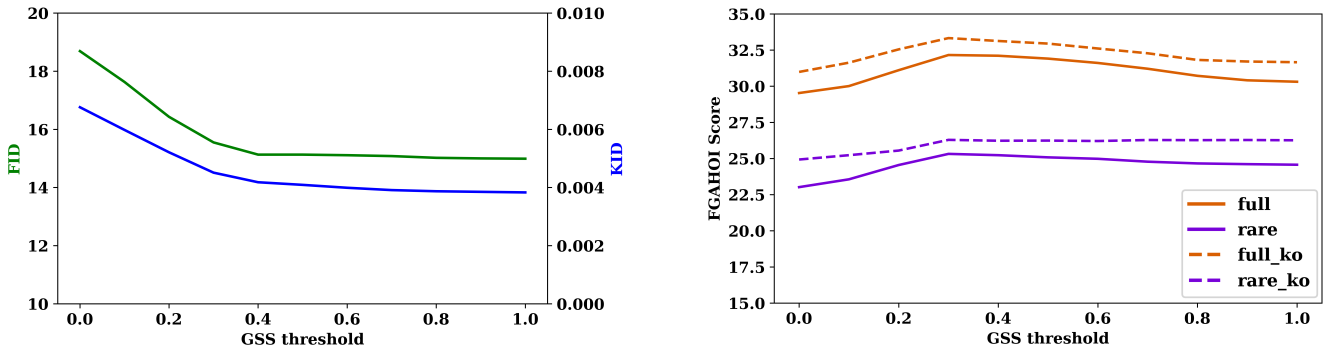
GR-Diffusion: Graph-guided Relational-aware Diffusion via Attention Alignment

Supplementary Material



(a) Image Fidelity vs. τ

(b) HOI Score vs. τ

Figure 1. Sensitivity analysis of the GSS threshold τ. The results demonstrate that $\tau = 0.3$ achieves the optimal trade-off between controllability and fidelity.

1. Implementation Details

Experimental Environment. Our framework is based on Diffusers library. All inference experiments and time consumption evaluations are conducted on a workstation equipped with a single GeForce RTX 3090 GPU. To optimize memory usage and inference speed without compromising generation quality, we utilize FP16 half-precision for the entire inference process.

Model Configuration. As mentioned in the main paper, we use InteractDiffusion v1.2 (based on Stable Diffusion v1.5) and InteractDiffusion-XL v1.0 (based on Stable Diffusion XL) as our backbones. For the guidance mechanism, the Node Alignment Guidance (NAG) and Edge Enhancement Guidance (EEG) are applied to the attention layers within the down-sampling blocks and the middle block of the U-Net. To ensure a fair comparison, the key hyperparameters are held constant across all experiments: the Graph Similarity Detector (GSD) activation threshold is set to $\tau = 0.3$, the base self-attention injection strength is set to $\beta_{base} = 1.0$, and the GSS scaling factor for attention warping is set to $\gamma = 1.0$.

2. Additional Ablation Studies

We perform a comprehensive analysis to evaluate the key hyperparameters in GR-Diffusion: the Graph Structural Similarity (GSS) threshold τ, the self-attention injection strength β_{base}, and the attention warping scale γ.

Impact of GSS Threshold τ. The threshold τ acts as the primary control mechanism of our graph-guided alignment. By varying $\tau \in [0.0, 1.0]$ as illustrated in Fig. 1, we analyze the trade-off between interaction controllability and image fidelity. We observe that while increasing τ imposes stricter structural constraints and consistently improves the HOI Detection Score, a fully active guidance ($\tau = 1.0$) introduces a minor degradation in image fidelity (FID). This suggests that enforcing spatial corrections on steps that are already structurally well-aligned may disrupt the natural coherence of the diffusion process. Consequently, we identify $\tau = 0.3$ as the optimal point, maximizing interaction accuracy while effectively preserving the generative quality of the pre-trained backbone.

Robustness of β_{base} and γ. We further examine the sensitivity of injection strength β_{base} (EEG) and warping scale γ (NAG). Our analysis reveals broad stable operating ranges where guidance is balanced with feature preservation, eliminating the need for per-sample fine-tuning. Specifically, the framework remains robust with $\beta_{base} \in [0.5, 1.5]$ and $\gamma \in [0.8, 1.2]$, consistently yielding effective control without overshadowing semantic features or introducing spatial distortions. Given this stability, we fix $\beta_{base} = 1.0$ and $\gamma = 1.0$ across all experiments.

3. More Qualitative Results

To further demonstrate the versatility and robustness of GR-Diffusion, we provide additional qualitative results focusing on two key aspects: performance on the SDXL backbone and transferability to community models.

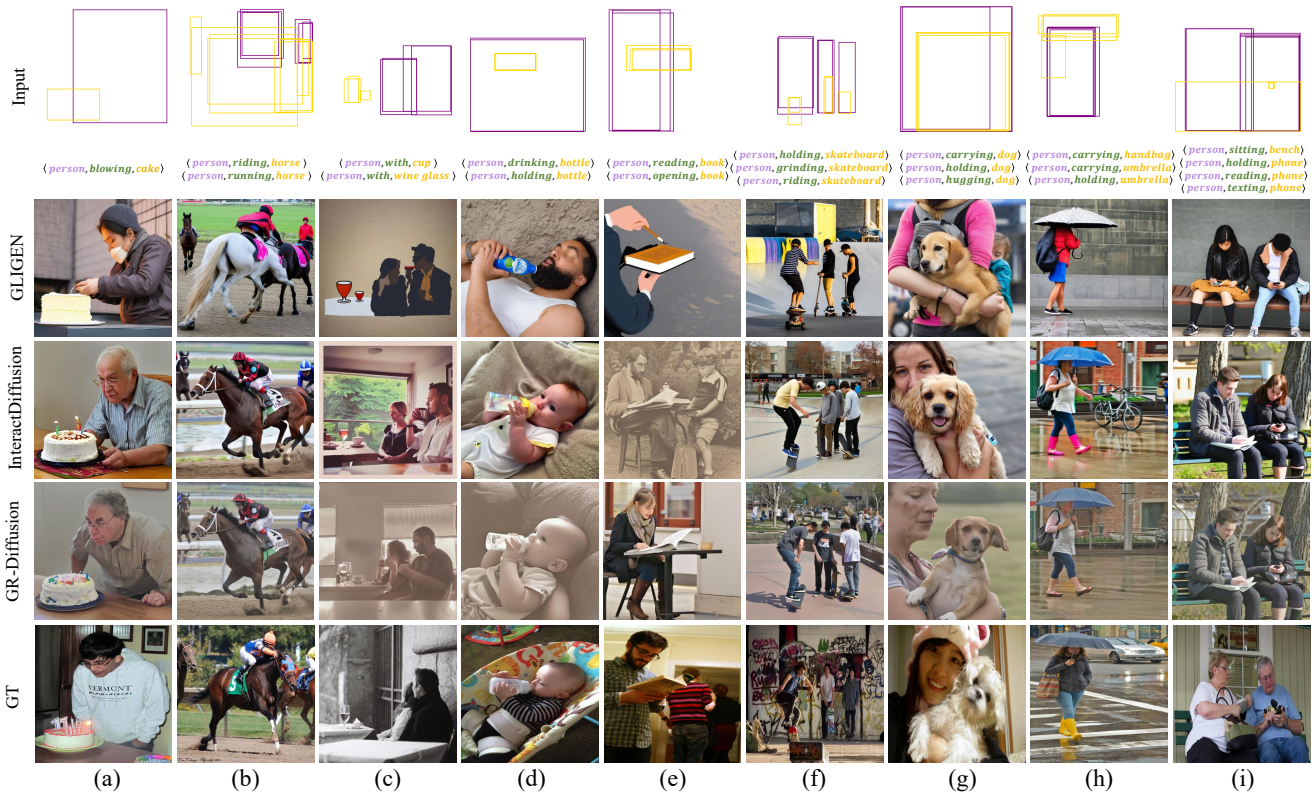


Figure 2. Qualitative comparison on the SDXL backbone. We compare GR-Diffusion with GLIGEN and InteractDiffusion. Our method effectively guides the SDXL model to follow complex interaction without compromising the high-quality generation capabilities.

3.1. Qualitative Results on SDXL Backbone

We extend our evaluation to the advanced Stable Diffusion XL (SDXL) backbone to assess performance in a larger latent space. As shown in Fig. 2, we conduct a comparative analysis against GLIGEN and InteractDiffusion adapted for SDXL. Qualitatively, while SDXL offers superior high-frequency details, baseline methods often lose spatial precision due to the expanded latent dimension. In contrast, GR-Diffusion effectively mitigates these issues through graph-guided alignment. By enforcing strict structural constraints via the NAG module, our method eliminates the spatially detached artifacts common in baselines, successfully grounding the strong semantic priors of SDXL into the specified layout. Furthermore, in complex scenarios requiring simultaneous interactions, GR-Diffusion maintains distinct attention focus for each action through the Target Scene Graph, ensuring multi-interaction coherence where baselines often suffer from semantic merging or action neglect.

3.2. Model Transferability

To verify the generalization capability of our training-free mechanism, we directly apply GR-Diffusion to various personalized community models including ChilloutMix, Cyberpunk, DreamShaper, EpicRealism, DarkSushiMix, ToonYou and Lyriel. As illustrated in Fig. 3, the results confirm that our graph-guided alignment operates in a generalized attention space, independent of low-level stylistic features. The guidance mechanism successfully injects interaction logic without disrupting the distinct artistic styles of the fine-tuned models, effectively decoupling interaction control from visual style. This robustness across diverse domains ensures that the interactions defined in the Target Scene Graph are accurately rendered regardless of whether the underlying distribution is photorealistic or stylized.

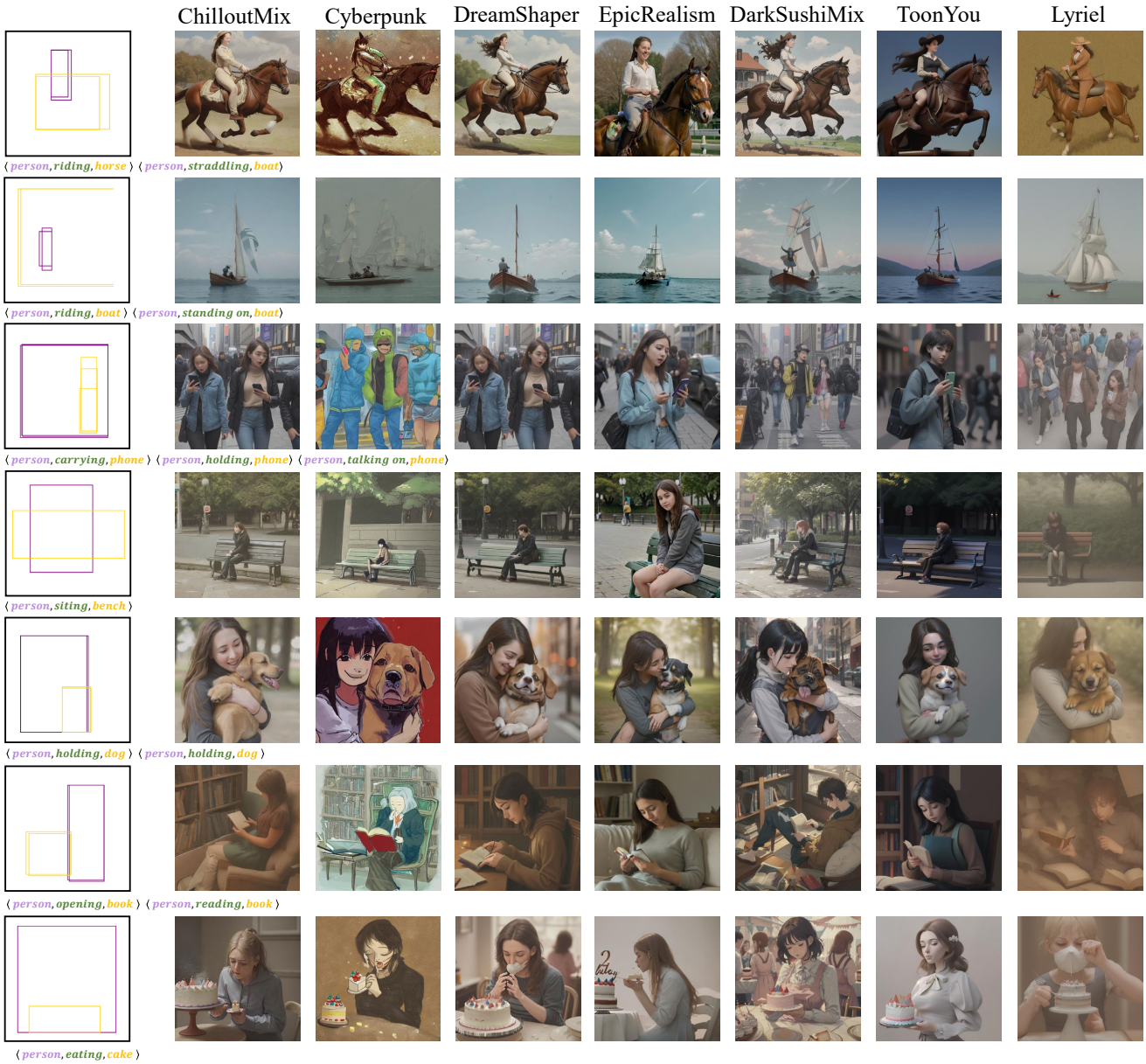


Figure 3. Demonstration of Model Transferability. We apply GR-Diffusion to various personalized community models. The results show that our method successfully injects interaction control while respecting the diverse artistic styles of the base models.