FreeControl: Training-Free Spatial Control of Any Text-to-Image Diffusion Model with Any Condition
(Supplementary Material)

In the supplementary material, we present additional qualitative results (Section A) and ablation experiments (Section B), and discuss the limitations (Section C) and societal impact of our method (Section D). For sections and figures, we use numbers (*e.g*., Sec. 1) to refer to the main paper and capital letters (*e.g*., Sec. A) to refer to this supplement. We hope this document complements the main paper.

A. Additional Qualitative Results

Handling conflicting conditions. We study cases where spatial conditions have minor conflicts to input text prompts. We assume that a text prompt consists of a concept (*e.g*., batman) and a style (*e.g*., cartoon), and contrast a conflicting case with its aligned version. Specifically, a conflicting case includes (a) a text prompt with a feasible combination of concept and style; and (b) a spatial condition (*i.e*. an edge map) derived from real images without the text concept. The corresponding aligned case contains a similar text prompt, yet using a spatial condition from real images with the same concept. We input those cases into ControlNet, T2I-Adapter, and FreeControl, using a set of pre-trained and customized models.

Figure A shows the results. Our training-free FreeControl consistently generates high quality images that fit the middle ground of spatial conditions and text prompts, across all test cases and models. T2I-Adapter sometimes fails even with an aligned case (see *Batman* examples), not to mention the conflicting cases. Indeed, T2I-Adapter tends to disregard the condition image, leading to diminished controllability, as exemplified by *Emma Watson* example (conflicting). ControlNet can generate convincing images for aligned cases, yet often fall short in those conflicting cases. A common failure mode is to overwrite the input text concept using the condition image, as shown by *skeleton bike* or *house in a bubble* examples (conflicting).

Extension to Image-to-Image Translation FreeControl can be readily extended to support image-to-image (I2I) translation by conditioning on a detailed / real image. A key challenge here is to allow FreeControl to preserve the background provided by the condition, *i.e*., the input content image. To this end, we propose two variants of FreeControl. The first removes the mask $\mathbf{M}$ in structure guidance (*i.e*., w/o mask), and the second generates from the inverted latent $\mathbf{x}_T^g$ of the condition image (*i.e*., fixed seed). We find that removing the mask helps extract and maintain the background structure, and starting inference from $\mathbf{x}_T^g$ retains the appearance from the condition image.

Figure B evaluates FreeControl and its two variants for text-guided I2I, and compares to strong baselines for the I2I task including PnP [6], P2P [2], pix2pix-zero [4] and SDEdit [3]. The vanilla FreeControl, as we expect, often fails to preserve the background. However, our two variants with simple modification demonstrate impressive results as compared to the baselines, generating images that adhere to both foreground and background of the input image.

Further, we evaluate the *self-similarity distance* and *CLIP score* of FreeControl, its variants, and our baselines on the ImageNet-R-TI2I dataset. The results are summarized in Figure B. Variants of FreeControl outperform all baselines with significantly improved structure preservation and visual fidelity, following the input text prompts.

Continuous control. Real-world content creation is a live experience, where an idea develops from a sketch into a more refined and finished piece of work. The intermediate states throughout this process may be interpreted as continuously evolving control signals. Figure D illustrates how FreeControl may assist an artist in his or her content creation experience. It produces spatially accurate and smoothly varying outputs guided by constantly changing conditions, thus serving as a source of inspiration over the course of painting.

Compositional control. By combining structure guidance from multiple condition images, FreeControl readily supports compositional control without altering the synthesis pipeline. Figure E presents our results using different combinations of condition types. The generated images are faithful to all input conditions while respect the text prompt.

Combination with ControlNet. Figure C demonstrates the results of combining FreeControl and ControlNet(canny), using the wireframe of a teapot and the mesh of a bunny as the condition. We use FreeContorl to denoise the latent for 30 steps, ControlNet for the next 70 steps, and the vanilla Stable Diffusion for the rest 100 steps. This hybrid approach improves the structural alignment of FreeContorl, unlocks the appearance customization, improves textual alignment, and accommodates un-trained conditions

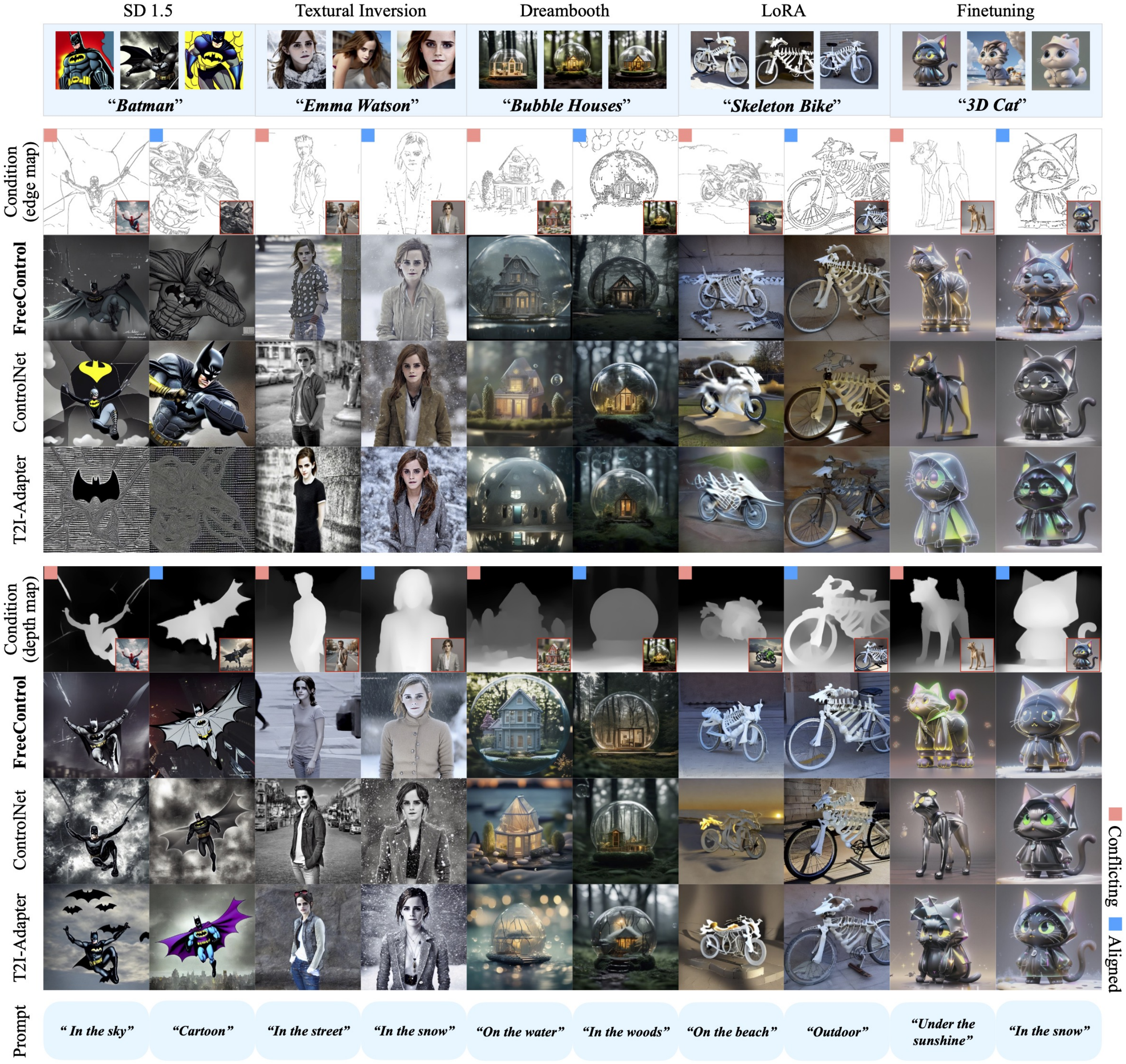

Figure A. **Controllable T2I generation of custom concepts.** FreeControl is compatible with major customization techniques and readily supports controllable generation of custom concepts without requiring spatially-aligned condition images. By contrast, ControlNet fails to preserve custom concepts given conflicting conditions, whereas T2I-Adapter refuses to respect the condition image and text prompt.

for ContorlNet.

B. Additional Ablation Study

We now present additional ablations of our model.

Size of semantic bases N_b. Figure G presents generation results over the full spectrum of N_b. A larger N_b improves structure alignment yet triggers the unintended transfer of appearance from the input condition. Hence, a good balance is achieved with N_b's in the middle range.

Number of seed images N_s. Figure H suggests that N_s has minor impact on image quality and controllability, allowing the use of *as few as* 1 *seed image* in the analysis stage. Large N_s diversifies image content and style, which helps perfect structural details (*e.g.*, limbs) in the generated images.

Choice of threshold τ_t. Figure I demonstrates that no *hard* threshold within the range of $[0, 1]$ can fully eliminate spuri-

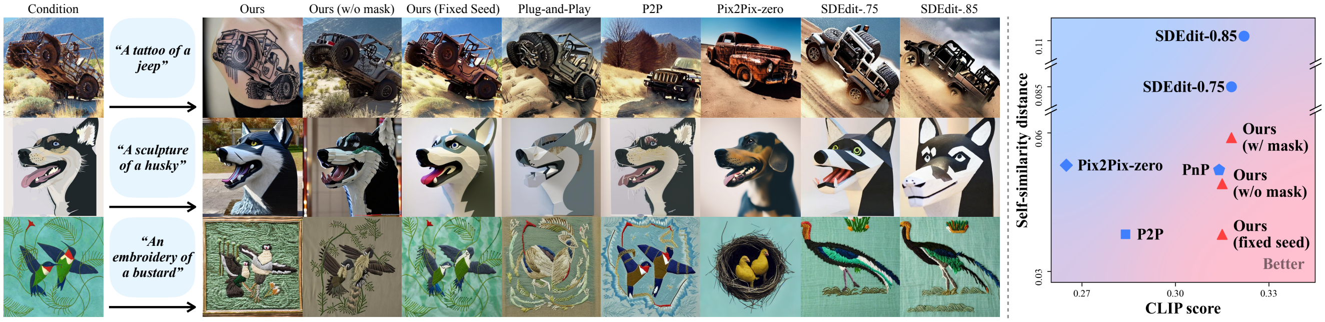

Figure B. **Qualitative and quantitative comparison on text-guided image-to-image translation.** FreeControl enables flexible control of image composition and style through guidance mask $\mathbf{M}$ and random seed (*left*). It strikes a good balance between structure preservation (self-similarity distance) and image-text alignment (CLIP score) in comparison to the baselines (*right*, better towards bottom right).

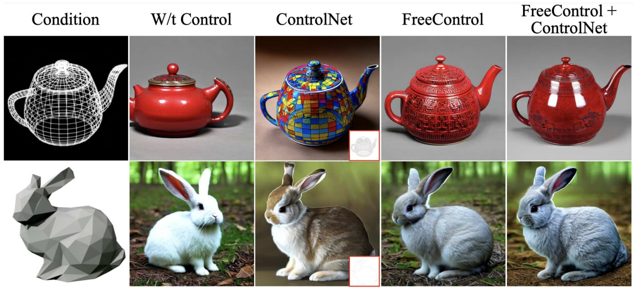

Figure C. **Qualitative results of combining ControlNet and FreeControl.** Top: *"A Chinese teapot, red"*; Bottom: *"A bunny, in the forest"*.

ous background signal while ensure a foreground structure consistent with the condition image. By contrast, our *dynamic* thresholding scheme, implemented as a per-channel `max` operation, allows FreeControl to accurately carve out the foreground without interference from the background.

Number of guidance steps. Figure J reveals that the first 40% sampling steps are key to structure and appearance formation. Applying guidance beyond that point has little to no impact on generation quality.

Choice of guidance weights λ_s and λ_a. Figure L confirms that FreeControl produces strong results within a wide range of guidance strengths. In particular, the output images yield accurate spatial structure when $\lambda_s \geq 400$ and rich appearance details when $\lambda_a \geq 0.2\lambda_s$. We empirically found that these ranges work for all examples in our experiments.

Basis reuse across concepts. Once computed, the semantic bases $\mathbf{S}_t$ can be reused for the control of semantically related concepts. Figure L provides one such example, where $\mathbf{S}_t$ derived from seed images of `man` generalize well on other mammals including `cat`, `dog` and `monkey`, yet fail for the semantically distant concept of `bedroom`.

C. Limitations

One limitation of FreeControl lies in its inference speed. Without careful code optimization, structure and appearance guidance result in 66% longer inference time (25 seconds) on average compared to vanilla DDIM sampling [5] (15 seconds) with the same number of sampling steps (200 in our experiments) on an NVIDIA A6000 GPU. This is on par with other training-free methods.

Another issue is that FreeControl relies on the pretrained VAE and U-Net of a Stable Diffusion model to encode the semantic structure of a condition image at a low spatial resolution (16 × 16). Therefore, it sometimes fails to recognize inputs with missing structure (*e.g.*, incomplete sketch), and may not accurately locate fine structural details (*e.g.*, limbs). Representative failure cases of FreeControl are illustrated in Figure F.

D. Societal Impact and Ethical Concerns

This paper presents a novel training-free method for spatially controlled text-to-image generation. Our method provides better control of the generation process with a broad spectrum of conditioning signals. We envision that our method provides a solid step towards enhancing AI-assisted visual content creation in creative industries and for media and communication. While we do not anticipate major ethical concerns, our method shares common issues with other generative models in vision and graphics, including privacy and copyright concerns, misuse for creating misleading content, and potential bias in the generated content.

References

[1] Civitai. `https://civitai.com/`. 4

[2] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, 2023. 1

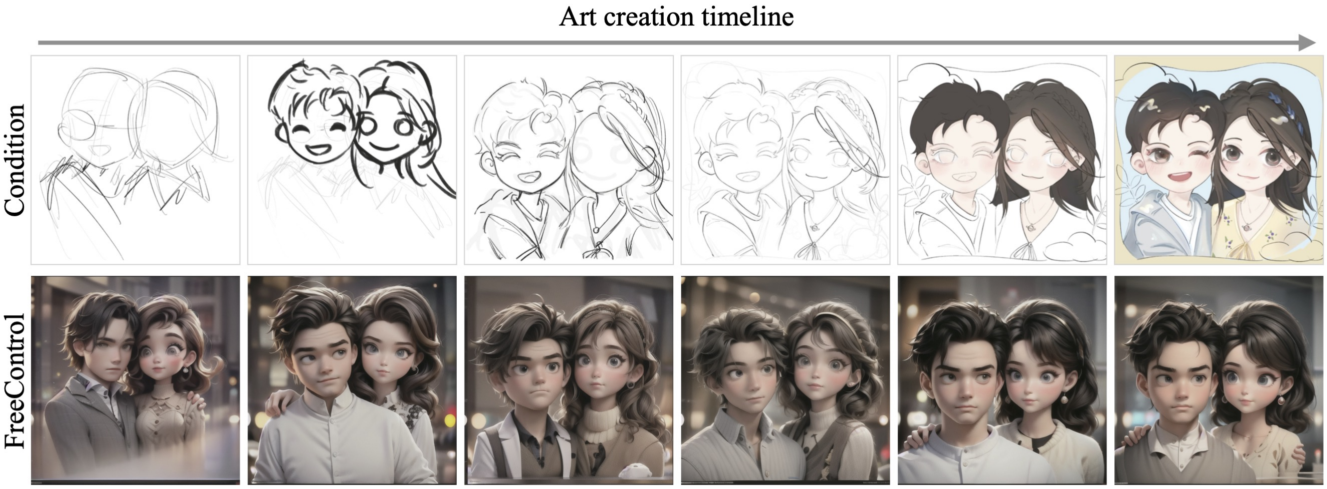

Figure D. **Controllable generation over the course of art creation.** Images are generated from the same seed with the prompt *"a photo of a man and a woman, Pixar style"* with a customized model from [1]. FreeControl yields accurate and consistent results despite evolving control conditions throughout the art creation timeline.

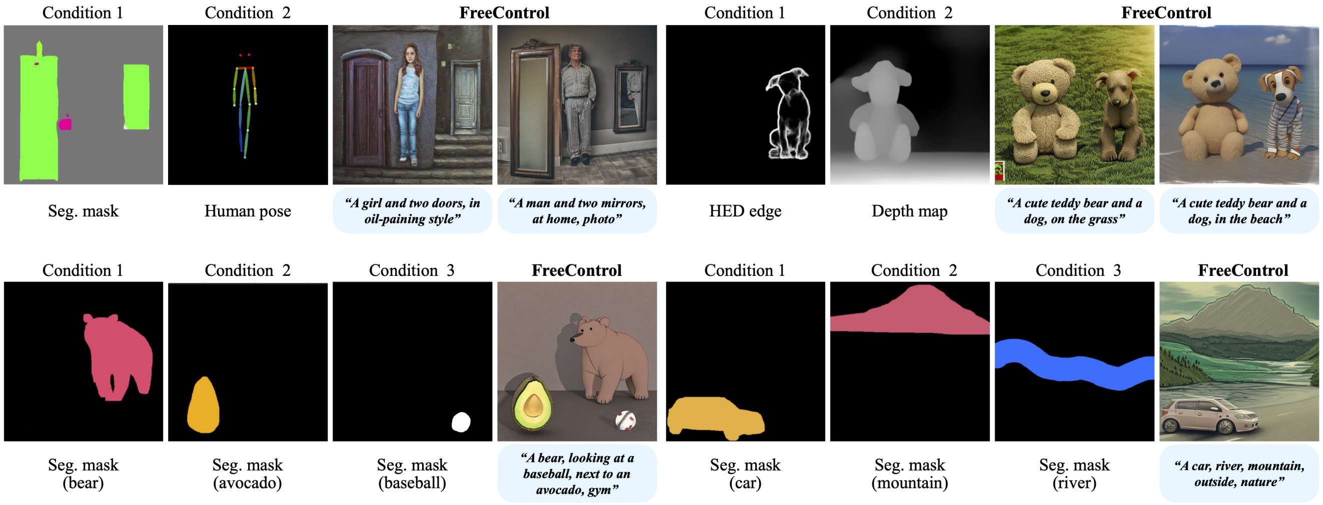

Figure E. **Qualitative results on compositional control**. FreeControl allows compositional control of image structure using multiple condition images of potentially different modalities.

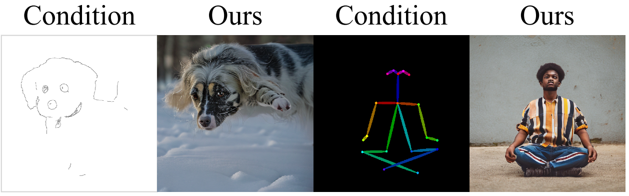

Figure F. **Failure cases.** FreeControl does not anticipate missing structure in the condition image (*left*) and may not accurately position fine structural details (limbs) in the output image (*right*).

[3] 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. In *ICLR*, 2022. 1

[4] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *SIGGRAPH*, 2023. 1

[5] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 3

[6] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023. 1

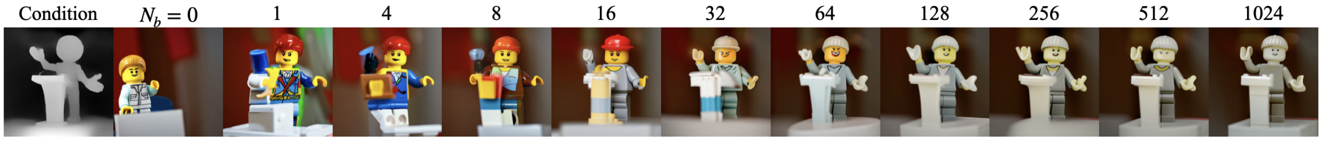

Figure G. **Ablation on size of semantic bases N_b.** Images are generated using the prompt *"a Lego man giving a lecture"*. They illustrate an inherent tradeoff between structure and appearance quality. A good balance can be achieved with N_b's in the middle range.

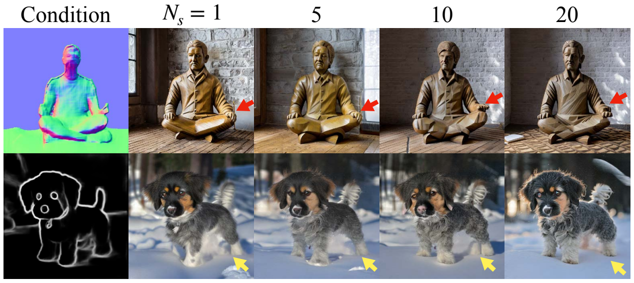

Figure H. **Ablation on number of seed images N_s.** Top: *"wooden sculpture of a man"*; Bottom: *"dog, in the snow"*. Larger N_s brings minor improvement on structure alignment.

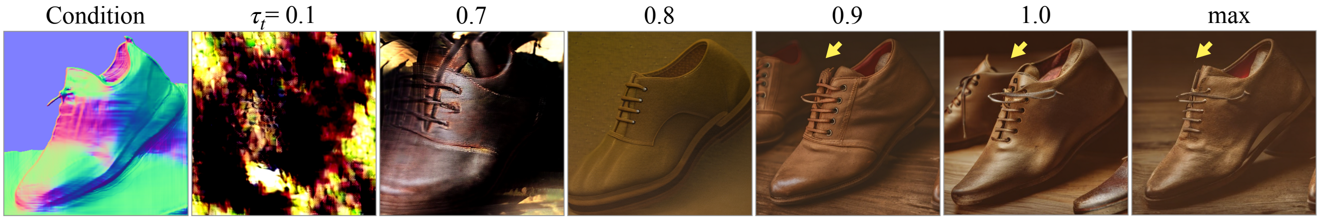

Figure I. **Ablation on threshold τ_t.** Images are generated using the prompt *"leather shoe on the table"*. Our dynamic threshold (max) encourages more faithful foreground structure and cleaner background in comparison to various hard thresholds (*e.g.*, 0.1).

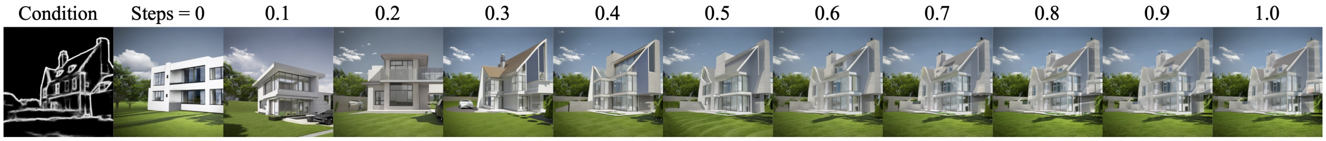

Figure J. **Ablation on number of guidance steps.** Images are generated using the prompt *"a modern house, on the grass, side look"*. Applying guidance beyond the first 40% diffusion steps (0.4) has little to no impact on the generation result.

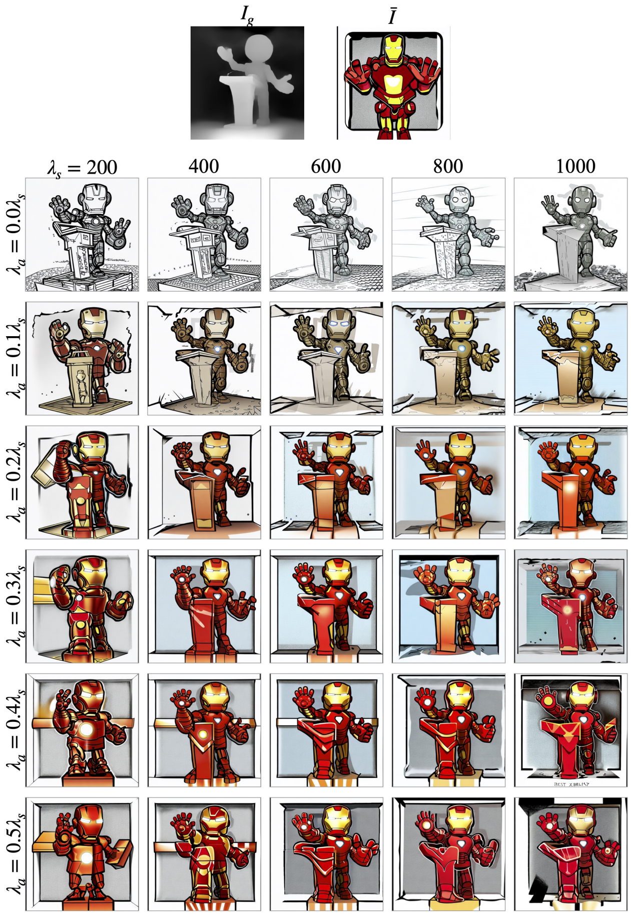
I_g
$\bar{I}$
λ_s = 200
400
600
800
1000
λ_a = 0.0λ_s
λ_a = 0.1λ_s
λ_a = 0.2λ_s
λ_a = 0.3λ_s
λ_a = 0.4λ_s
λ_a = 0.5λ_s

Figure K. **Ablation on guidance weights λ_s and λ_a.** Images are generated with the prompt *"an iron man is giving a lecture"*. FreeControl yields strong results across guidance weights.

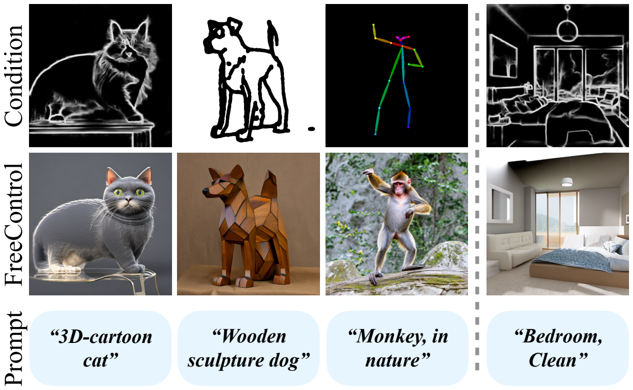

Figure L. **Ablation on basis reuse.** The semantic bases computed for *"man"* enable the controllable generation of semantically related concepts (cat, dog, and monkey) while falling short for unrelated concepts (bedroom).