FreeControl: Efficient, Training-Free Structural Control via One-Step Attention Extraction

Jiang Lin[1,†], **Xinyu Chen**[1], **Song Wu**[2], **Zhiqiu Zhang**[1], **Jizhi Zhang**[1], **Ye Wang**[4], **Qiang Tang**[3], **Qian Wang**[2], **Jian Yang**[1], **Zili Yi**[1*]

[1]Nanjing University, Suzhou, China
[2] JIUTIAN Research, Beijing, China
[3]University of British Columbia, Vancouver, Canada
[4]Jilin University, Changchun, China

lin@smail.nju.edu.cn, yi@nju.edu.cn

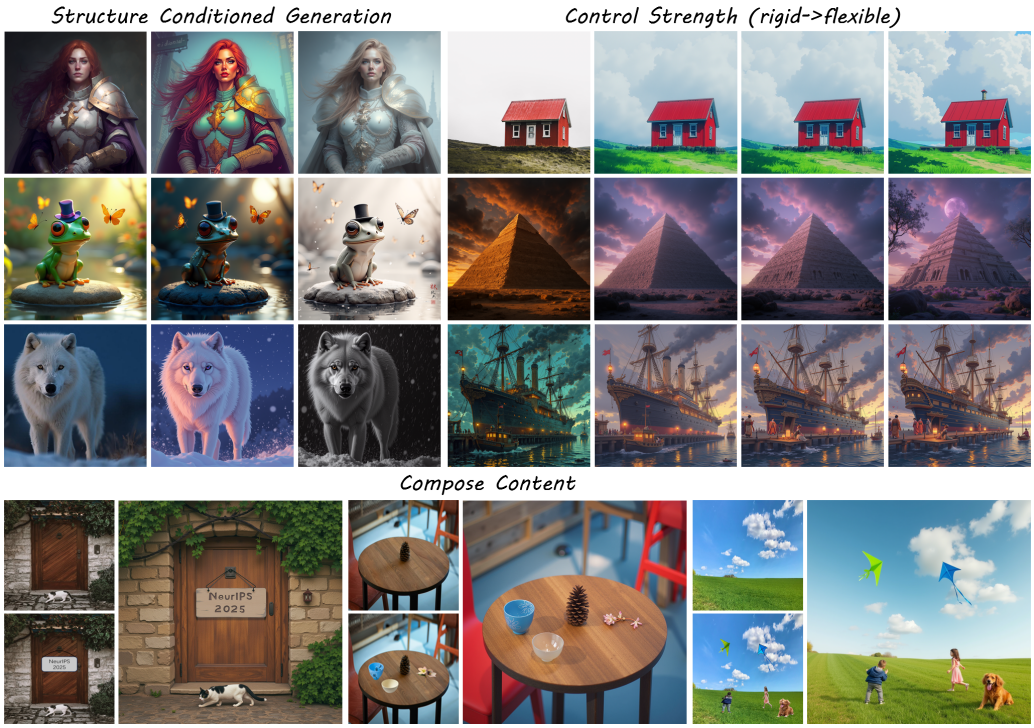

Figure 1: FreeControl enables efficient, structure-aware generation from raw image references. Top-left: structure-conditioned generation using reference image on the left. Top-right: Tunable Control strength via adjustable attention injection. Bottom: compositional generation from user-assembled reference images enables intuitive spatial and semantic layout control.

Abstract

Controlling the spatial and semantic structure of diffusion-generated images remains a challenge. Existing methods like ControlNet rely on handcrafted condition maps and retraining, limiting flexibility and generalization. Inversion-based approaches offer stronger alignment but incur high inference cost due to dual-path denoising. We present **FreeControl**, a training-free framework for semantic struc-

*Corresponding authors: yi@nju.edu.cn; †: project lead

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

tural control in diffusion models. Unlike prior methods that extract attention across multiple timesteps, FreeControl performs *one-step attention extraction* from a single, optimally chosen key timestep and reuses it throughout denoising. This enables efficient structural guidance without inversion or retraining. To further improve quality and stability, we introduce *Latent-Condition Decoupling (LCD)*: a principled separation of the key timestep and the noised latent used in attention extraction. LCD provides finer control over attention quality and eliminates structural artifacts. FreeControl also supports compositional control via reference images assembled from multiple sources, enabling intuitive scene layout design and stronger prompt alignment. FreeControl introduces a new paradigm for test-time control—enabling structurally and semantically aligned, visually coherent generation directly from raw images, with the flexibility for intuitive compositional design and compatibility with modern diffusion models at ~5% additional cost.

1 Introduction

While diffusion models [8, 33, 36, 32, 5, 28, 2] have revolutionized generative image synthesis, they remain difficult to control. There often lacks intuitive ways to specify what appears where, or how objects should relate spatially and semantically—making them less suitable for tasks like scene layout, object rearrangement, or design prototyping.

A prevalent approach to this challenge involves conditioning generation on external control maps, as exemplified by ControlNet [47] and T2I-Adapter [24]. These methods inject spatial guidance via edge maps, depth cues, or segmentation masks. While effective, they depend on handcrafted pre-processing and require separate training for each control type and base model. In particular, ControlNet demands large-scale paired datasets and substantial training resources per condition, making it expensive to scale across modalities or architectures. Moreover, the control signals themselves are limited: Canny edges are often overly rigid and may conflict with prompt semantics, while segmentation-based guidance is restricted by limited category labels, preventing nuanced or open-ended structure control. In contrast, test-time augmented methods [38, 23, 19, 40] such as DDIM inversion [38] and TRAC [19] extract structure from reference images by reconstructing their latent trajectory and injecting features throughout denoising. These techniques incur high inference cost, requiring full or dual-path denoising and considerable memory.

We propose a training-free, test-time augmented framework for semantic structural control using raw reference images. Our method performs a single additional denoising step at a model-specific key timestep, chosen to extract maximally informative self-attention. This attention matrix captures both spatial structure and semantic intent, and is consistently injected into the main generation process to guide the arrangement and content of the generated image. The strength and scope of guidance are tunable, enabling both flexible layout guidance and strong structural adherence, depending on user intent. Unlike prior test-time augmentation methods [7, 4], our approach eliminates the need for inversion or reconstruction entirely—removing both the computational burden and architectural complexity of dual-path denoising. With only 5% additional cost over baseline inference, it delivers high-quality structural control without retraining, making it directly compatible with fine-tuned [35] or LoRA-augmented models [10].

By collapsing multi-step extraction into a single attention signal, our one-step approach creates a tractable point of analysis—enabling us to systematically study and refine the quality of structural guidance through Latent-Condition Decoupling (LCD). LCD separates the roles of the noised latent and the key timestep, revealing how each factor shapes the extracted structure. This lets us improve alignment quality, reduce artifacts, and offer tunable control over structural granularity—from coarse layout to fine semantic detail.

To support intuitive, layout-aware control beyond segmentation maps or prompt tuning, we introduce a composition-based conditioning strategy. As shown in fig. 1, users can directly assemble reference images by cropping and combining objects from different sources, enabling them to express both spatial layout and semantic intent in a natural visual form, and generate images with content that aligns with their expectations. This flexibility transforms structural and semantic conditioning into a designable interface for high-level scene control.

In experiments, FreeControl outperforms existing structural control methods in both spatial alignment and visual fidelity, while maintaining high efficiency. Qualitative results further demonstrate its advantage in semantic-level control, producing generations that more faithfully adhere to the intended prompts.

Our contributions are as follows:

- We present a training-free, test-time augmented method for semantic structural control from raw reference images, eliminating the need for handcrafted inputs, inversion, or retraining.
- We propose a one-step attention extraction framework that uses a single denoising step at a key timestep to guide generation, with attention maps injected across layers during inference.
- We introduce Latent-Condition Decoupling (LCD), a principled method that separates the key timestep from the noised latent in attention extraction, enabling stronger control and improved stability.
- We introduce a composition-based conditioning approach that allows users to define both spatial layout and semantic intent through assembled reference images, enabling intuitive control beyond segmentation maps or prompt tuning.

2 Related Work

Diffusion Models. Diffusion models [8, 38, 5, 28] have emerged as a leading framework for high-quality image synthesis, with success across tasks such as text-to-image generation [33, 36, 32], image-to-image translation [12, 22, 15], and image editing [7, 3, 45, 13, 37]. Foundational models like DDPM [8] and DDIM [38] introduced the core denoising process, while later developments such as LDM [33], DiT [27], and SD3 [6] have scaled diffusion to high-resolution, semantically rich generation. The field has also moved from U-Net-based backbones [34] to more expressive transformer-based architectures [41].

Structural Guidance via Training-Based Conditioning. Training-based methods such as ControlNet [47] and T2I-Adapter [24] guide spatial structure using condition maps like edges or segmentation. While they achieve strong low-level alignment, they require retraining for each control type and base model—introducing high computational cost and model proliferation. Their reliance on handcrafted inputs also leads to brittle performance when structure maps are noisy or conflict with text prompts. T2I-Adapter is lighter but similarly struggles with complex scenes and still demands per-condition training. High-level alternatives like GLIGEN [18] and IP-Adapter [46] support layout-aware and visual conditioning using bounding boxes or global image features. However, IP-Adapter still needs base-model-specific training and shows varied quality depending on the dataset. Though these methods improve compositional flexibility, they lack fine-grained structural control—e.g., for object contours or pose—limiting their utility in dense structure transfer tasks.

Test-Time Control via Inversion and Attention Reuse. Test-time approaches offer another path to structure control. DDIM inversion [38, 40, 21] and Null-Text Inversion [23] reconstruct noise latents from reference images, enabling attention reuse for editing. Though effective, they are computationally heavy and depend on prompt alignment. Prompt-to-Prompt [7] modifies cross-attention for semantic edits while preserving layout but cannot incorporate visual references.

Plug-and-Play [40] injects image features during inference but offers coarse structure control. It requires dual attention modulation and ResNet backbones, limiting efficiency and compatibility with transformer-based models like DiTs [27]. TRAC [19] improves efficiency by avoiding inversion, but still extracts attention at many timesteps, incurring high cost.

Crucially, both inversion-based and inversion-free methods operate under the same assumption: that structure must be extracted progressively across a denoising trajectory. Yet, across timesteps, the role of attention remains consistent—capturing spatial layout and semantic structure. This raises a fundamental question: if attention serves the same purpose at every step, is repeated extraction truly necessary?

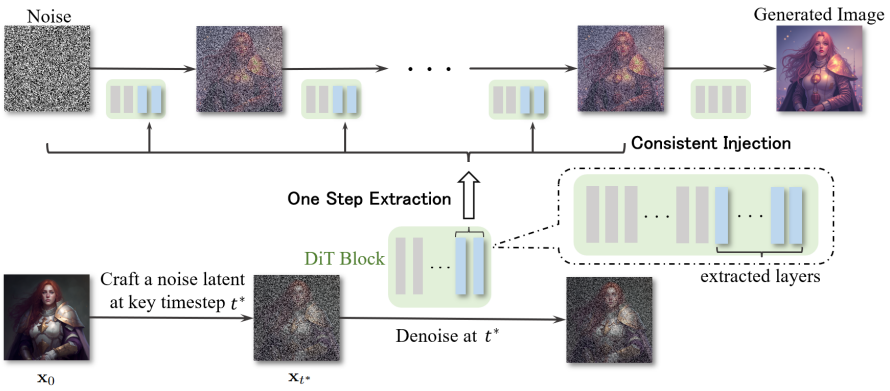

Figure 2: The illustration of one-step attention extraction framework. The query attention matrices in the later layers (blue layers) are extracted from a forward-simulated latent at a single key timestep and are injected consistently in the generation process to enable structural guidance.

3 Methods

3.1 Semantic Structural Control via One-Step Extraction

One-Step Attention Extraction Motivated by the insight that structural information remains conceptually consistent across timesteps, we introduce a *one-step attention extraction strategy* to replace multi-step guidance. As demonstrated in fig. 2, rather than accumulating structure through repeated attention capture, we extract attention matrices once from a single, designated timestep and reuse them throughout the denoising process. This approach preserves structural alignment while significantly reducing computational overhead.

A key design consideration in this framework is the selection of the optimal timestep t^* for attention extraction. We conduct an empirical evaluation across a range of candidate steps and identify the one (661) that yields the strongest structural alignment in the final output. In contrast to inversion-based methods, which require traversing a full reverse denoising trajectory to reach t^*, we adopt a lightweight forward simulation strategy: we apply the forward noise process directly to the reference latent $\mathbf{x}_0$ to simulate the noised latent at timestep t^*, bypassing reverse diffusion entirely. Specifically, we compute the noised latent as:

$$\mathbf{x}_{t^*} = \sigma_{t^*} \cdot \boldsymbol{\epsilon} + (1 - \sigma_{t^*}) \cdot \mathbf{x}_0 \tag{1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ is standard Gaussian noise, and $\sigma_{t^*} \in [0, 1]$ is the timestep-dependent noise scale factor at optimal timestep t^*. A single denoising step is then applied to $\mathbf{x}_{t^*}$ to produce intermediate attention maps. From this, we extract the self-attention query matrices $\mathbf{Q}_{t^*}^{(l)}$ at each transformer layer l. During generation, these matrices are injected at every timestep t by replacing the model's dynamically computed queries:

$$\mathbf{Q}_t^{(l)} \leftarrow \mathbf{Q}_{t^*}^{(l)} \tag{2}$$

The key ($\mathbf{K}$) and value ($\mathbf{V}$) matrices remain dynamically computed from the evolving latent $\mathbf{x}_t$, preserving responsiveness to the generative context while maintaining consistent structural queries. This procedure introduces no per-image tuning and requires only a single additional denoising step, making our method highly efficient, training-free, and broadly applicable across diffusion architectures.

Layer-Aware Injection for Preserving Appearance Quality. While one-step attention extraction provides effective structural control, indiscriminate injection across all layers can degrade visual quality. In particular, injecting structural $\mathbf{Q}$ matrices into early layers of the diffusion transformer interferes with low-level synthesis tasks—such as color, lighting, and texture modeling—often resulting in desaturation, flat textures, or unnatural shading. This occurs because early layers are primarily responsible for fine appearance features, and rigid structural guidance can disrupt their generative flexibility.

In contrast, deeper transformer layers capture higher-level semantic and spatial information, making them more suitable targets for structural injection. To balance structure and appearance, we adopt a layer-aware injection strategy that applies structural $\mathbf{Q}$ matrices only to mid-to-late layers (the blue layers in fig. 2). This preserves structure alignment while allowing early layers to focus on generating detailed and visually rich content.

Figure 3: Left: Noisy artifacts induced by the noise term. Right: Different granularity of structural control under different key timesteps.

3.2 Latent-Condition Decoupling (LCD) for Enhanced Attention Quality

With attention now extracted from a single forward-simulated timestep, we gain a stable and isolated point of intervention for improving control. We introduce *Latent-Condition Decoupling (LCD)* to exploit this opportunity. Rather than varying the key timestep t^*, LCD disentangles the two core influences on attention quality: (1) the noised latent $\mathbf{x}_{t^*}$ provided as input, and (2) the key timestep passed to the model. By isolating and manipulating these factors independently, we gain deeper insight into how structural guidance arises—and unlock both improved fidelity and fine-grained control.

To isolate the contribution of the noised latent, we fix the key timestep (at previously optimal 661) and vary the construction of $\mathbf{x}_{t^*}$. As shown in fig. 3, we find that latents generated with high noise levels (i.e., large σ_{t^*}) tend to introduce visible noise artifacts such as scattered dots in the final image. These high-noise latents also degrade the attention maps extracted for structural guidance, likely due to the model's inability to reason clearly over heavily corrupted input.

Based on this insight, we propose a simplified latent construction that removes the stochastic noise term entirely. Rather than performing forward diffusion with sampled noise, we directly construct a scaled $\tilde{\mathbf{x}}$, which serves as a substitute for $\mathbf{x}_{t^*}$.

$$\tilde{\mathbf{x}} = (1 - \sigma) \cdot \mathbf{x}_0,$$

where σ here becomes a tunable scale factor, independent of the key timestep. The removal of the noise term improves the stability of the proposed method. This noise-free latent simulates the amplitude characteristics of an intermediate timestep while preserving the spatial coherence of the original image latent. Empirical evaluation shows that moderate values (e.g., $\sigma \in [0.25, 0.5]$) yield the best results.

We then fix the latent input $\tilde{\mathbf{x}}$ and vary the key timestep passed to the diffusion transformer. As shown in fig. 3, the choice of key timestep affects the granularity of structural control. Conditioning with a timestep near zero yields prominent but coarse structure—capturing large shapes and global layout while omitting fine detail. In contrast, using a timestep closer to the original key timestep (e.g., 661) achieves finer structure transfer, preserving contours, texture boundaries, and detailed object shapes.

This observation opens the door to user-driven structural tradeoffs: by adjusting the conditioning timestep, one can control the rigidity of structural guidance. Lower key timesteps provide more compositional flexibility—suitable for creative reinterpretations or stylistic variation—while higher key timesteps enforce stricter alignment with the reference structure. This tunable granularity makes LCD not only a tool for quality improvement, but also a mechanism for interactive control.

3.3 Unlocking the Full Potential of FreeControl with Compositional Generation

With the core attention control mechanism in place, FreeControl serves as a flexible framework for structural guidance, enabling users to intuitively define spatial and semantic layout without modifying

the model. Rather than relying on hand-crafted segmentation maps or prompt engineering, users can directly control both the content and position of visual elements through image composition.

Compositional Reference Images for Semantic Layouts. While structural control methods like ControlNet or T2I-Adapter are effective at enforcing edges or spatial layouts, they often struggle to preserve semantic intent—particularly in complex scenes. FreeControl addresses this by enabling compositional reference images, where users can define both *what* should appear and *where* it should appear using direct visual assembly.

For instance, a user can extract an object (e.g., a kite) from a source image using a segmentation tool like SAM [14] and paste it onto a new background (e.g., sky). The resulting image encodes both spatial structure and semantic intent, and serves as a direct condition for generation. This "design by composition" approach allows users to guide the layout in a natural, intuitive way—without requiring segmentation maps or prompt engineering. Examples are shown in fig. 1, which comprise cases such as transferring digital text to real writing and scene composition.

To improve robustness when assembling such references, Gaussian blur could be optionally applied to the compositional image before passing it to the model. This lightweight preprocessing step reduces sharp boundaries and high-frequency noise, helping the model focus on the intended structure while avoiding artifacts from copy-paste seams.

4 Experiments

4.1 Implementation details

We conduct all experiments using the FLUX.1-dev [16] model with the FlowMatchEulerDiscrete scheduler, a timestep range of 1000 to 400, and a guidance scale of 6.5. Quantitative results use 25 denoising steps; 50 steps are used elsewhere for improved visual quality. The key timestep t^* is fixed at 661. Attention is extracted once and injected into the last 25 transformer layers of the model's single transformer block in the quantitative evaluations, and may be reduced elsewhere to demonstrate results of lower structural control. Compositional image generation is disabled unless specifically ablated. Inference is performed on a single NVIDIA RTX A6000 GPU with 48 GB of memory, and the inference time is measured over 100 runs.

4.2 Quatitative Comparison

Dataset. We evaluate on 5,000 images sampled from the COCO 2017 [20] validation set, resized to 512×512. Each image is paired with its corresponding caption, which is used as the input text prompt for controlled generation.

Metrics. We report **FID** for visual fidelity, **SSIM** and **PSNR** for low-level similarity, and **CLIP-Text Similarity** [31] for semantic alignment between images and prompts. For Canny-conditioned models, we quantify structural fidelity with the **F1 score** computed between the input Canny edge map and the Canny edge map extracted from each generated image. For depth-conditioned models, we report pixel level accuracy as the **mean squared error (MSE)** between the input depth map and the depth map predicted from the generated image.

Comparison Methods. We compare FreeControl against five strong baselines: ControlNet [47], UniControlNet [49], UniControl [30], ControlNet++ [17], and Flux-ControlNet [43, 44]. The first four baselines are implemented on Stable Diffusion v1.5, while Flux ControlNet is built on FLUX.1-dev[16]. All SD 1.5-based models are run with 20 denoising steps, and Flux-based methods—including FreeControl—use 25 steps, following the respective official configurations.

Note that ControlNet-style methods require pre-processed condition maps (e.g., Canny edge or depth), while FreeControl directly uses the raw image as structural input, with no preprocessing(we also did not count the preprocessing time for the comparison methods in table 2). The Canny edge is computed with the high threshold set to 200, and the low threshold set to 100.

Results. Table 1 reports quantitative comparisons across several metrics. FreeControl outperforms all baselines in terms of structural similarity (SSIM and PSNR), while maintaining competitive CLIP-Text alignment with prompt semantics. Compared to ControlNet and UniControl-style meth-

Table 1: Comparison with controlled-generation methods. The best scores are in bold, and the second scores are under lined.

Configuration	F1 ↑ / MSE ↓	FID ↓	SSIM ↑	PSNR ↑	CLIP-T ↑
ControlNet SD1.5 (Canny) [47]	0.23 / *	18.18	0.2585	10.55	0.3083
ControlNet++ (Canny) [17]	<u>0.30</u> / *	22.06	0.2784	10.59	0.2986
UniControl (Canny) [30]	**0.35** / *	21.22	0.3714	11.66	0.3103
UniControlNet (Canny) [49]	0.26 / *	17.97	0.2783	10.59	<u>0.3137</u>
FLUX.1-dev ControlNet (Canny) [43]	0.16 / *	27.11	0.2515	10.65	0.3009
ControlNet SD1.5 (Depth) [47]	* / 30.64	18.09	0.2383	10.22	0.3107
FLUX.1-dev ControlNet (Depth) [44]	* / 47.04	19.27	0.1968	10.74	0.3087
ControlNet++ (Depth) [17]	* / 27.79	23.23	0.2093	9.71	0.3020
UniControl (Depth) [30]	* / 33.51	28.24	0.2255	10.09	0.3105
UniControlNet (Depth) [49]	* / 34.72	22.25	0.2038	10.12	**0.3156**
Ours (Iterative Extraction)	<u>0.30</u> / **20.76**	<u>16.43</u>	**0.8078**	**19.11**	0.3043
Ours (One Step Extraction)	0.28 / <u>21.18</u>	**15.64**	<u>0.7564</u>	<u>17.49</u>	0.3087

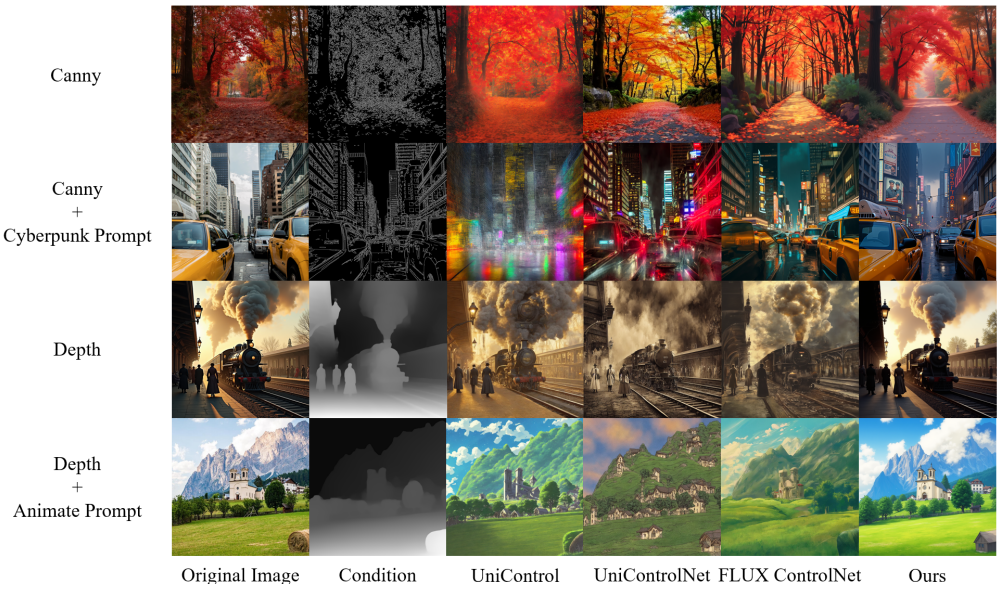

Figure 4: Qualitative comparisons on structure-conditioned image generation. Rows 1 and 3 show results where all methods are conditioned using the original caption of the reference image. Rows 2 and 4 present generations under stylized prompts to evaluate each method's ability to generalize beyond the original content.

ods—which rely on handcrafted edge or depth inputs—our method achieves higher visual fidelity without requiring retraining or specialized condition maps.

In edge-conditioned tasks, FreeControl achieves an F1 score of 0.30 with lower mean squared error (MSE), produces Canny edge results comparable to UniControl and ControlNet++ while using only the raw reference image. Additionally, compared to the iterative extraction, our method performs competitively while exerting significantly less computational resources.

These results demonstrate that FreeControl not only preserves spatial structure and semantic content effectively but also serves as a lightweight, training-free alternative to existing structure-conditioned generation pipelines. For further reference regarding the flexibility and fidelity of our method, we also provide quantitative comparisons that benchmark under different settings in the Appendix.

4.3 Qualitative Results

We present qualitative comparisons in fig. 4, where FreeControl is conditioned on raw reference images, while baseline methods rely on preprocessed control signals. Under the original prompt,

FreeControl delivers superior structural alignment and visual fidelity. In contrast, comparison methods either fail to align accurately with the intended structure or generate artifacts such as blur or noise, undermining image quality. We further demonstrate results on representative stylized prompts to evaluate generalization beyond the original setting. FreeControl successfully preserves structural integrity while adapting to new prompts, demonstrating robust guidance under prompt variation. Canny-based methods rigidly adhere to edge maps, often at odds with prompt semantics—resulting in unnatural appearances and ghost artifacts. Depth-based methods suffer from insufficient detail in the control signal, leading to misalignment, semantic drift, and diminished image fidelity. Overall, the results underscore FreeControl's ability to maintain consistent structural control and prompt adherence, even when guided by raw image inputs rather than handcrafted control maps.

4.4 Inference Time

We benchmark inference speed for FLUX ControlNet (Canny), FLUX ControlNet (Depth), the vanilla FLUX pipeline, and our method. All models are run with 25 denoising steps and produce 1024×1024 px outputs on an NVIDIA RTX A6000. For each pipeline, we fix a single prompt and a single source image—together with its corresponding condition map (canny edges or depth)—and execute the generation 100 times, recording the elapsed time at every run. Note that we exclude the pre-process time of the comparison methods for a fair comparison. The aggregate statistics from these 100-run trials are reported in table 2, and our method, being a test-time augmented method, performs equally efficiently as the training-based methods. Beyond that, the additional memory usage brought by our method is around 1 GB, which is also negligible.

Table 2: Inference time of different methods.

Configuration	**Average Inference Time**	Max	Min	Variance
FLUX Original Pipeline FLUX.1-dev [16]	24.89	24.98	24.18	0.0101
FLUX.1-dev ControlNet (Canny) [43]	26.01	26.52	25.61	0.0054
FLUX.1-dev ControlNet (Depth) [44]	26.01	26.09	25.80	0.0034
Ours	26.11	26.16	25.32	0.0117
Ours(Iterative Extraction)	45.16	48.09	45.01	0.1012

4.5 Compatibility with Fine-Tuned or LoRA-Augmented Models

ControlNet [47] often exhibit limited compatibility with finetuned or augmented via LoRA [9]. This instability arises because ControlNet relies heavily on the original backbone's parameters—its control branches are trained jointly with the base model and assume specific internal feature distributions. However, our method is not dependent on any specific model architecture or weights, demonstrating strong adaptability across different model variants. To validate this, we conduct experiments on both fine-tuned [1] and LoRA-augmented [26] models, comparing our method with ControlNet FLUX (Canny) and FLUX ControlNet (Depth). As shown in fig. 5, our method demonstrates superior compatibility by providing stable results with consistent structure and fine adaptation to the model changes in the community models, whereas ControlNet fails to be compatible with them and produce artifacts and distorted results. More results can be found in the Appendix.

5 Ablation Study

5.1 One-Step vs. Iterative Attention Extraction

To validate the effectiveness of extracting attention from a single timestep, we compare our method against a baseline that mimics iterative attention extraction across multiple denoising steps—similar to inversion-based or reconstruction-based strategies. In this baseline, attention matrices are extracted and injected step-by-step, rather than reused. As shown in table 1 and table 2, one-step injection achieves comparable structural fidelity while significantly reducing computational overhead. This result supports our hypothesis that structural information can be captured once and reused without loss of guidance, due to the shared purpose of structural encoding across timesteps.

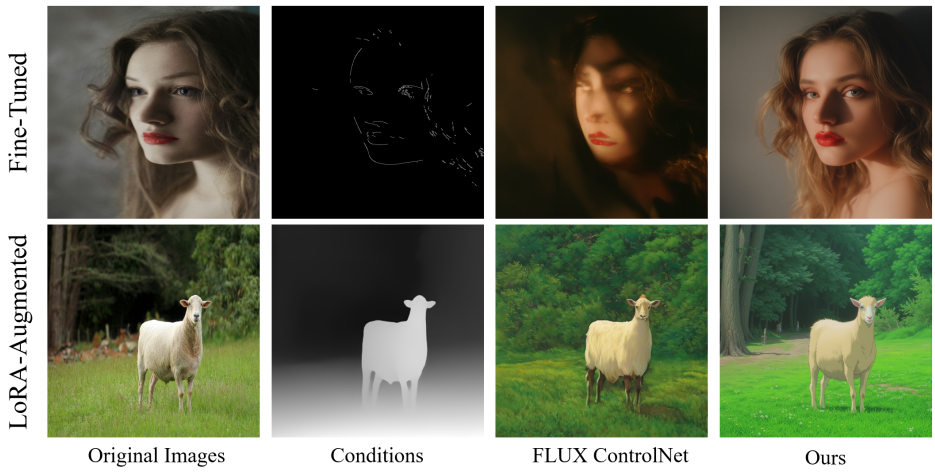

Figure 5: Examples of compatibility with fine-tuned or LoRA-augmented models.

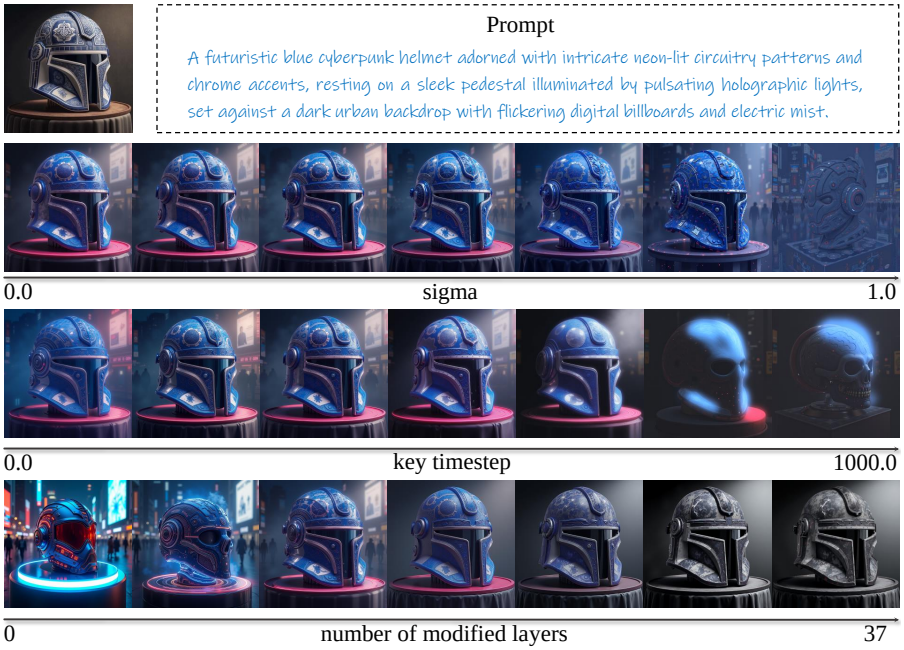

Figure 6: Visual analysis of structural control effects by varying injection depth, sigma, and key timestep in FreeControl.

5.2 Injection Depth vs. Sigma vs. Key timestep

The injection depth (number of transformer layers) influences the strength of structural control, while the injection quality (i.e., the content of the attention matrix) determines its focus. To isolate the effects of each factor, we vary it individually while holding others fixed at empirically optimal values. As shown in fig. 6, varying the injection depth reveals a different trade-off: injecting into fewer transformer layers relaxes structure in less critical regions, improving texture and color fidelity, whereas deeper injection increases rigidity at the cost of visual richness—especially in color saturation. The choice of key timestep, on the other hand, influences the focus of attention—that is, what kind of structural information is being injected. Earlier timesteps (e.g., $t = 0$) yield more abstract, layout-level attention that allows greater freedom in fine details; later timesteps (e.g., $t = 661$) still capture high-level structure but with greater specificity and finer granularity, resulting in more detailed structural alignment. The sigma value, in contrast, remains relatively stable at moderate settings, with

Figure 7: Qualitative Comparison between Layer-Aware Injection and Full-Layer Injection

structural control gradually diminishing as it approaches 1—first affecting fine details, then larger structures. Based on these observations, we recommend adjusting layer depth and injection range to tune structural strength, while selecting the appropriate key timestep and sigma to steer the level of structural detail captured in the generation.

5.3 Impact of Layer-Aware Injection

As discussioned in section 3.1, mid-to-late layer queries encode structural layout rather than raw appearance, aligning better with our structural control objective. However, early-layer queries primarily encode low-level appearance statistics, especially color. When injected, they conflict with the prompt-conditioned key/value features, which often imply a different color palette. This feature mismatch causes the model to lose chromatic fidelity, leading to muted or grayscale-like outputs. To verify this, we test prompts with deliberately shifted color themes. As shown in fig. 7, full-layer injection causes localized color fading in regions where prompt colors diverge from the reference, while our layer-aware injection maintains both structure and visual quality.

6 Limitations

While FreeControl is efficient, training-free, and offers strong structural control, it does not support condition maps like edges or segmentation. This limits scenarios where users prefer editing sketches or symbolic inputs. Although compositional image assembly provides flexibility, some use cases may still benefit from explicit support for sparse conditions.

7 Conclusion

This paper revisits a central assumption in attention-based structural control for diffusion models: that effective guidance requires multi-step extraction. We show that a single-step extraction—when properly conditioned—can offer strong, reusable structural signals without inversion or retraining. Our Latent-Condition Decoupling (LCD) reveals that attention quality depends not just on the timestep, but on how the noised latent and conditioning signal are configured. This enables more stable and controllable generation. Beyond efficiency, FreeControl supports intuitive control by allowing users to compose reference images that express both layout and intent—bridging structure and semantics without relying on edge maps or segmentation masks. Overall, our findings suggest attention can serve not just as an internal mechanism, but as a practical, tunable approach for control diffusion models.

8 Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62406134), Jiangsu Provincial Science & Technology Major Project (Grant No. BG2024042), the Suzhou Key Technologies Project (Grant No. SYG2024136) and the Nanjing University-China Mobile Communications Group Co. Ltd. Joint Institute.

References

[1] AWplanet. Awportraitxl, 2025. Accessed: 2025-05-22.

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023.

[4] 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 transactions on Graphics (TOG)*, 42(4):1–10, 2023.

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[6] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[7] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[9] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[11] Ifmain. Ultrareal_fine-tune, 2025. Accessed: 2025-05-22.

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[13] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6007–6017, 2023.

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[15] Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. *arXiv preprint arXiv:2412.08629*, 2024.

[16] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[17] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback: Project page: liming-ai. github. io/controlnet_plus_plus. In *European Conference on Computer Vision*, pages 129–147. Springer, 2024.

[18] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22511–22521, 2023.

[19] Jiang Lin and Zili Yi. Inversion-free video style transfer with trajectory reset attention control and content-style bridging. *arXiv preprint arXiv:2503.07363*, 2025.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[21] Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7817–7826, 2024.

[22] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[23] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047, 2023.

[24] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 4296–4304, 2024.

[25] OpenAI. gpt-4o, 2025. Accessed: 2025-05-22.

[26] Openfree. flux-chatgpt-ghibli-lora, 2025. Accessed: 2025-05-22.

[27] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[28] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[29] PrithivMLmods. Canopus-pixar-3d-flux-lora, 2025. Accessed: 2025-05-22.

[30] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023.

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

[32] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

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

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[35] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.

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

[37] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879, 2024.

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[39] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023.

[40] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[42] Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing, 2025.

[43] XLabs-AI. Flux-controlnet-canny-diffusers. `https://huggingface.co/XLabs-AI/flux-controlnet-canny-diffusers`, 2024.

[44] XLabs-AI. Flux-controlnet-depth-diffusers. `https://huggingface.co/XLabs-AI/flux-controlnet-depth-diffusers`, 2024.

[45] Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with language-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9452–9461, 2024.

[46] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[47] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[48] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer, 2025.

[49] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:11127–11150, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: I believe they are accurately reflected.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: It is addressed in the Limitations section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The information provided in the method and experiments section is sufficient to reproduce the results presented in this paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The paper has clearly provided the details needed to reproduce the method proposed to the extent to support our claim; however, we are yet unable to provide a properly formulated code regarding the method and all the experiments conducted. This paper will, however, open-source the code regarding its main method upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All the details can be found either in the method section or the implementation details section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Proper information, if needed, is provided to justify the statistical significance of the results, as in table 2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Relevant information is provided in the implementation details, and there is also a table regarding the time of executiontable 2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The author believe there is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper itself does not release any pretrained models, or datasets.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The datasets used are properly cited, and the assets are properly referenced if necessary.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Quantitative Results on Stylized Prompts (COCO Dataset)

FreeControl conditions generation directly on the raw image, rather than on derived conditions like edges or depth. This design provides a rich and detailed structural prior, which has proven effective in preserving the original layout and composition during generation. While prior experiments focus on tasks where the target image shares semantic similarity with the reference, real-world use cases often involve prompts that diverge stylistically or conceptually from the original image. To broaden the evaluation scope and assess FreeControl's robustness in more diverse generative settings, we conduct a benchmark using stylized prompts.

We construct this benchmark on the COCO [20] validation set. For each image, we retain the original as the structural reference and generate stylized prompts by combining the original caption with one of five target styles (e.g., Cyberpunk, Vaporwave). These stylized prompts are created using GPT-4o [25], conditioned on both the caption and the style keyword.

This setup introduces a challenging mismatch between the style and the visual structure, making it difficult for models to retain both prompt adherence and structural fidelity. We compare FreeControl against Flux-ControlNet [47], which serves as a strong baseline for structure-conditioned generation on the same backbone. As shown in table 3, FreeControl consistently preserves structure better (as measured by F1 and MSE) while generating content more semantically aligned with the stylized prompts (via CLIP-T). This demonstrates the strength of our method in transferring structure faithfully even when prompt semantics diverge from the original image.

B More Results on Compatibility with Fine-Tuned or LoRA-Augmented Models

To further support the findings discussed in the main paper, we provide additional qualitative results on fine-tuned and LoRA-augmented diffusion models. Specifically, we evaluate FreeControl and FLUX ControlNet variants on community models that are fine-tuned [1, 11] or LoRA-augmented [26, 29].

As shown in figs. 8 to 11, our method consistently preserves structure and semantic fidelity across diverse model variants, producing stable and visually coherent outputs. In contrast, ControlNet-based approaches exhibit visible artifacts, color shifts, or loss of structural alignment under the same settings.

These results further confirm that FreeControl maintains strong compatibility across both fine-tuned and LoRA-augmented backbones, benefiting from its training-free nature and independence from specific model weights or feature distributions.

C Additional Visual Results

We provide more visual results for the readers to reference. fig. 12 showcases additional generation results based on compositional reference images. Users can crop and paste objects into a layout to specify spatial intent, allowing for precise control over the scene's composition. With a touch of creativity, FreeControl empowers users to bring their imaginative visions to life, generating stunning, dynamic visuals that reflect their unique concepts (like a giant floating whale in the sky). fig. 13 and fig. 14 provide further comparisons between FreeControl and FLUX ControlNet.

Table 3: Quantitative evaluation on the COCO validation set using stylized prompts. The best scores are in bold.

Method	**F1** ↑	**MSE** ↓	**SSIM** ↑	**PSNR** ↑	**CLIP-T** ↑
FLUX ControlNet (Canny)	0.19	N/A	0.2629	9.72	0.2646
FLUX ControlNet (Depth)	N/A	41.41	0.2029	9.98	0.2448
Ours	**0.25**	**26.24**	**0.4825**	**15.07**	**0.2981**

D Expanded Baseline Comparison Results

We have added quantitative comparison experiments with two image editing models, In-Context Edit [48] and Taming Rectified Flow [42], to improve the fairness and completeness of the evaluation. The test images are consistent with the previous comparison experiments. We use image captions and stylized prompts as text inputs respectively, and the corresponding results are shown in table 4 and table 5.

According to the results, Our method performs comparably to, and often surpasses, Taming Rectified Flow across several metrics. ICEdit, built on the FLUX-Fill model [16] for image inpainting, achieves relatively high similarity metrics (e.g., PSNR) primarily because it keeps all content outside the edited region untouched. However, this strategy limits its ability to satisfy the desired balance between structural control and free content generation. As a result, its CLIP-T score is lower, and it often struggles with stability and controllability when following editing instructions.

Table 4: Quantitative Results with image-editing methods on the COCO validation subset.

Method	**F1** ↑	**MSE** ↓	**SSIM** ↑	**PSNR** ↑	**CLIP-T** ↑	**FID** ↓
In-Context Edit	0.47	17.30	0.7781	21.34	0.3024	8.64
Taming Rectified Flow	0.19	28.31	0.4390	16.90	0.3164	16.35
Ours	0.28	21.18	0.7564	17.49	0.3087	15.64

Table 5: Quantitative Results with image-editing methods on the COCO validation subset using stylized prompts.

Method	**F1** ↑	**MSE** ↓	**SSIM** ↑	**PSNR** ↑	**CLIP-T** ↑
In-Context Edit	0.30	35.96	0.5436	14.70	0.2543
Taming Rectified Flow	0.17	38.49	0.4034	16.37	0.2585
Ours	0.25	26.24	0.4825	15.07	0.2981

E Additional Quantitative Ablation Analysis

We have supplemented more comprehensive ablation studies to justify the choice of key parameters in our method, such as the key timestep t^*, numbers of modified transformer layers and σ. The results are presented in table 6. Thanks to LCD, by flexibly tuning the hyperparameters, we can achieve structural control of varying strength and granularity, producing stable and controllable results that cater to different user requirements.

Table 6: Ablation results on on the COCO validation subset. Each entry in the Parameters column indicates the number of modified layers, the key timestep t^*, and the σ, in that order.

Parameters	**F1** ↑	**MSE** ↓	**SSIM** ↑	**PSNR** ↑	**CLIP-T** ↑	**FID** ↓
20-661-0.25	0.27	23.01	0.5251	16.61	0.3083	17.84
25-561-0.25	0.27	22.44	0.5232	16.61	0.3077	17.42
25-661-0.0	0.30	21.74	0.5630	17.20	0.3061	17.78
25-661-0.25	0.28	21.86	0.5438	16.87	0.3048	18.00
25-661-0.5	0.24	24.64	0.5097	16.54	0.3072	22.51
25-761-0.25	0.28	23.47	0.5492	17.07	0.3048	21.40
30-661-0.25	0.30	22.20	0.5461	16.88	0.3020	20.20

F Evaluation under Challenging Structural Scenarios

F.1 Semantic Entanglement and Object Occlusion

We conduct stress tests on both real and synthetic images featuring semantic entanglement and severe object overlap. Selected results are shown in fig. 15. For each example, the left image is the original input, and the right image is the generated result. As observed, FreeControl consistently avoids distorted or unrealistic artifacts such as extra limbs or warped body structures. The outputs remain natural-looking and visually coherent, demonstrating strong robustness even under highly challenging compositional scenarios.

F.2 Preservation of Facial Identity

Our method provides flexible control over facial identity preservation, allowing users to adjust the strength of identity retention via hyperparameters. Under non-conflicting text-image guidance, tuning parameters such as the number of modified transformer layers enables strong structural control while preserving facial details, making FreeControl suitable for identity-sensitive tasks. As shown in fig. 16, with higher control strength, our approach demonstrates a strong ability to retain facial identity. Conversely, for artistic creation or diversity-oriented generation, relaxing the control allows for slight, intentional changes in facial features, leading to more expressive results.

G Applicability to UNet-based Models (i.e., Stable Diffusion)

Our method is designed to operate purely at the attention level, making it architecture-agnostic. We have implemented it on UNet–based models (e.g., SD1.5 [33], SDXL [28]) and observed strong structural control behavior after only minimal hyperparameters adjustments to fit the model. Several qualitative examples of structural control are presented in fig. 17 and fig. 18. We further note that, due to the inherent capacity limitations of UNet-based models, the degree of controllability can diminish in highly complex scenarios. In practice, we find that leveraging FLUX models [16] yields more stable and visually coherent generations, and we recommend their use when high-fidelity control is desired.

H Further Discussion on the Design Space

H.1 Independence from ROPE

The query matrices FreeControl extracts are captured before RoPE [39] is applied, so the injected queries contain no positional encoding — they are entirely image-driven. While RoPE still affects key and value during generation, it does not alter what FreeControl injects. Furthermore, FreeControl works identically on U-Net architectures (which do not use RoPE), showing that structural consistency stems from the extracted queries themselves, not from positional priors.

To directly confirm this point, we ran a controlled test by removing RoPE entirely from the FLUX model. As expected, the base model collapsed into near-random noise, since it was never trained to operate without positional encoding. Crucially, when we applied FreeControl under the same no-RoPE setup, the one-step injection still imposed clear, image-driven structure on the output. The result looked like "structured noise" faithfully echoing the condition image's layout — strong evidence that FreeControl's guidance originates from the injected queries themselves, not from RoPE.

H.2 Key timestep Choice

The key timestep fundamentally governs the granularity of structural information that FreeControl can extract. In diffusion models, each denoising step is influenced not only by progressively refined latents but also by a changing timestep input that biases the network toward different levels of detail. Conceptually, the key timestep acts like a focus knob: adjusting it continuously shifts the model's representational emphasis from global layout patterns to fine-grained textures.

By holding the latent fixed and sweeping only the key timestep, our experiments reveal a natural progression in structural granularity encoded within the query matrices. Both quantitative metrics and

visual evidence show a smooth evolution—from coarse shape-level representation toward detailed texture-level encoding. Among all tested values, key timestep (661) emerges as a sweet spot, offering the best trade-off between global consistency and structural precision, making it the most suitable extraction point for query-based control.

H.3 Layer-wise Query Matrices Similarity

We extract query matrices at different depths under two configurations: with LCD and without LCD, where the only difference lies in whether noise is added to x_0 during the forward diffusion process (as defined in Sec. 3.1 of main paper). We then compare these two sets of query matrices with the query matrices from every timestep of the multi-step extraction variant and compute the cosine similarity, as reported in table 7. The layer-wise similarity shows a clear low-to-high trend: shallow layers tend to have lower similarity, while deeper layers converge more. This aligns with our layer-aware injection choice in Sec. 3.1 — early layers focus more on appearance elements rather than structure, diverging more across timesteps and contributing less to shared structural signals.

Table 7: Cosine Similarity Between One-Step and Multi-Step Extracted Query Matrices. The "Global" row reports the average similarity of query matrices across all layers.

Layer Depth	Cosine Similarity	
	w/ LCD	w/o LCD
Early	0.5418	0.6680
Mid	0.5861	0.6853
Last	0.6011	0.7638
Global	0.5769	0.7063

Figure 8: Visual results comparing our method and FLUX ControlNet on Lora-Augmented models (Ghibli-Style LoRA).

Figure 9: Visual results comparing our method and FLUX ControlNet on Lora-Augmented models (Canopus-Pixar-3D-Style LoRA).

Figure 10: Visual results comparing our method and FLUX ControlNet on finetuned models (AWPortrait Fine-Tune).

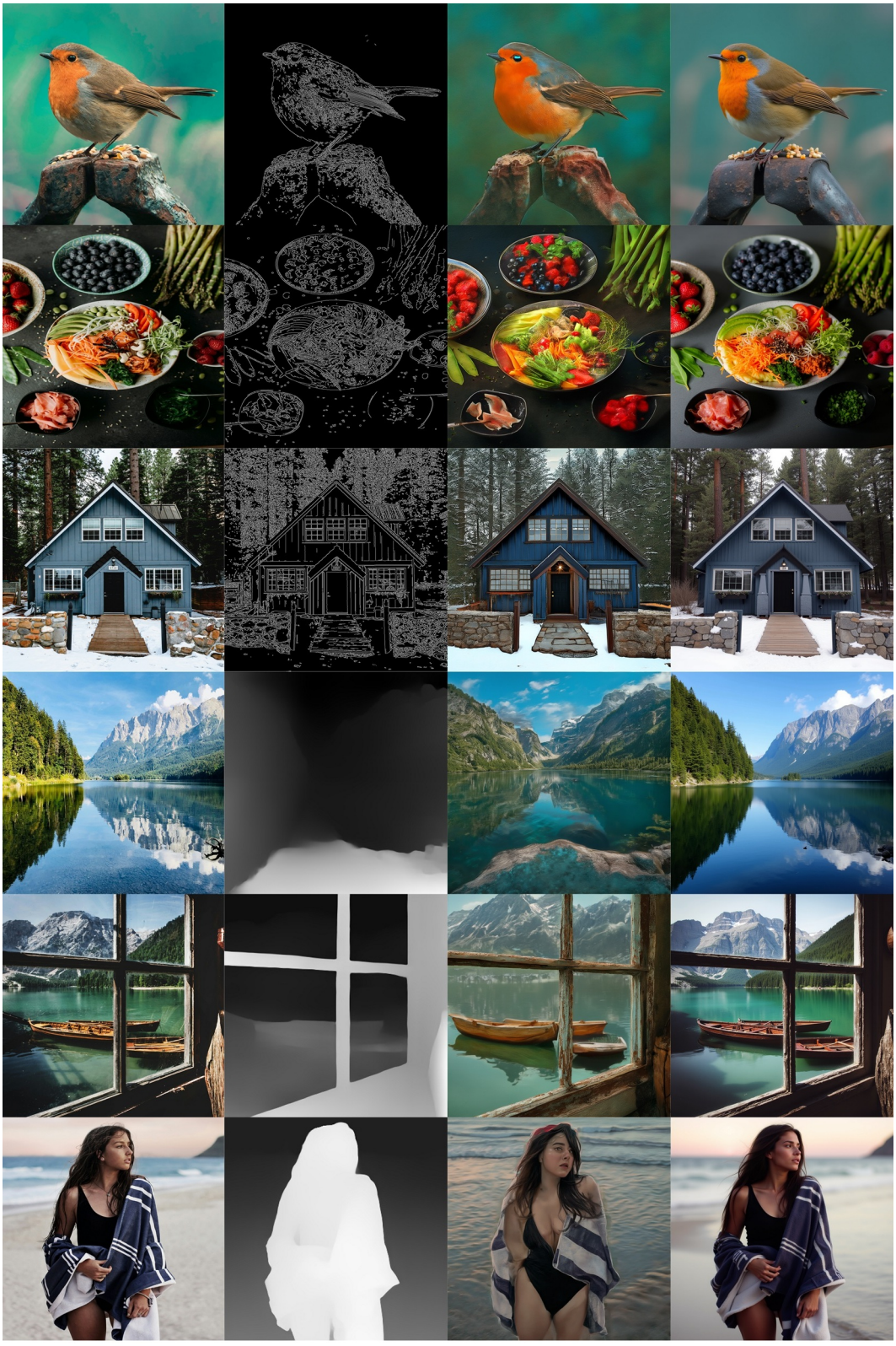

Figure 11: Visual results comparing our method and FLUX ControlNet on finetuned models (UltraReal Fine-Tune).

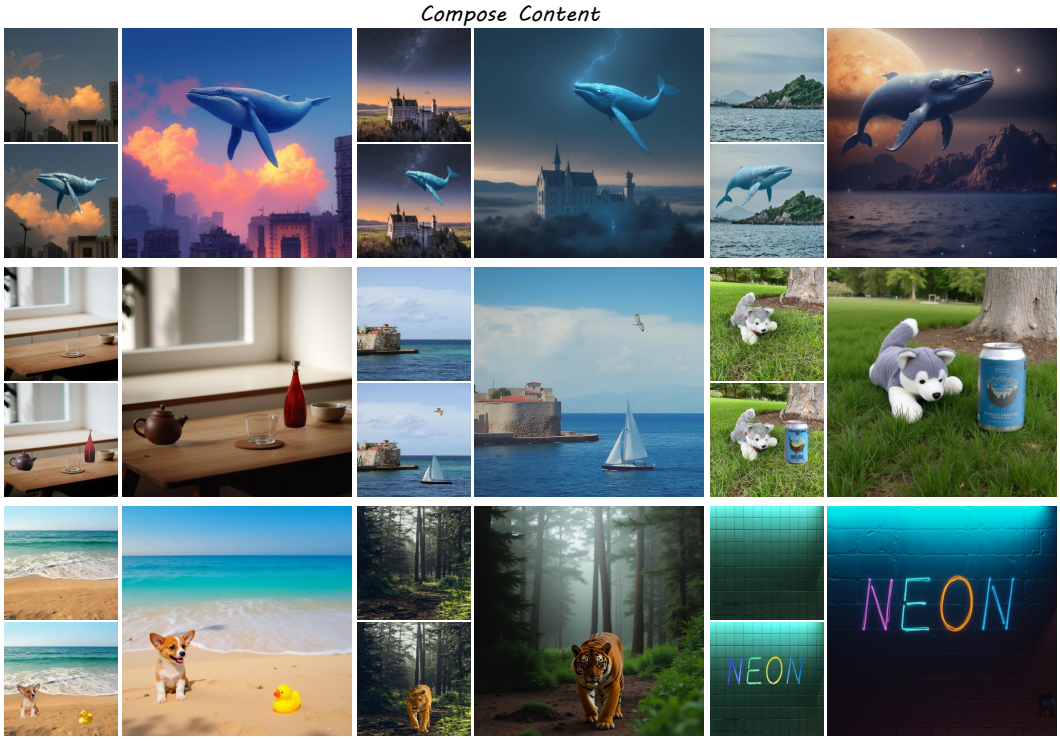

Figure 12: More visual results on compositional generation.

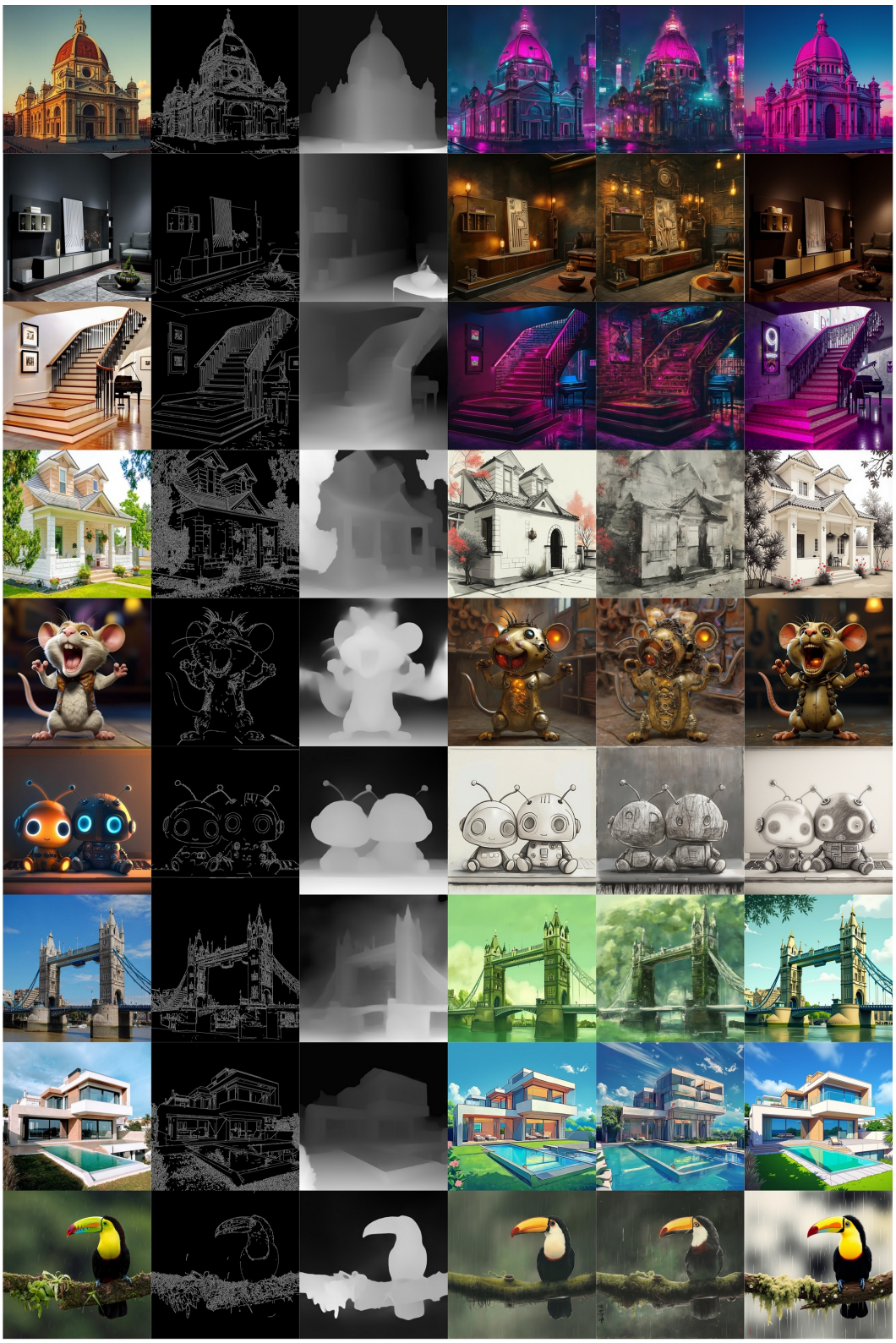

Figure 13: Qualitative comparisons on structure-conditioned image generation.

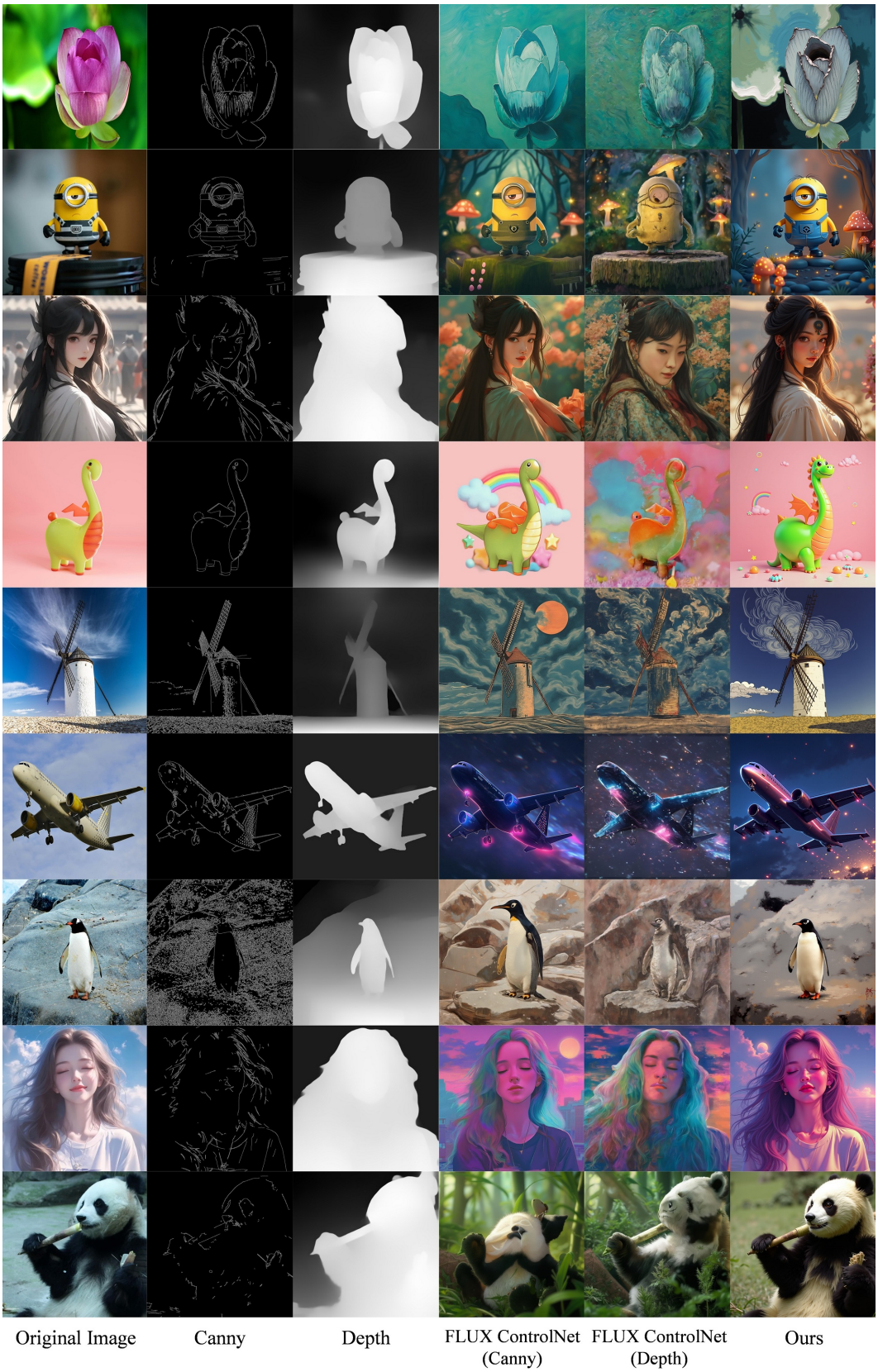

Figure 14: Qualitative comparisons on structure-conditioned image generation.

(a) (b) (c)

Figure 15: Examples of generated images under Semantic Entanglement and Object Occlusion. For each pair, the image on the left is the original image, and the image on the right is the generated result.

(a) (b) (c)

Figure 16: Examples of Facial Identity Control. Adjusting control strength, together with a suitable prompt, enables strong structural preservation of facial features. For each pair, the image on the left is the original image, and the image on the right is the generated result.

(a) (b) (c)

Figure 17: Generation Examples on SDXL model Using Our Method. For each pair, the image on the left is the original image, and the image on the right is the generated result.

(a) (b) (c)

Figure 18: Generation Examples on SD-1.5 model Using Our Method. For each pair, the image on the left is the original image, and the image on the right is the generated result.