TAUE: Training-free Noise Transplant and Cultivation Diffusion Model

Supplementary Material

8. Ablation Study

We analyze the effect of key hyperparameters through controlled ablations: the denoising crop ratio r_{crop}, the attention threshold τ_A, and the background activation threshold τ_{BG}. All experiments are conducted under the same configuration as in the main paper.

8.1. Effect of Crop Ratio r_{crop}

Table 4 shows that smaller crop ratios (e.g., $r_{\text{crop}} = 0.3$) lead to poor overall quality with a high FID of 92.32 and degraded reconstruction metrics ($\text{PSNR}_{\text{fg}} = 15.59$, $\text{SSIM}_{\text{fg}} = 0.81$). Increasing the ratio gradually improves both foreground and background reconstruction, with $r_{\text{crop}} = 0.4$ achieving the lowest FID (58.81) and high CLIP alignment (CLIP-I$= 0.642$, CLIP-S$= 0.321$). For larger values ($r_{\text{crop}} \geq 0.6$), PSNR and SSIM continue to increase, but FID slightly worsens (up to 63.52), indicating potential overfitting to the foreground. Overall, the mid-point configuration $r_{\text{crop}} = 0.5$ provides the best trade-off between structure preservation and global coherence.

8.2. Effect of Attention Threshold τ_A

As reported in Table 5, FID remains stable across different τ_A values (60.06–60.96). Lower thresholds (0.1–0.2) yield better reconstruction performance (PSNRfg up to 21.6 and SSIMfg $= 0.91$), while higher values (0.4–0.5) slightly improve CLIP alignment (CLIP-I$= 0.647$) but degrade pixel-wise quality (PSNRfg $= 19.1$, LPIPSfg $= 0.151$). The default setting $\tau_A = 0.3$ achieves balanced performance across all metrics, maintaining stable FID and CLIP scores while avoiding under- or over-activation of attention.

8.3. Effect of Background Threshold τ_{BG}

The results in Table 6 show that lowering τ_{BG} to 0.6 achieves the best FID (58.75) and the highest CLIP-I (0.648), suggesting improved overall quality. As τ_{BG} increases, both foreground and background reconstruction metrics (PSNR and SSIM) steadily improve, and perceptual distortion (LPIPS) decreases slightly. However, FID gradually worsens beyond $\tau_{\text{BG}} = 1.0$, indicating excessive smoothing or loss of separation. Thus, the default $\tau_{\text{BG}} = 1.0$ offers a stable balance between layer separation and contextual harmony.

9. Additional Results

Figure 7 provides additional qualitative examples across diverse categories. For each case, we show the *Foreground Object*, the *Background*, and the final *Composite*. TAUE maintains semantic and spatial consistency among all three outputs while preserving fine object details and realistic scene context.

Table 4. **Ablation on the crop ratio r_{crop}.** Results show that $r_{\text{crop}} = 0.5$ achieves the best balance between reconstruction and overall visual quality.

Setting	FID↓	CLIP-I↑	CLIP-S↑	PSNR_{fg} ↑	PSNR_{bg} ↑	SSIM_{fg} ↑	SSIM_{bg} ↑	LPIPS_{fg} ↓	LPIPS_{bg} ↓
$r_{\text{crop}} = 0.3$	92.32	0.600	0.305	15.59	20.73	0.813	0.793	0.215	0.216
$r_{\text{crop}} = 0.4$	58.81	0.642	0.321	19.10	23.87	0.872	0.858	0.166	0.152
$r_{\text{crop}} = 0.5$ (default)	60.53	0.646	0.323	20.46	25.86	0.901	0.895	0.137	0.106
$r_{\text{crop}} = 0.6$	62.02	0.644	0.323	21.01	27.19	0.916	0.917	0.120	0.077
$r_{\text{crop}} = 0.7$	63.52	0.643	0.322	21.19	28.05	0.922	0.930	0.113	0.061

Table 5. **Ablation on the attention threshold τ_A.** Performance remains stable around $\tau_A = 0.3$, which we use as default.

Setting	FID↓	CLIP-I↑	CLIP-S↑	PSNR_{fg} ↑	PSNR_{bg} ↑	SSIM_{fg} ↑	SSIM_{bg} ↑	LPIPS_{fg} ↓	LPIPS_{bg} ↓
$\tau_A = 0.1$	60.96	0.645	0.323	21.62	25.93	0.911	0.896	0.126	0.104
$\tau_A = 0.2$	60.29	0.646	0.323	21.17	25.88	0.907	0.895	0.131	0.105
$\tau_A = 0.3$ (default)	60.53	0.646	0.323	20.46	25.86	0.901	0.895	0.137	0.106
$\tau_A = 0.4$	60.56	0.646	0.323	19.79	25.76	0.895	0.894	0.144	0.107
$\tau_A = 0.5$	60.06	0.647	0.324	19.10	25.75	0.888	0.893	0.151	0.108

Table 6. **Ablation on the background threshold τ_{BG}.** The default $\tau_{\text{BG}} = 1.0$ offers the best balance between separation and harmony.

Setting	FID↓	CLIP-I↑	CLIP-S↑	PSNR_{fg} ↑	PSNR_{bg} ↑	SSIM_{fg} ↑	SSIM_{bg} ↑	LPIPS_{fg} ↓	LPIPS_{bg} ↓
$\tau_{\text{BG}} = 0.6$	58.75	0.648	0.323	19.59	25.57	0.885	0.893	0.154	0.108
$\tau_{\text{BG}} = 0.8$	60.26	0.647	0.323	20.08	25.75	0.894	0.894	0.145	0.106
$\tau_{\text{BG}} = 1.0$ (default)	60.53	0.646	0.323	20.46	25.86	0.901	0.895	0.137	0.106
$\tau_{\text{BG}} = 1.2$	61.00	0.645	0.323	20.76	25.96	0.907	0.896	0.130	0.105
$\tau_{\text{BG}} = 1.4$	61.58	0.644	0.323	20.97	26.09	0.911	0.897	0.125	0.104

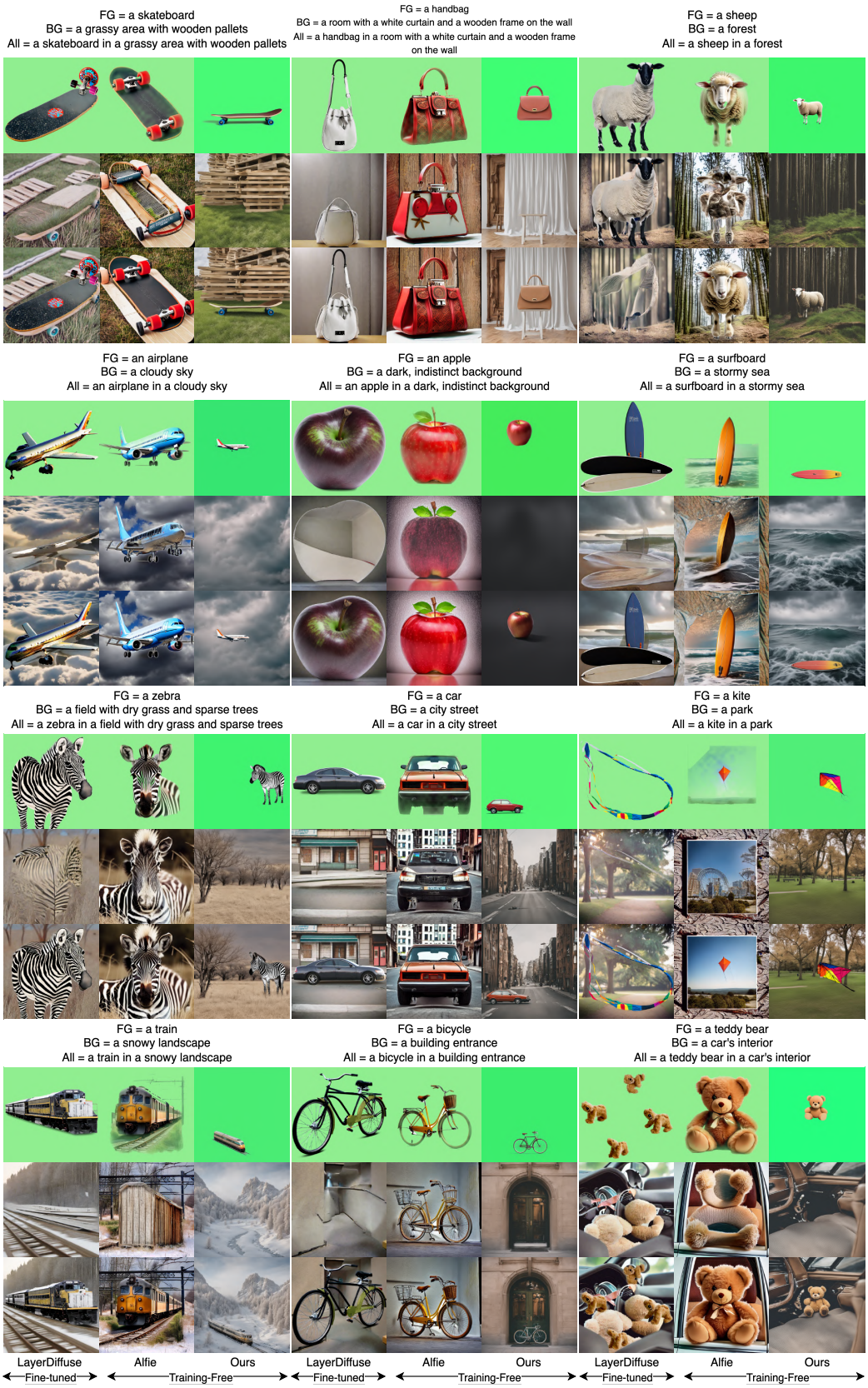

Figure 7. **More qualitative results.** Each triplet shows foreground, background, and composite. Across vehicles, animals, man-made objects, and natural scenes, TAUE produces crisp foregrounds, harmonized backgrounds, and composites with consistent geometry, lighting, and shadows, without fine-tuning or inpainting.