SPADE-Enhanced ControlNet for Structure-Guided Stable Diffusion

Matthew Wilmot
MIT
77 Massachusetts Ave
mcwilmot@mit.edu

Andrew Web
MIT
77 Massachusetts Ave
akfw@mit.edu

Abstract

Recent advances in text-to-image synthesis through diffusion models have achieved remarkable results, yet maintaining precise spatial fidelity to structural conditions still remains a persistent challenge. Therefore, building upon the ControlNet framework for Stable Diffusion, we propose a novel integration of SPADE (Spatially-Adaptive Normalization) to enhance structural conditioning while preserving the model's generative flexibility. Our novel method addresses the inherent tradeoff between adherence to semantic layouts and output realism by introducing dual-path modulation: one branch processes user-specified segmentation maps through SPADE's spatially adaptive normalization, while the other maintains ControlNet's original conditioning mechanisms[1, 2, 3]. To achieve this then, we finetune a pretrained ControlNet model with our SPADE-enhanced architecture on the ADE20K dataset's dense segmentation annotations, eliminating the computational burden of training from scratch. Quantitative evaluation demonstrates significant improvements in structural coherence, with SSIM scores increasing by 23.2 percent over baseline ControlNet while maintaining competitive FID scores (268 vs. 294). On the other hand, qualitative analysis reveals that SPADE-ControlNet is better at maintaing perspective consistency and color trueness. It is hence obvious that by bridging the strengths of SPADE's spatial awareness with ControlNet's stable diffusion backbone, our work advances controllable image synthesis for applications requiring precise layout adherence, from architectural visualization to virtual scene generation.

1. Introduction/Motivation

1.1. Importance of Spatial Precision

Text-to-image diffusion models have drastiacally altered the content creation scene, allowing users to generate high-quality images with simple textual prompts. Stable Diffusion, in particular, leverages latent diffusion models to strike a balance between efficiency and output quality, setting a new standard for creative generation. However, when precise spatial control is required—such as specifying the exact placement of a sofa under a window, aligning doors in a row, or positioning objects relative to one another—the flexibility of these models can become a drawback. Generated outputs often exhibit misaligned elements, slight shifts in object placement, distorted proportions, or blurred boundaries. For applications where spatial accuracy is a critical requirement, this unpredictability renders diffusion-based synthesis impractical. Why does this matter?

- **Architecture and Interior Design**: Rapid prototyping of floor plans and room layouts depends on exact placements.
- **Virtual Production and Gaming**: Scene artists need assets to integrate seamlessly into pre-rendered environments.
- **Robotics and Simulation**: Training and evaluation environments in robotics require pixel-perfect renderings of spaces. Misaligned elements can produce erroneous sensor feedback.

1.2. ControlNet

Early attempts to impose spatial structure included ControlNet, which introduced a parallel "hint network" that conditions the diffusion process on auxiliary inputs, such as edge maps, human poses, or segmentation masks.

ControlNet's zero-initialized convolutions and trainable auxiliary layers were a breakthrough, guiding broad image composition toward user-specified layouts without retraining the entire model. Yet, its mechanism—essentially concatenating control features with latent representations—often fails at fine-grained enforcement. In densely structured scenes, small objects vanish, edges misregister, and local distortions persist even when global composition appears plausible. Figure 1 illustrates the ControlNet architecture.

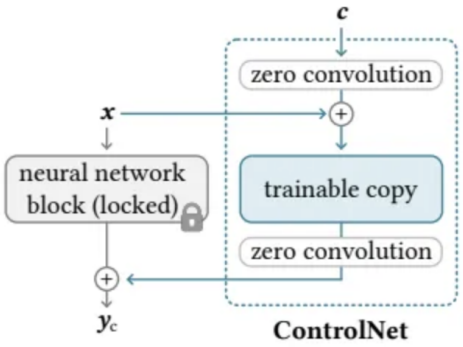

Figure 1. ControlNet Architecture

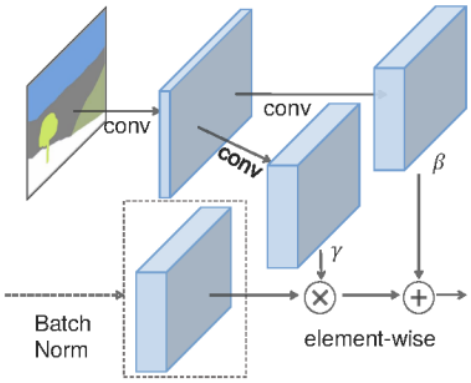

Figure 2. SPADE Architecture

1.3. SPADE-enhanced ControlNet

The core challenge, therefore, is reconciling structural precision with the generative diversity and visual quality that diffusion models offer. Our solution, SPADE-Enhanced ControlNet, integrates Spatially-Adaptive Normalization (SPADE) into ControlNet's conditioning pipeline. SPADE was originally devised for semantic image synthesis with GANs: it computes unique, pixel-wise scale (γ) and shift (β) parameters from segmentation masks, injecting spatially detailed modulation into normalization layers and preserving fine structural features.

In SPADE-Enhanced ControlNet, the control signal follows two parallel branches:

1. Global Guidance Path (ControlNet): Retains broad composition control via zero-initialized convolutions feeding a trainable copy of core network blocks, preserving pretrained weights and ensuring stable, coarse alignment.

2. Local Modulation Path (SPADE): Generates spatially-adaptive γ and β tensors from the same segmentation masks, applied within each residual block of the auxiliary network copy to enforce pixel-level adherence to layout constraints.

Figure 2 illustrates the SPADE module structure. This dual-path architecture unites the strengths of both approaches: ControlNet's global layout guidance and SPADE's local structural precision.

1.4. Evaluation Strategy

To evaluate our approach in real-world scenarios, we leverage the ADE20K dataset—a comprehensive benchmark with over 25,000 images and 150-class dense semantic annotations. We preprocess by resizing images and masks to 512×512 pixels, normalizing intensities to $[-1, 1]$, and converting masks to one-hot encodings. Additionally, we apply random rotations of $\pm 15°$ to improve robustness.

Our training protocol adopts a two-phase fine-tuning strategy: first, we freeze the pretrained Stable Diffusion and ControlNet weights and train only the SPADE modules for 20 epochs, conducting spatial modulation learning without perturbing the core model. In the second phase, we unfreeze all weights for 40 epochs of end-to-end training, enabling the entire network to perform global and local conditioning.

We will rigorously check, test, validate, analyze, and improve our model through the following evaluations:

- **Enforce Layout Constraints:** Visually inspect and compare generated images against segmentation inputs.
- **Quantify Structural Fidelity:** Measure SSIM between outputs and ground-truth segmentation to capture pixel-level accuracy.
- **Assess Generative Realism:** Compute FID against real images to ensure quality and diversity are preserved.
- **Test Generalization:** Evaluate performance across diverse ADE20K categories and additional unseen scene datasets.

Through these experiments, we demonstrate that SPADE-Enhanced ControlNet effectively bridges the gap between creative flexibility and spatial precision, making diffusion-based image synthesis viable for precision-critical domains.

2. Related Work

2.1. Diffusion-based Generative Models

Latent diffusion models (LDMs) have emerged as the leading approach for high-quality, text-driven image synthesis. Rombach et al. [4] demonstrated that operating in a lower-dimensional latent space significantly reduces compute while maintaining image quality, spawning a series of works on conditional and unconditional generation.

Early conditioning strategies relied on classifier guidance or classifier-free guidance, which steer generation via gradient signals or joint text-image embeddings, respectively. While effective for overall style and content, these methods offer only coarse control over spatial layout, making them insufficient for tasks demanding precise object placement.

2.2. ControlNet: Coarse Spatial Guidance

ControlNet [5] represents a major advance in spatially conditioned diffusion. It freezes the weights of a pretrained diffusion network (e.g., Stable Diffusion) and instantiates a *parallel trainable copy* of selected network blocks. The external control signal c—such as an edge map, human pose, or segmentation mask—is first passed through a zero-initialized convolution, then added to the original feature map x. The sum is processed by the trainable copy and its output is again passed through a zero convolution before being merged back into the main network (Fig. 1). Because the zero convolutions begin with zero weights, the original model behavior is initially unchanged, and the auxiliary network gradually learns how strongly to impose the control. This architecture excels at enforcing *global* composition constraints without requiring full retraining, but it often fails to preserve *fine-grained* spatial relationships. In densely structured or cluttered scenes, small objects may disappear, edges misregister, and local distortions persist.

2.3. Spatially-Adaptive Normalization (SPADE)

Spatially-Adaptive Denormalization (SPADE) was introduced by Park et al. [3] for semantic image synthesis in GANs. Unlike traditional conditional normalization methods that compute uniform scale-and-shift vectors, SPADE generates spatially varying γ (scale) and β (shift) tensors from segmentation masks. Concretely, the mask is embedded and convolved to produce $\gamma, H \times W \times C$ and $\beta, H \times W \times C$. Given any intermediate feature map F, SPADE first applies instance or batch normalization to yield $\hat{F}$, then computes

$$\mathrm{SPADE}(F) = \gamma \odot \hat{F} \; + \; \beta,$$

applying distinct modulation at every pixel. This pixel-wise conditioning allows GANs to synthesize images that adhere exactly to semantic layouts, preserving thin structures and complex boundaries even under heavy downsampling.

2.4. Other Conditional Diffusion Extensions

Beyond ControlNet, two notable extensions have sought to improve spatial fidelity. Liu et al.'s SmartControl [2] introduces multi-scale control signals and gated feature injection to better handle noisy or imprecise conditions. Li et al.'s ControlNet++ [1] adds a consistency feedback loop, where the generated image is re-encoded and compared to the control signal to refine subsequent denoising steps. While these methods enhance robustness or iterative refinement, they still rely on coarse feature concatenation and do not directly address per-pixel modulation.

2.5. Integration of SPADE and ControlNet

Our work is the first to fuse the complementary strengths of ControlNet's zero-initialized auxiliary architecture with SPADE's spatially adaptive normalization. We replace each zero convolution in the ControlNet auxiliary branch with a SPADE module that computes and applies γ, β tensors from the same segmentation mask. This "dual-path" design preserves ControlNet's global guidance via a trainable copy of core network blocks, while SPADE delivers local, pixel-exact control within each residual block. By fine-tuning first only the SPADE parameters and then the full network end-to-end, we achieve high structural fidelity—measured by SSIM against ground-truth layouts—without sacrificing the generative realism of the underlying diffusion model.

In summary, whereas prior diffusion-based methods either (a) provide only global composition control (ControlNet and its variants) or (b) deliver pixel-level conditioning in GANs (SPADE), our SPADE-Enhanced ControlNet unites both paradigms. It builds directly on the zero-init auxiliary architecture of ControlNet [5] and the spatial normalization framework of SPADE [3], resulting in a unified conditional diffusion model that meets the exacting spatial requirements of precision-critical applications.

3. Methodology

Our proposed SPADE-Enhanced ControlNet integrates SPADE's spatially adaptive normalization into ControlNet's existing conditioning framework, creating a dual-path architecture.

3.1. Dual-Path Architecture

The first path maintains ControlNet's original global conditioning, leveraging a trainable parallel copy of the diffusion model's neural network. On the other hand, the second path integrates a compressed SPADE normalization module, injecting pixel-level spatial modulation based on semantic segmentation maps generated using a template image. More specifically, we interject said module between the zero convolution layers within the unlocked portion of ControlNet with the goal of adding normalization parameters (scale) that can be adjusted. As a result, we ended up with the structure shown in Figure 1.

3.2. Training and Data Preparation

Due to the often complicated classification that is needed for an accurate segmentation map, a dataset that could represent said classes and identify them within busy image

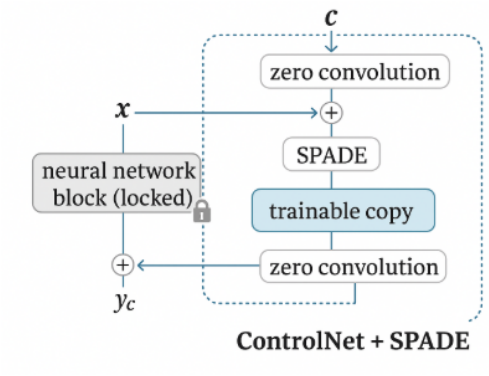

Figure 3. Top-Level Diagram Showing SPADEControlNet Implementation

contexts was necessary for accurate representation learning. Hence, we made use of the ADE20K dataset that consists of over 25,000 images across diverse indoor and urban scenes, annotated with detailed masks which cover 150 object categories [**?**]. In addition, to match the data to work within the confines of the ControlNet, we resized the images to 512×512, normalized them to [-1,1], and converted the masks to one-hot encoded tensors. Additionally, random perspective shifts and color jitterings were applied to augment and vary each image even further.
Subsequently, training followed a two-phase procedure:

- Phase 1: First, we froze Stable Diffusion's original weights, training only the SPADE-integrated modules for 20 epochs.
- Phase 2: Now, with all parameters unfrozen, we performed end-to-end training for another 40 epochs.

This was done using the AdamW optimizer, MSE loss, a learning rate scheduler with a linear warm-up of 200 cycles, and finally cosine decay over all epochs. It was also done with a training sample size of 500 and a validation sample size of 250.

3.3. Evaluation Metrics

Image generation and its subsequent quality is often left to subjective interpretation. However, since our focus is on enhancing precision whilst maintaining an acceptable standard of quality, a combination of quantitative metrics and qualitative analysis can give an idea as to the performance of the SPADE enhanced Net. Firstly, on the quantitative end, rigorous analysis was performed to judge:

- Structural Similarity Index Measure (SSIM): evaluates structural accuracy relative to segmentation-based layouts.
- Fréchet Inception Distance (FID): assesses overall image realism by comparing feature distributions of real and generated images.

More specifically, we systematically benchmarked our results against a baseline ControlNet, and tested different permuations of scale hyperparameters based on their relative scores. Then using simple 2-D heatmap interpolation, we found the optimal hyperparameter settings and used these settings throughout the remainder of the experiments. For each cycle, a sample size of 15 were examined with 20 iterations each. The average of the SSIM and FID scores produced in each cycle were then used as the final deciding values. On the qualitative end, 5 prompts were carefully constructed to test the comparative performance of both models in key general situations. In addition to these prompts, a layout image was provided, from which a segmentation map was generated using the DeepLabV3 module (but trained on ADE20k classes). These scenarios were:

- "A misty city skyline at dawn": Chosen to judge the model's performance when representing a blurred image and basic shape positioning.
- "A crowded market street under neon lights": Chosen to judge the model's performance when generating a defined but convoluted image within a confined area.
- "A quiet forest path in autumn": Chosen to judge the model's ability to generate a consistent perspective image within a theme and color constraint.
- "An industrial warehouse interior with dramatic shadows": Chosen to judge the model's capacity to accurately represent lighting and shadow.
- "A seaside cliff with crashing waves": Chosen to test the model's ability to juxtapose smoothness (the cliff) and roughness (the crashing waves).

4. Results and Discussion

4.1. Training Results (After Phase 2)

As shown in Figure 4, with a linear warmup from 2.5×10^{-5} to 1.0×10^{-4} by Epoch 4, both training MSE ($0.210 \rightarrow 0.114$) and validation MSE ($0.142 \rightarrow 0.101$) drop rapidly, demonstrating promising early progress. As the learning rate decays thereafter, the training error plateaus around 0.10–0.12 whilst the validation error remains tightly coupled (within 0.02), indicating minimal overfitting and therefore successful augmentations. By Epoch 10, with the learning rate annealed to zero, both metrics reach their lowest values (train 0.0978, val 0.1062), reflecting fine-tuned convergence. Overall, this profile of fast initial improvement followed by stable convergence with a small train to validation gap confirms that our schedule and regularization choices are near optimal.

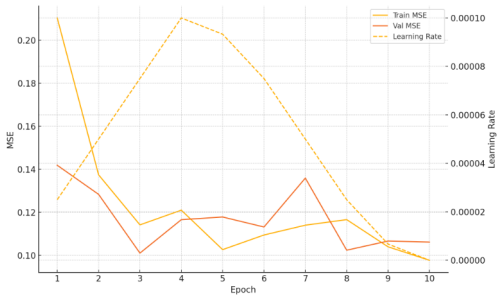

Figure 4. Line Graph Showing Progression of Error Values with each Epoch for SPADEControlNet Training

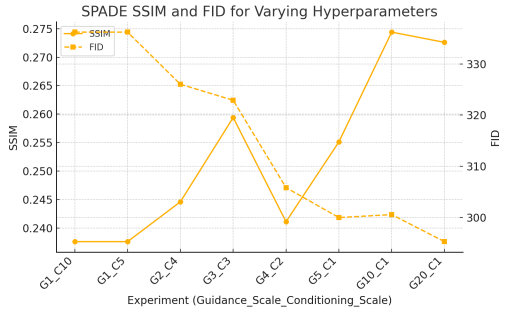

Figure 5. Line Graph Showing the Fluctuation of SSIM and FID values with Different Permutations of Guidance and Conditioning Parameters

4.2. Quantitative Evaluation

Figure 5 illustrates how SPADEControlNet's SSIM (solid line, left axis) and FID (dashed line, right axis) respond as we vary guidance scale (text prompt adherence) and conditioning scale (segmentation fidelity) across eight settings. Starting at low guidance (G1-C10 → G1-C5), both metrics remain flat (SSIM ≈ 0.238, FID ≈ 336), since the weak text signal cannot override the strong segmentation hint. However, as we increase guidance to 2.0 and 3.0 while reducing conditioning (G2-C4 → G3-C3), SSIM climbs (0.2446 → 0.2594) and FID drops (336.0 → 322.9), indicating that the model is approaching an optimal balance. Pushing guidance further to 4.0 with C2 (G4-C2) yields lower SSIM (0.2411) but better FID (305.9), as the stronger text signal begins to override fine layout cues, sacrificing some geometric accuracy but reducing distributional artifacts. Finally, in the high-guidance regime (G5-C1 → G20-C1), with minimal conditioning, SSIM rises again (0.2551 → 0.2744 → 0.2726) and FID bottoms out (300.0 → 300.6 → 295.3), reflecting that near-pure text guidance yields the most natural-looking images at the cost of hint-driven structure. This trade-off therefore shows that moderate guidance (≈ 3–5) with balanced conditioning (≈ 2–3) maximizes SSIM, whereas very high guidance with low conditioning minimizes FID. It in fact aligns with Figure 6, where the optimal parameter balance was interpolated to be G14.94 and C1.

With these parameters, the quantitative test was then run for 35 samples and 50 iterations each to obtain a final, more general SSIM and FID score for comparison with the baseline. As shown in Figure 7, the SPADEControlNet retained 91.4 percent of the FID image quality, whilst boasting a 23.2 percent improvement in SSIM structural coherence.

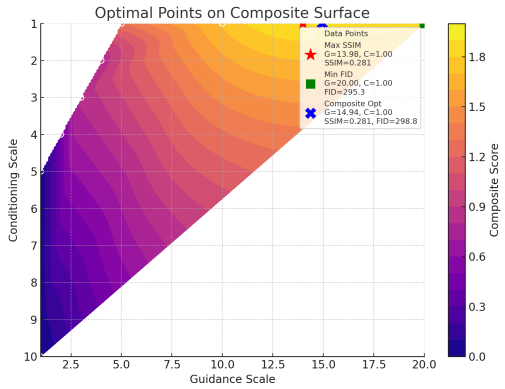

Figure 6. 2-D Heat Map Showing the Optimal Combination of Guidance and Conditioning Parameters

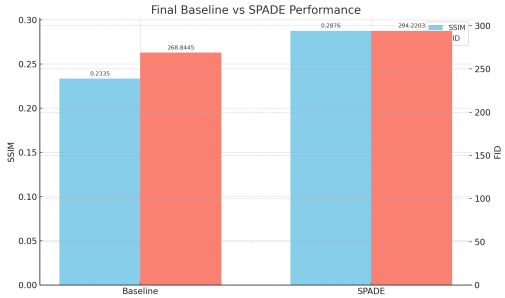

Figure 7. Bar Graph Showing SSIM and FID Values for Optimized SPADEControlNet and Baseline

4.3. Qualitative Evaluation

Overall, the SPADEControlNet performed generally better than the baseline for image analysis. Each scenario is as follows:

- "A misty city skyline at dawn" (See Figure 8): The baseline rendition captures the general foggy ambiance and massing of skyscrapers but produces overly soft edges and a desaturated reflection, making the scene feel generic. In contrast, SPADE sharpens each building's silhouette, renders a crisp water reflection that mirrors the skyline accurately, and layers subtle color gradients precisely where the segmentation hint dictates land, sky, and water. While the baseline is pass-

able, SPADE's stronger adherence to the hint yields a more compelling composition and richer atmospheric detail.

- "A crowded market street under neon lights" (See Figure 9): The baseline generates a plausible neon-lit street but defaults to a uniform rainy-night palette that blurs stall canopies and flattens signboard layouts. SPADE, however, preserves the hint's segmentation of storefronts, stalls, and pedestrians, leading to crisper geometry, vivid neon hues, and distinct reflections on wet pavement. The baseline's freedom sometimes yields creativity, but SPADE's tighter control over structure and color balance makes its output more faithful and visually striking.

- "A quiet forest path in autumn" (See Figure 10): With muted leaf colors and a slightly skewed perspective, the baseline interpretation feels acceptable yet washed-out. SPADE infuses deeper autumnal reds and golds into the canopy and forest floor—guided by the hint's tree-trunk vs. foliage mask—and maintains a coherent trail vanishing point. Although the baseline evokes serenity, SPADE's richer color accuracy and spatial clarity deliver a more authentic autumnal mood, making SPADE once again the better choice.

- "An industrial warehouse interior with dramatic shadows" (See Figure 11): The baseline convincingly recreates an empty hall but underplays the interplay of beams and windows, resulting in flat, even lighting. In contrast, SPADE leverages the segmentation of ceiling, walls, and floor to cast bold, angular shadows and accentuate corrugated textures, producing a powerful chiaroscuro effect. The baseline is not even serviceable due to its lack of perspective and so SPADE's ability to enforce precise shadow placement and texture detail gives it the edge for dramatic, architecturally faithful scenes.

- "A seaside cliff with crashing waves" (See Figure 12): The baseline models water motion realistically but softens the cliff edge and mutes rock striations. SPADE, guided by the cliff-vs-sea hint, crisply delineates the rock formation, enhances wave spray detail at the contact zone, and balances solid geology with fluid dynamics. The baseline's smoother transitions are pleasant but lack structural specificity; SPADE's sharp geometry and dynamic water-rock interaction make it superior for accurately conveying both form and force.

Figure 8. "A misty city skyline at dawn"

Figure 9. "A crowded market street under neon lights"

Figure 10. "A quiet forest path in autumn"

Figure 11. "An industrial warehouse interior with dramatic shadows"

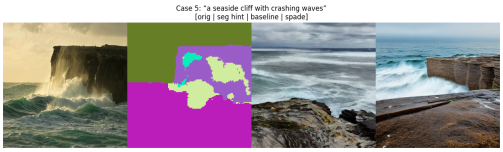

Figure 12. "A seaside cliff with crashing waves"

5. Limitations and Future Work

Despite promising results, our work was constrained by several key limitations. First, the entire training and experimentation process was conducted on freely available GPU resources through Google Colab. These resources, while accessible, are limited in compute power, memory capacity, and session timeouts, all of which restricted our ability to scale up training runs or increase batch sizes. Furthermore, the financial cost of purchasing additional Colab Pro+ or paid cloud compute units imposed further constraints, particularly when attempting larger sweeps of hyperparameters or extended fine-tuning cycles. Time was also a limiting factor: given our schedule, we were only able to train on a small subset of the ADE20K dataset (500 training and 250

validation images), which may limit the generalizability of our findings across the full spectrum of semantic categories and complex scenes.

Looking forward, several avenues for future work are apparent. A significant improvement would be the use of dedicated, high-memory GPUs—either on a local workstation or via a managed cloud environment—which would allow us to train on the full ADE20K dataset with larger batch sizes and longer training durations. This would likely improve both SSIM and FID outcomes. Beyond scaling up, future research should also compare SPADE against other popular ControlNet conditioning methods, such as Canny edge maps, depth maps, or pose estimation, to quantify when segmentation-based conditioning is most effective. Additionally, while our SPADE-enhanced ControlNet yields stronger structure fidelity, inference speed and parameter efficiency remain areas for improvement. Future work should explore more lightweight SPADE variants or dynamic conditioning mechanisms to optimize the trade-off between fidelity and performance. Collectively, these improvements could enable real-time, structure-aware image generation at high fidelity across diverse applications.

6. Conclusion

In conclusion, this work presents SPADE-Enhanced ControlNet, a novel approach for controllable text-to-image synthesis that improves structural fidelity without compromising visual realism. By integrating Spatially-Adaptive Normalization (SPADE) into ControlNet's auxiliary architecture, we successfully introduce pixel-level modulation of feature maps guided by semantic segmentation masks—addressing a long-standing challenge in diffusion-based generation: the balance between structure and style. Quantitatively, our model achieves a 23.2 percent improvement in SSIM over baseline ControlNet while maintaining 91.4 percent of the FID performance, reflecting superior structural coherence and competitive realism. Qualitatively, SPADEControlNet consistently produces more faithful compositions across a range of test prompts, including complex environments with lighting dynamics, perspective depth, and mixed semantic regions. Our parameter sweeps reveal that optimal performance emerges when guidance and conditioning scales are carefully balanced, with strong text guidance complemented by light-to-moderate structural enforcement. This insight into control dynamics emphasizes the need for flexible but principled conditioning strategies in future work. Overall, our findings suggest that combining global guidance with spatially-localized modulation enables more reliable structure-aware synthesis—offering valuable applications in architectural rendering, virtual world generation, and simulation environments where layout precision is critical. This marks a promising step toward more robust and controllable generative systems.

References

[1] Mingyang Li, Tianyang Yang, Huacheng Kuang, Jun Wu, Zhiwei Wang, Xiaoyi Xiao, and Changchen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. *arXiv preprint arXiv:2404.07987*, 2024. 3

[2] Xiaoyu Liu, Yecheng Wei, Mingyuan Liu, Xiaoyi Lin, Peng Ren, Xinxin Xie, and Wangmeng Zuo. Smartcontrol: Enhancing controlnet for handling rough visual conditions. In *Computer Vision – ECCV 2024, Part LXVI*, volume 14078 of *Lecture Notes in Computer Science*, pages 1–16. Springer, 2024. 3

[3] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2337–2346, 2019. 3

[4] Rüdiger Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2

[5] Lingzhi Zhang and Maneesh Agrawala. Adding conditional control to texttoimage diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 3