Dual Recursive Feedback on Generation and Appearance Latents for Pose-Robust Text-to-Image Diffusion

Supplementary Material

S1. Introduction

This supplementary material is intended to support the main paper. We provide comprehensive ablation studies that substantiate the chosen hyper-parameters and methodological decisions underpinning the Dual Recursive Feedback framework. Moreover, we report additional experiments with complementary evaluation metrics that rigorously quantify the contribution of each feedback mechanisms.

S2. Ablation studies on DRF

Steps for DRF. To balance both efficiency and generated image fidelity, we apply DRF during a subset of the total inference steps (50 steps). We compare the scenarios of applying DRF for 10 steps, 20 steps (ours), and all 50 steps, with results summarized in Fig. S1.

- 10 steps: While inference time (36.09s) is reduced, certain structural details (e.g., the statue's head orientation) are not sufficiently captured.
- All 50 steps: It takes 135.07s but does not yield substantially better outcomes than the 20 steps setting.
- **20 steps (Ours)**: Although slightly longer (56.87s), it preserved fine-grained structure effectively.

Figure S1. **Ablation study on the number of steps.** DRF is applied during intermediate 20 steps after the first five steps.

Number of DRF iterations. We investigate how many times DRF should be invoked at each timestep within its recurrence loop to maximize its impact. As illustrated in Fig. S2, increasing the number of DRF applications at each timestep more robustly transforms the appearance into the intended structural form. This confirms that elevating the iteration count can further strengthen the alignment between the structure and the generated image. Furthermore, we analyze the cost of these gains by measuring time as a function of the DRF iteration count, as summarized in Fig. S3. Inference latency grows almost linearly with additional feedback passes, allowing practitioners to trade fidelity for speed by selecting an iteration budget that fits their deployment constraints.

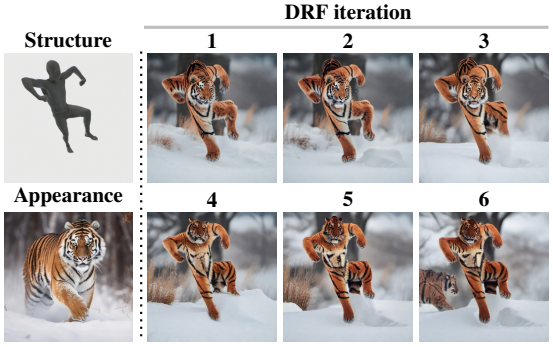

Figure S2. **Ablation study on the number of iterations for DRF.** The fusion of appearance and structure image is formed by iteration number of DRF.

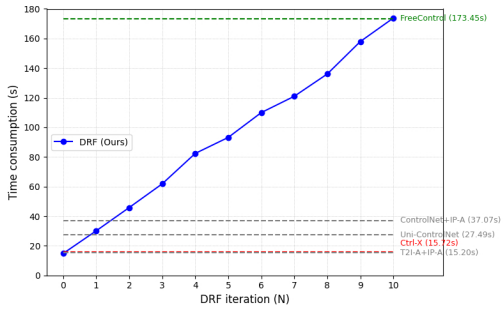

Figure S3. **DRF iteration cost.** Inference time rises roughly linearly with the number of DRF passes (blue), staying below FreeControl and matching faster baselines at low iteration counts.

Hyper-parameters of DRF. We performed a parameter sweep over λ, ρ, and k to validate our chosen settings. Across all three metrics, CLIP, Self-Sim, and DINO-I, each hyper-parameter exhibits a clear, monotonic sweet spot shown in Fig. S3). Increasing the update weight λ mildly affects CLIP and Self-Sim but raises DINO-I, peaking at λ=1.0. For the feedback balance ρ, smaller values favor appearance consistency; ρ=0.001 delivers the lowest Self-Sim and the highest DINO-I without degrading CLIP. Finally, the recursion depth K provides diminishing returns: quality improves up to K=5 and plateaus thereafter, while further iterations incur extra latency. Accordingly, we adopt $\lambda, \rho, K = (1.0, , 0.001, , 5)$ in all subsequent experiments.

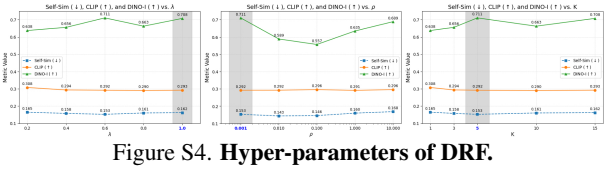

Figure S4. **Hyper-parameters of DRF.**

Weight for DRF loss. Building on the theoretical analysis in Sec. 4, we empirically verify that amplifying the generation feedback term as the recursive feedback iteration index i grows is crucial for harmonising appearance and structure features. Specifically, the exponential schedule of Eq. (10) progressively shifts the optimisation focus from low-level appearance injection in early iterations to high-level structural refinement in later ones. As shown in Fig. S5, this strategy consistently yields sharper details and more faithful pose alignment than linear or uniform weighting, confirming that a larger weight on generation feedback at higher recursion depths is the most effective way to fuse appearance and structural constraints in the final image.

Figure S5. **Comparison of weight method of DRF.** Exponential weight schedule determines the optimized weight for Generation feedback.

S3. Additional Experiments

To more rigorously confirm that DRF preserves both appearance and structure, we add two complementary metrics. ArcFace [3] similarity quantifies identity retention by measuring the cosine distance between face embeddings of the generated image and the appearance reference, while SAM [13]-IoU assesses pose fidelity by comparing structure-aware mask segmentations of the structure reference with those of each synthesis. As reported in Tab. S1, DRF achieves the highest ArcFace similarity, demonstrating superior identity preservation alongside accurate pose alignment. While DRF matches Ctrl-X in IoU, it delivers superior overall image quality.

	DRF (Ours)	Ctrl-X	FreeControl	Uni-ControlNet	ControlNet + IP-Adapter	T2I-Adapter + IP-Adapter
ArcFace ↓	**0.6221**	0.6497	0.7043	0.6961	0.7089	0.6469
IoU ↑	**0.8048**	0.8205	0.7984	0.6983	0.7498	0.4767

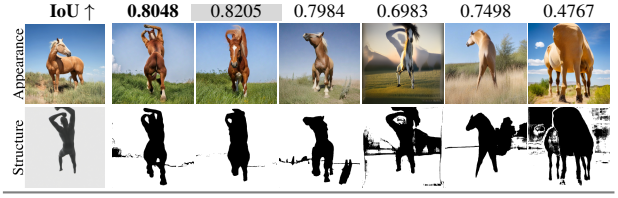

Table S1. **Additional experiments of DRF.** DRF attains the lowest ArcFace [3] (↓) and strong SAM [13] (↑), visibly fusing appearance and structure more faithfully than baseline models.