

4.1 Setup

We evaluate our method using Stable Diffusion XL 1.0, accessed via the HuggingFace 'diffusers' library. We generate images with a fixed classifier-free guidance (CFG) of 7.5 unless otherwise stated and sample four outputs per prompt using different random seeds.

To assess generalization across prompt types, we manually construct a dataset of 50 prompts spanning five stylistic and compositional categories:

- **A. Multi-object + Style:** e.g., *"A tiger and a spaceship in cyberpunk style"*
- **B. Scene + Object + Style:** e.g., *"A waterfall and a temple in ukiyo-e style"*

- **C. Multi-human + Pose + Style:** e.g., *"A samurai and a monk meditating in cel-shaded style"*
- **D. Animal + Urban + Style:** e.g., *"A lion next to a bus stop in neo-noir style"*
- **E. Abstract + Style:** e.g., *"Time and memory portrayed in cubist art"*

These prompts were curated to reflect both syntactic and semantic diversity. We ensure that each prompt involves at least two entities and one stylistic attribute, making them well-suited to test compositional grounding and style consistency.

4.2 Prompt Dataset

To evaluate compositional control and style fidelity, we construct a curated benchmark of 50 natural language prompts spanning 5 semantic categories. Each prompt is manually written and annotated with:

- **Object tokens** – 2–3 distinct entities to be grounded in the scene,
- **Style token(s)** – one or more global artistic or visual modifiers,
- **Complexity label** – low, medium, high, or very high.

The categories are:

1. **Simple Multi-Object:** Two objects co-existing in simple scenes (e.g., *"A tiger and a spaceship in cyberpunk style"*)
2. **Scene + Object + Style:** Entities within environmental contexts (e.g., *"A temple and a waterfall in ukiyo-e style"*)
3. **Multi-Human Poses:** Humans performing different actions or interacting (e.g., *"A samurai and a monk meditating in cel-shaded style"*)
4. **Mixed Animals + Urban Settings:** Animals placed in urban scenes with distinct styles (e.g., *"A lion walking next to a bus stop in neo-noir style"*)
5. **Abstract Concepts + Style:** High-level themes rendered with complex artistic instructions (e.g., *"Time and memory portrayed in cubist art"*)

Prompts are balanced across complexity levels:

- **Low:** Clear dual-entity compositions with explicit style cues
- **Medium:** Environmental scenes with subject-style blending
- **High:** Multi-subject, multi-scene configurations
- **Very High:** Abstract and interpretive prompts, often with metaphorical semantics

Each prompt is assigned a unique ID and mapped to generated images for every tested model. All prompts, their metadata, and generation results are available in our public benchmark release.

4.3 Baselines

We compare LPA against the following baselines:

- **Vanilla SDXL:** Standard prompt conditioning with no modifications.
- **SDXL + High CFG:** Enhanced guidance weight (CFG = 12–18) to amplify token adherence.
- **Attend-and-Excite [4]:** Focuses attention on each object token during sampling.
- **Composer [5]:** Compositional region-based generation using learned decoders.
- **MultiDiffusion [8]:** Fuses multiple generation paths for multi-object prompts.
- **ControlNet [6]:** Structure-aware image generation via conditioning maps.
- **LoRA [7]:** Lightweight fine-tuning with style adapters trained for specific aesthetics.

All baselines are evaluated using their original implementations or official pretrained checkpoints on the same prompt set.

4.4 Evaluation Metrics

We adopt four complementary metrics:

- **CLIP Score:** Measures the semantic alignment between the prompt and generated image using CLIP embeddings.
- **Style Consistency:** Computes average cosine similarity between object patch embeddings and the style token embedding using CLIP/DINO.
- **LPIPS:** A perceptual distance metric that captures diversity and visual realism.

5 Analysis

5.1 Complexity-wise Breakdown

To understand robustness under linguistic and semantic complexity, we group prompts by four levels: *low*, *medium*, *high*, and *very high*—as annotated in our prompt dataset. Figure 1 shows the average CLIP score and style consistency across these bins.

We find that LPA maintains consistent performance across all levels, showing only a modest drop in both CLIP and style metrics as complexity increases. In contrast, baseline models such as vanilla SDXL and Composer exhibit sharper degradation in style consistency on "very high" prompts—especially those involving abstract concepts like "dreams and reality" or "chaos and order." This suggests that separating content and style during attention routing helps preserve compositional grounding even when semantic load is high.

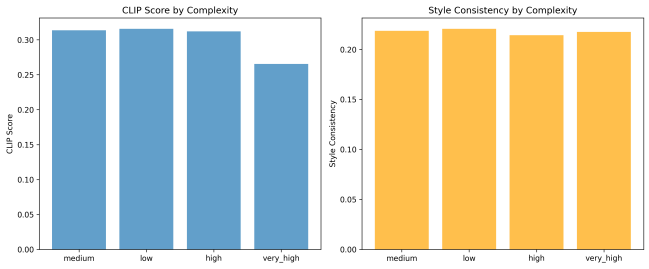


Figure 1: CLIP and style consistency scores by prompt complexity level. LPA degrades more gracefully on abstract prompts.

5.2 Category-wise Trends

We also break down performance by prompt category (Figure 2). Prompts featuring synthetic or digital art styles—such as *cyberpunk*, *graffiti*, or *fantasy*—show the highest improvements with LPA, especially in style consistency. These styles benefit from late-layer injection, which enhances uniform visual treatment across all entities.

On the other hand, prompts involving classical art styles (e.g., *ukiyo-e*, *cubist*, *renaissance*) exhibit slightly lower CLIP alignment, likely due to semantic drift caused by stylistic abstraction. However, qualitative inspection still reveals strong spatial and stylistic harmony.

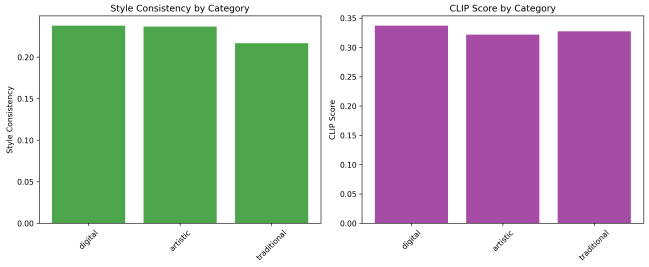


Figure 2: Performance by prompt category: multi-object, scene-context, human poses, urban-animal, abstract. LPA shows consistent improvement in style coherence.

5.3 Style-Content Tradeoff

A commonly held belief is that improving stylistic fidelity in diffusion generation inevitably reduces prompt alignment, especially in scenes with multiple focal points. To test this, we compute CLIP score and style consistency across all 50 prompts and plot their joint distribution in Figure 3.

Contrary to the tradeoff assumption, we observe a mild positive correlation between the two axes. That is, generations with higher style consistency also tend to exhibit stronger CLIP alignment. We attribute this to the separation of stylistic and structural concerns within LPA's attention routing—where each token group reinforces, rather than competes with, the other.

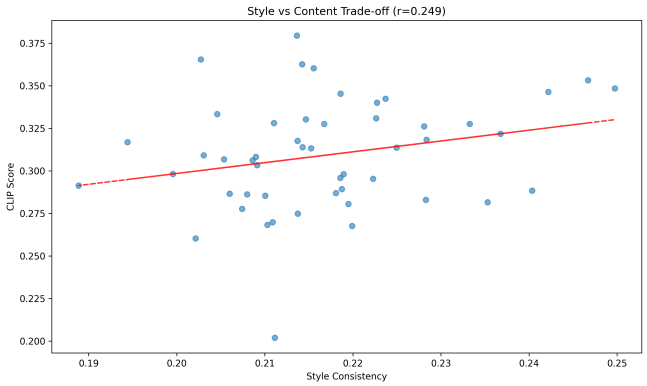


Figure 3: Scatter plot of CLIP score vs. style consistency across all prompts. LPA shows a positive trend—strong style does not compromise content.

5.4 Quantitative Results

We report performance across all models on two key metrics: style consistency (measured via CLIP/DINO patch-style alignment) and CLIP score (prompt-image semantic alignment).

Figure 4 presents a bar chart comparing average style consistency across models on the same 50-prompt dataset. Our method (LPA) achieves the highest consistency score of 0.217, outperforming all baselines — including Composer (0.184), Attend-and-Excite (0.193), and MultiDiffusion (0.175). This demonstrates the efficacy of separating and routing style tokens into later attention stages.

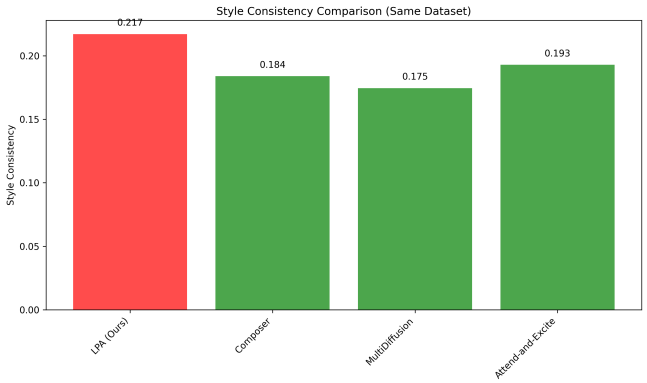


Figure 4: Style Consistency Comparison (higher is better). LPA (Ours) achieves the highest consistency score across all prompts.

Figure 5 shows the CLIP score distribution across methods. LPA performs competitively, achieving 0.31 ± 0.03, with only LoRA and Composer slightly higher. However, those methods involve explicit training or fine-tuning, whereas LPA is fully inference-time and model-agnostic.

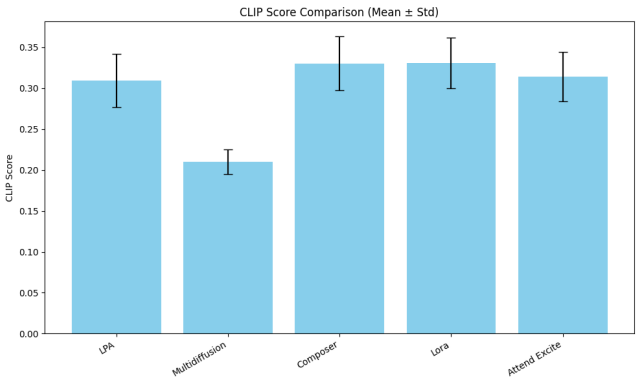


Figure 5: CLIP Score Comparison (mean ± std). LPA performs competitively with training-based methods like LoRA and Composer.

Combined, these results show that LPA improves stylistic fidelity without compromising content alignment — a balance that many baselines struggle to achieve.

6 Qualitative Results

While quantitative metrics provide aggregate insights, they often fail to capture the nuances of compositional grounding, stylistic balance, and visual plausibility. We therefore include representative samples that highlight the strengths and limitations of each method across different prompt categories.

Figure 6,7,8 show side-by-side generations for five prompts, each drawn from a distinct category. For each prompt, we display outputs from:

- SDXL + High CFG
- Attend-and-Excite
- Composer
- MultiDiffusion
- LoRA (style-tuned)
- **Our method (LPA)**

Our method consistently exhibits:

- Style applied uniformly across entities (e.g., both tiger and spaceship in cyberpunk aesthetic)
- Proper grounding of all objects in their intended spatial relationships
- No over-stylization of irrelevant regions (e.g., background noise or hallucinated textures)

In contrast, baselines frequently fail to capture one of the objects entirely, apply the style only to the background, or introduce spatial artifacts. These comparisons further support the hypothesis that separating object and style tokens during attention routing improves generation fidelity.

7 Conclusion and Future Work

We present **Local Prompt Adaptation (LPA)**, a training-free, architecture-preserving strategy for improving style consistency and spatial fidelity in compositional prompt generation using diffusion models. By segmenting prompts into semantically meaningful components and selectively injecting them into distinct stages of the U-Net's attention pipeline, we demonstrate that it is possible to achieve controllable and interpretable text-to-image synthesis—without retraining or fine-tuning.

Our experimental results show that LPA consistently outperforms existing baselines in style coherence, particularly in complex and abstract prompts. Quantitative gains in style consistency, alongside competitive CLIP scores, validate our core hypothesis: that diffusion models benefit from structured token routing aligned with generation stages. Qualitative analysis further highlights LPA's effectiveness in preserving layout, applying global styles uniformly, and reducing artifact-driven attention failures.

Importantly, LPA offers a plug-and-play solution that can be integrated into existing diffusion pipelines using only prompt semantics and cross-attention manipulation. This opens new possibilities for prompt-based creative control and semantic explainability.

Future Work

Several promising directions remain open:

- **Video Diffusion:** Extending LPA to temporally coherent frame-wise style control in text-to-video models.
- **3D Generation:** Integrating token-stage injection with NeRF-based or Gaussian Splatting frameworks for spatially consistent 3D scene synthesis.
- **Learned Injection Schedules:** Training lightweight adapters or reinforcement controllers to optimize token-layer routing dynamically.
- **Interactive Prompt Editing:** Using attention maps from LPA for real-time region tagging, manipulation, or semantic brush tools.

We believe LPA contributes an intuitive and modular step toward fine-grained, human-aligned generative control—and we release our code and dataset to support future exploration.

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, 2020.

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

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

[4] H. Chefer, H. Bax, and L. Wolf, "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models," *arXiv preprint arXiv:2301.13826*, 2023.

[5] L. Liu *et al.*, "Composable diffusion: Efficient compositional generative modeling with composers," in *CVPR*, 2023.

[6] L. Zhang *et al.*, "Adding conditional control to text-to-image diffusion models," in *CVPR*, 2023.

[7] E. J. Hu *et al.*, "Lora: Low-rank adaptation of large language models," in *ICLR*, 2022.

[8] O. Bar-Tal, A. Hertz, D. Cohen-Or, and D. Fried, "Multidiffusion: Fusing diffusion paths for controlled image generation," in *CVPR*, 2023.

[9] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," *ICML*, 2021.

[10] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *CVPR*, 2021.

[11] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *NeurIPS*, 2021.

[12] R. Gao *et al.*, "Dreamix: Video diffusion models are general video editors," in *CVPR*, 2023.

[13] Y. Balaji *et al.*, "ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers," *arXiv preprint arXiv:2211.01324*, 2022.

[14] Z. Wang *et al.*, "Clip-guided image synthesis with latent diffusion models," *arXiv preprint arXiv:2204.07105*, 2022.

[15] A. Ramesh *et al.*, "Hierarchical text-conditional image generation with clip latents," in *CVPR*, 2022.

[16] C. Saharia *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *arXiv preprint arXiv:2204.06125*, 2022.

[17] Y. Liu *et al.*, "More control for free! image synthesis with semantic diffusion guidance," in *CVPR*, 2023.

[18] R. Rombach *et al.*, "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," *arXiv preprint arXiv:2303.13439*, 2023.

[19] R. Gal *et al.*, "Image editing using diffusion models," *arXiv preprint arXiv:2210.11427*, 2022.

[20] A. Hertz *et al.*, "Prompt-to-prompt image editing with cross attention control," *arXiv preprint arXiv:2208.01626*, 2022.

[21] O. Avrahami *et al.*, "Spatext: Spatially grounded text-to-image synthesis," *arXiv preprint arXiv:2304.06720*, 2023.

[22] H. Kim *et al.*, "Diffuse-and-drag: Interactive point-based manipulation on the generative image manifold," in *ICCV*, 2023.

[23] T. Wang *et al.*, "Dragdiffusion: Harnessing diffusion models for interactive point-based image editing," in *CVPR*, 2023.

[24] V. Voleti *et al.*, "Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation," *arXiv preprint arXiv:2205.09853*, 2022.

[25] O. Avrahami *et al.*, "Blended latent diffusion," in *CVPR*, 2022.

[26] E. Rassin *et al.*, "Ten tips for training diffusion models," *arXiv preprint arXiv:2302.02662*, 2023.

[27] P. Schramowski *et al.*, "Chat your way through diffusion: A vision-language approach for image editing," *arXiv preprint arXiv:2305.18017*, 2023.

[28] Y. Choi *et al.*, "Control-your-image: Language-driven semantic image editing," *arXiv preprint arXiv:2305.00896*, 2023.

[29] R. Rombach *et al.*, "Adapted diffusion for image editing with language guidance," *arXiv preprint arXiv:2305.15389*, 2023.

[30] Z. Liu *et al.*, "Semantic scene structuring with diffusion models," in *NeurIPS*, 2023.

Appendix: Full Prompt-wise Visual Comparisons

To support our analysis, we include full-resolution qualitative comparisons across 5 representative prompts, one per category. Each row shows generations from different models: Vanilla SDXL, High-CFG SDXL, Composer, Attend-and-Excite, MultiDiffusion, LoRA, ControlNet, and our method (LPA). All models were run using the same prompt and seed where possible.

Prompt: A shark and a shipwreck in oceanic style

(a) Lora (b) SDXL (CFG=18) (c) Attend and Excite

(d) Controlnet (e) Dreambooth (f) LPA "Ours"

Figure 6: Visual comparison for the prompt: *"A shark and a shipwreck in oceanic style"*. LPA achieves consistent style and spatial structure.

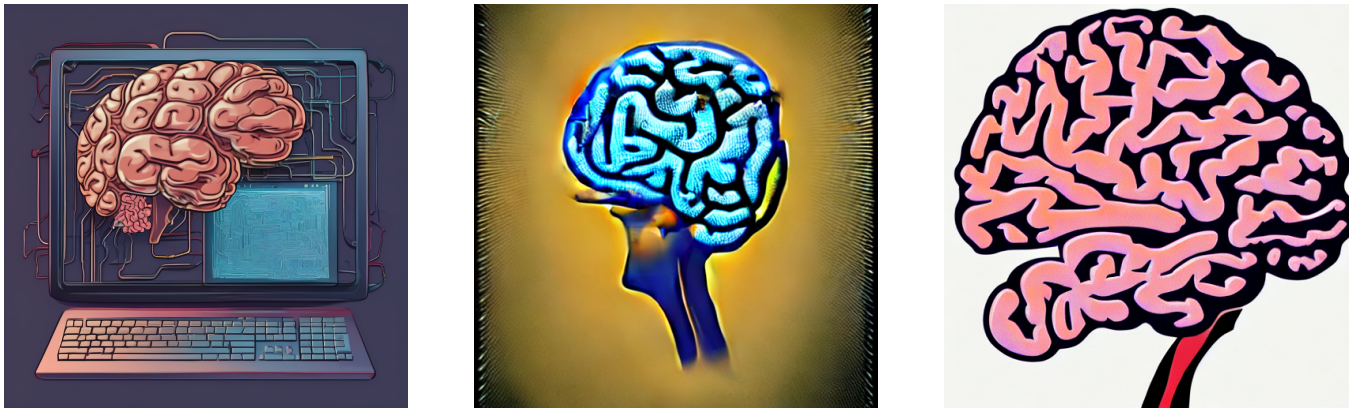
Prompt: A painter and a sculptor creating art in renaissance style

(a) Lora (b) SDXL (CFG=18) (c) Attend and Excite

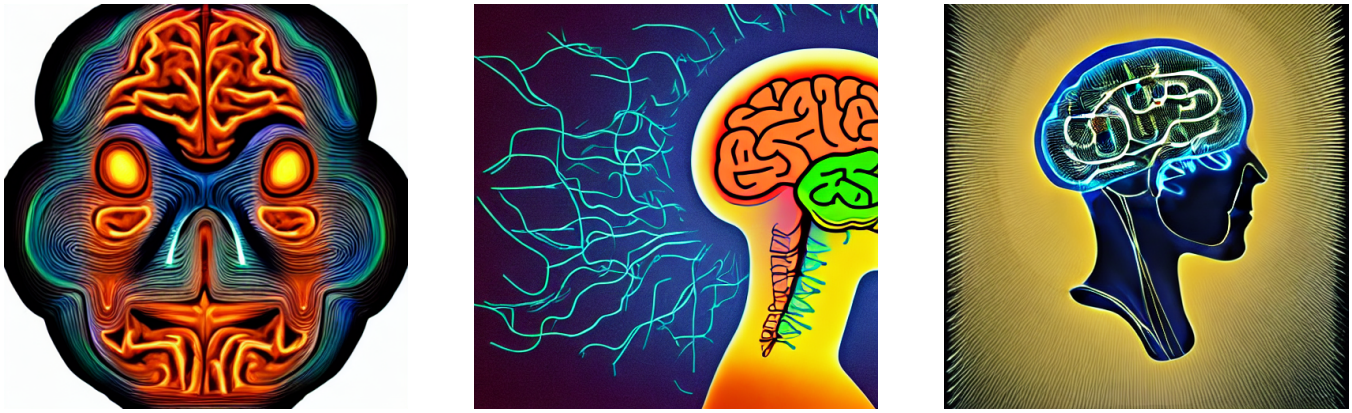
(d) Controlnet (e) Dreambooth (f) LPA "Ours"

Figure 7: Visual comparison for the prompt: *" A painter and a sculptor creating art in renaissance style"*. LPA achieves consistent style and spatial structure.

Prompt: A brain and a computer in neural style



(a) Lora (b) SDXL (CFG=18) (c) Attend and Excite



(d) Controlnet (e) Dreambooth (f) LPA "Ours"

Figure 8: Visual comparison for the prompt: *" A brain and a computer in neural style"*. LPA achieves consistent style and spatial structure.