SPECTRA: SOURCE-PRESERVING EDITING WITH CROSS-SCALE TRANSFORM AND REPRESENTATION ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-guided image editing must follow the target instruction while retaining source information that should remain unchanged. Existing training-free methods use feature injection, attention control, or predefined frequency regions, but typically apply coarse or fixed preservation policies. We study source trajectories and find that preservation evidence differs across transform scales, is non-uniform across orientations within a scale, and changes over sampling time. SPECTRA therefore defines preservation over scale, direction, space, and time. At each native step, source-only multiscale transform energies determine threshold-free scale–direction weights, while target-token attention supplies spatial evidence through Semantic Structure Mass-Odds (SSMO). The fused clean prediction is returned to the backbone's native DDIM or FlowMatch scheduler, enabling one interface across Stable Diffusion and transformer flow backbones; appearance-preservation tasks optionally use decoder-cycle residual chroma injection (DCRCI). Across two appearance-changing and two appearance-preserving benchmarks, SPECTRA shifts the preservation–editability operating point toward stronger source retention. Controlled analysis separates a strong cross-scale effect from a smaller conditional directional refinement: directional benefit tends to increase with source anisotropy, while a shuffled-direction control shows that structural preservation depends on matching source-derived weights to their corresponding bands.

1 INTRODUCTION

Text-guided image editing must change a source according to a target prompt without unnecessarily altering geometry, identity, background, or appearance. Training-free methods address this preservation–editability tension through inversion/attention control (Hertz et al., 2023; Mokady et al., 2023), feature injection (Tumanyan et al., 2023), spatial guidance (Mo et al., 2024), or feed-forward structure/appearance control (Lin et al., 2024). Frequency-aware methods instead manipulate spectral content: FBSDiff substitutes frequency bands (Gao & Liu, 2024), FreeDiff progressively truncates frequencies (Wu et al., 2024), FSI-Edit injects high-frequency reconstruction features (Yang et al., 2025), and FGD separates low-frequency style from high-frequency structure guidance (Gao et al., 2025). These works establish spectral control as useful, but leave a finer question: *how should preservation be allocated when source information is unevenly distributed inside the representation?*

Our starting observation is that source-preservation evidence is not uniform at a fixed frequency magnitude or even at a fixed transform scale. In a multiscale directional representation, different scales carry complementary source information, while the energy distribution over orientations within a scale can be strongly anisotropic. Moreover, this directional profile evolves along the native denoising or flow trajectory. The resulting preservation problem is therefore structured along four axes: **scale**, **direction**, **space**, and **time**. A fixed band selection cannot express this structure, and a scale-only policy ignores variation among directions that share the same scale.

Figure 1 summarizes the per-step policy. A matched multiscale directional transform decomposes source-reference and target-conditioned clean predictions. Source coefficients alone determine scale and within-scale directional importance, without learned strengths, top-k selection, frequency cutoffs, or direction-specific time windows. Native target attention instead answers *where* this source prior

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

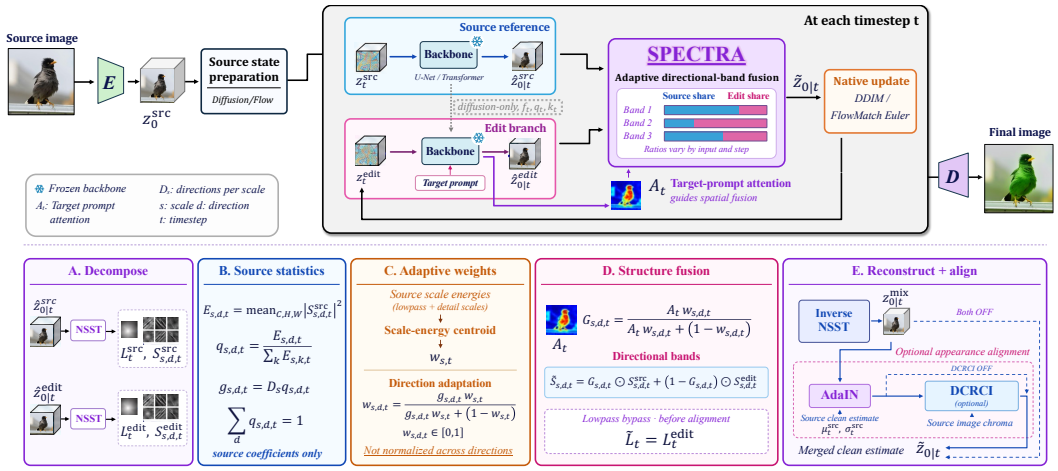


Figure 1: Overview of SPECTRA. Source-reference and target-conditioned clean predictions are decomposed at each native step. Source energies determine threshold-free scale–direction weights (*what/when*); target-token attention supplies spatial evidence (*where*) through SSMO. The fused clean prediction is reconstructed, optionally aligned by AdaIN/DCRCI, and returned to the native DDIM or FlowMatch scheduler. Self Attn-Conv is U-Net-specific; clean-prediction fusion is shared across backbones.

should act through a continuous SSMO gate. After reconstruction, the fused clean prediction is returned to the native scheduler. This interface is shared by DDIM-based Stable Diffusion (Rombach et al., 2022; Song et al., 2021) and FlowMatch-based transformers (Lipman et al., 2023; Esser et al., 2024; Black Forest Labs, 2024).

The results reveal a hierarchy. Increasing transform depth from one to three scales consistently improves source-preservation metrics, establishing cross-scale complementarity as the dominant effect. Direction-aware allocation is a smaller conditional refinement: its paired benefit tends to increase with source anisotropy, and a shuffled-direction control shows that correct band assignment gives the best DINO and Edge preservation. Thus the evidence links a measurable representation property and its assignment to when directional refinement is useful, without treating the anisotropy association as causal.

Our contributions are:

- **Cross-scale directional heterogeneity.** We characterize source-preservation information as heterogeneous across transform scales, orientations within each scale, and sampling time. Scale-depth and trajectory analyses establish cross-scale complementarity and unequal, time-varying directional importance.
- **Threshold-free adaptive scale–direction preservation.** Source transform energies directly determine scale- and direction-specific weights without hand-tuned thresholds or subband selection. The rule reduces exactly to Scale-only under uniform directional energy, while a shuffled-direction control supports the importance of correct band assignment.
- **Semantic spatial gating and diffusion-internal color correction.** SSMO couples source transform confidence with target-aware native attention to instantiate *where* preservation acts. For appearance-preservation tasks, DCRCI provides targeted chroma correction; an isolated AdaIN-versus-AdaIN+DCRCI ablation improves CSD and ΔE_{00} while exposing its metric trade-off. The same clean-prediction interface applies across diffusion and flow backbones.

2 Related Work

Training-free source-guided editing. Prompt-to-Prompt and Null-text inversion control attention or inversion states (Hertz et al., 2023; Mokady et al., 2023); PnP injects source spatial features/self-

attention (Tumanyan et al., 2023); PnP Inversion introduces PIE-Bench (Ju et al., 2024); FreeControl and Ctrl-X provide training-free spatial or structure/appearance control (Mo et al., 2024; Lin et al., 2024); LEDITS++ and DDPM-Edit offer localized editing and edit-friendly inversion (Brack et al., 2024; Huberman-Spiegelglas et al., 2024). These methods preserve source information through inversion, feature reuse, attention, or spatial constraints. SPECTRA instead studies how preservation should be allocated within a multiscale directional representation as sampling evolves.

Frequency-aware and directional control. Frequency-conditioned editing includes learned DCT branches (Gao et al., 2024), band substitution in FBSDiff/FBSDiff++ (Gao & Liu, 2024; Gao & Jia, 2026), progressive truncation in FreeDiff (Wu et al., 2024), high-frequency suppression/injection in FlexiEdit and FSI-Edit (Koo et al., 2024; Yang et al., 2025), frequency-interactive attention in FIA-Edit (Yang et al., 2026), and low-/high-frequency style/structure guidance in FGD (Gao et al., 2025). Our claim is not that these representations lack orientation information; their control policies are primarily organized by predefined bands, radial regions, or fixed frequency roles. SPECTRA instead treats preservation as state-dependent allocation: the source determines scale/direction weights at each step, and native target attention determines where they act. Thus the contribution is not a new transform, but an adaptive preservation policy. NSST provides the default anisotropic multiscale representation (Easley et al., 2008); Appendix G tests DTCWT, DWT, and NSCT.

3 SPECTRA

3.1 PROBLEM SETUP AND CLEAN-PREDICTION INTERFACE

Let x^{src} be a source image and p^{tgt} a target editing prompt. A frozen VAE maps the source to a clean latent z_0^{src}. At native sampling step t, the backbone produces a source-reference state z_t^{src} and an evolving target edit state z_t^{edit}. The source branch is conditioned with empty text; the target prompt conditions only the edit branch. The scheduler-specific model outputs are converted to clean latent predictions

$$\hat{z}_{0,t}^{\mathrm{src}}, \qquad \hat{z}_{0,t}^{\mathrm{edit}}. \tag{1}$$

SPECTRA intervenes only on these clean predictions. It does not replace the denoiser, transformer, or scheduler.

For SD1.5/SD2.1, an empty-text DDIM inversion provides a source trajectory matched to the 50-step generation schedule. For SD3 and FLUX.1-dev, the source state is reconstructed at the selected native image-to-image timestep using the pipeline's scheduler-native noising operation and the same seed-fixed noise realization as the edit initialization. This separation preserves each backbone's native state construction while exposing the common clean-prediction interface used below.

Figure 1 gives the complete per-step data flow. The key abstraction is the clean-prediction interface: backbone-specific state construction remains native, while the preservation policy acts on spatial clean latents before the scheduler update.

3.2 SOURCE-CONDITIONED SCALE AND DIRECTION ALLOCATION

Matched multiscale directional analysis. Let $\mathcal{T}$ denote a transform with terminal low-pass coefficient L and directional detail coefficients $S_{s,d}$ at scale s and direction d. We apply the same analysis to the two clean predictions:

$$(L_t^{\mathrm{src}}, \{S_{s,d,t}^{\mathrm{src}}\}) = \mathcal{T}(\hat{z}_{0,t}^{\mathrm{src}}), \tag{2}$$

$$(L_t^{\mathrm{edit}}, \{S_{s,d,t}^{\mathrm{edit}}\}) = \mathcal{T}(\hat{z}_{0,t}^{\mathrm{edit}}). \tag{3}$$

Our default is a three-scale NSST with the `dmaxflat4` filter and matched analysis/inverse reconstruction. The terminal low-pass is not treated as an additional directional scale.

Source-only directional statistics. All adaptation statistics are computed from source coefficients. For scale s with D_s native directions,

$$E_{s,d,t} = \mathrm{mean}_{c,h,w} \left| S_{s,d,t}^{\mathrm{src}}(c,h,w) \right|^2, \qquad q_{s,d,t} = \frac{E_{s,d,t}}{\sum_{j=1}^{D_s} E_{s,j,t} + \varepsilon}. \tag{4}$$

When the total energy is numerically zero, the implementation uses a uniform distribution. Equal directional energies are also mapped exactly to uniform weights. We define the mean-preserving relative directional importance

$$g_{s,d,t} = D_s q_{s,d,t}. \tag{5}$$

Thus $g_{s,d,t} = 1$ under a directionally uniform source scale.

Scale scheduler. Before directional refinement, we compute a source-energy scale allocation. For the three-scale NSST, let the normalized scale energies, including the terminal low-pass, have coordinates $c = (0, 1, \frac{1}{2}, \frac{1}{4})$. Their energy centroid gives

$$\mathrm{centroid}_t = \sum_s \bar{E}_{s,t} c_s, \qquad \rho_t = \mathrm{clip}(1 - \mathrm{centroid}_t, 0, 1), \tag{6}$$

followed by the complementary low-frequency scale weights

$$w_{0,t}^{\mathrm{LF}} = 1, \qquad w_{s,t}^{\mathrm{LF}} = \frac{\rho_t^2}{\rho_t^2 + c_s^2 + \varepsilon}. \tag{7}$$

The final configuration uses this LF branch; a separate HF fusion branch is disabled.

Threshold-free directional refinement. The Scale-only control broadcasts $w_{s,t}^{\mathrm{LF}}$ to every direction. Scale+Direction instead applies parameter-free odds modulation:

$$\boxed{w_{s,d,t} = \frac{g_{s,d,t} w_{s,t}^{\mathrm{LF}}}{g_{s,d,t} w_{s,t}^{\mathrm{LF}} + \left(1 - w_{s,t}^{\mathrm{LF}}\right)}.} \tag{8}$$

There is no learned direction strength, frequency threshold, top-k, exponent, temperature, or direction-specific start/end schedule. Importantly, if $q_{s,d,t} = 1/D_s$, then $g_{s,d,t} = 1$ and Eq. equation 8 reduces exactly to $w_{s,d,t} = w_{s,t}^{\mathrm{LF}}$. Directional adaptation therefore introduces no preference when the source scale itself is isotropic.

3.3 Semantic Structure Mass-Odds: Where to Preserve

The transform controller answers which source components are trusted, but not where they should act spatially. We derive a target-semantic map $A_t(p) \in [0, 1]$ from the native target-conditioned attention at each step. For U-Net backbones, A_t is the average cross-attention probability mass assigned to non-special target tokens across selected layers and heads. For SD3/FLUX.1-dev, we compute the analogous image-query probability mass over target-text tokens in native joint attention. Maps from different resolutions are aligned before aggregation. A_t is probability mass, not a learned segmentation mask and not min–max normalized.

We combine target-semantic mass with the source prior using Semantic Structure Mass-Odds:

$$\boxed{G_{s,d,t}(p) = \frac{A_t(p)\, w_{s,d,t}}{A_t(p)\, w_{s,d,t} + (1 - w_{s,d,t})}.} \tag{9}$$

The implementation uses numerically stable endpoint handling so $w = 0$ gives $G = 0$ and $w = 1$ gives $G = 1$. A large $A_t(p)$ does not independently force source copying: it can increase the source share only when the source-derived prior $w_{s,d,t}$ is already non-negligible. We use target-token mass to identify locations participating in target-conditioned semantic structure, while the source controller determines whether the corresponding directional component is trusted. The directional coefficients are then fused pointwise,

$$\widetilde{S}_{s,d,t}(p) = G_{s,d,t}(p) S_{s,d,t}^{\mathrm{src}}(p) + \big(1 - G_{s,d,t}(p)\big) S_{s,d,t}^{\mathrm{edit}}(p). \tag{10}$$

We deliberately keep the terminal low-pass from the edit branch,

$$\widetilde{L}_t = L_t^{\mathrm{edit}}, \qquad \widetilde{z}_{0,t} = \mathcal{T}^{-1}\Big(\widetilde{L}_t, \{\widetilde{S}_{s,d,t}\}\Big). \tag{11}$$

This yields the paper-wide interpretation

$$\underbrace{w_{s,d,t}}_{\text{WHAT + WHEN}} \times \underbrace{A_t(p)}_{\text{WHERE}} \longrightarrow \underbrace{G_{s,d,t}(p)}_{\text{source-preservation gate}}. \tag{12}$$

3.4 Task-Dependent Appearance Alignment

Structure preservation and appearance preservation require different controls. For appearance-changing tasks, AdaIN and DCRCI are disabled so the requested color/style change is not pulled back toward the source. For appearance-preservation tasks, channel-wise AdaIN (Huang & Belongie, 2017) first aligns the fused clean latent with the source clean prediction. DCRCI then performs a deterministic VAE decode–reencode cycle inside each guided step: it keeps the current lightness, inserts source CIELAB chroma at identity coordinates, and injects only the resulting latent cycle residual with unit weight. Its purpose is specifically to correct source-relative chroma drift, not to optimize every preservation metric. Appendix A gives the exact construction, and Appendix H isolates its effect from AdaIN.

3.5 Return to the Native Scheduler

The final fused clean prediction is converted back to the update quantity expected by the installed scheduler: epsilon (or the configured equivalent) for DDIM and velocity for FlowMatch. The unmodified native scheduler then advances the edit state. Thus backbone-specific source construction, attention extraction, and scheduler semantics remain native, while the transform controller, SSMO, reconstruction, and optional appearance alignment share one clean-prediction interface. Exact DDIM and FlowMatch conversions are given in Appendix A.

4 Experiments

4.1 Experimental Setup

Tasks and datasets. We evaluate two objectives. **Source-guided appearance changing** requires preserving geometry and identity while changing requested appearance; we use the official PIE-Bench Color40 subset ($n = 40$) from PnP Inversion (Ju et al., 2024) and an ImageNet-R RGB-fixed variant ($n = 87$ source–prompt pairs). The latter keeps the 29 ImageNet-R source images and first three official TI2I prompts fixed, then deterministically inserts red, blue, and green target colors into prompt indices 0, 1, and 2, respectively. It is a prompt-only derived subset, not an official ImageNet-R benchmark. **Source-guided appearance preservation** uses PIE-Bench Style-9 ($n = 80$) and the first three official ImageNet-R-TI2I prompts for each of the same 29 mapped sources ($n = 87$ pairs). ImageNet-R originates from Hendrycks et al. (2021). All frozen manifests are method-independent.

Metrics. For both tasks, structural preservation is measured by DINO key self-similarity distance (Caron et al., 2021) and HED-edge similarity (Xie & Tu, 2015; Radford et al., 2021), while target adherence is measured by CLIP image–text cosine similarity (CLIP-T) (Radford et al., 2021). For readability, **DINO distances are multiplied by** 100 **in all reported tables**. Appearance-preservation tables additionally report CSD style similarity (Somepalli et al., 2024), mean CIEDE2000 color difference ΔE_{00} (Sharma et al., 2005), and LPIPS-AlexNet (Zhang et al., 2018). CSD, ΔE_{00}, and LPIPS are deliberately excluded from appearance-changing rankings because they reward retaining source appearance even when the prompt requests changing it.

Backbones and implementation. The primary benchmark uses SD2.1 at 512^2 with 50 DDIM steps, CFG 7.5, and empty-text source inversion. Its U-Net-specific source self-attention Q/K injection (first 80% of steps) and convolutional feature injection (all steps) are denoted jointly as Self Attn-Conv. The final NSST configuration uses three scales, Scale+Direction adaptation, SSMO at every selected native step, LF scheduling, and HF fusion off. AdaIN/DCRCI are off for appearance changing and on with unit strength for appearance preservation. Cross-backbone experiments additionally use SD1.5, SD3-Medium at 1024^2 with 50 native FlowMatch steps and CFG 7.0, and FLUX.1-dev at 1024^2 with 28 native FlowMatch steps and guidance 3.5. SD3 and FLUX.1-dev use their native image-to-image strength contract and do not emulate Self Attn-Conv. All reported runs use seed 1. Experiments were executed on RTX 3090 GPUs; multi-GPU use only shards independent samples or configurations.

Baselines. We compare against PnP (Tumanyan et al., 2023), FSI-Edit (Yang et al., 2025), FreeDiff (Wu et al., 2024), FIA-Edit (Yang et al., 2026), Ctrl-X (Lin et al., 2024), FBSDiff (Gao & Liu, 2024), FBSDiff++ (Gao & Jia, 2026), DDPM-Inversion/Edit (Huberman-Spiegelglas et al., 2024), and

Table 1: Source-guided appearance changing on PIE-Bench Color40 and ImageNet-R RGB-fixed. Lower DINO is better; higher Edge and CLIP-T are better. Bold denotes the best value within each dataset.

	PIE-Bench Color40 ($n = 40$)			ImageNet-R RGB-fixed ($n = 87$)		
Method	DINO ↓	Edge ↑	CLIP-T ↑	DINO ↓	Edge ↑	CLIP-T ↑
PnP (Tumanyan et al., 2023)	0.3635	0.933883	0.320520	0.4692	0.923408	0.320213
FSI-Edit (Yang et al., 2025)	0.2529	0.954036	0.308202	0.3047	0.946249	0.289174
FreeDiff (Wu et al., 2024)	0.4014	0.912496	0.305592	0.5519	0.901492	0.300771
FIA-Edit (Yang et al., 2026)	0.1776	0.958813	0.313540	0.2710	0.942602	0.299892
Ctrl-X (Lin et al., 2024)	0.4487	0.926016	0.312394	0.4841	0.930784	0.296206
FBSDiff (Gao & Liu, 2024)	0.5948	0.909538	**0.328082**	0.6283	0.910846	0.317961
FBSDiff++ (Gao & Jia, 2026)	0.5828	0.920354	0.323535	0.6576	0.908733	0.322893
DDPM-Inversion (Huberman-Spiegelglas et al., 2024)	0.3012	0.940370	0.317586	0.4009	0.919397	0.316845
LEDITS++ (Brack et al., 2024)	0.4208	0.914714	0.325675	0.6076	0.899558	**0.333123**
SPECTRA (ours)	**0.1483**	**0.970599**	0.310811	**0.1977**	**0.965040**	0.302801

LEDITS++ (Brack et al., 2024). Competing methods use their official or locked local configurations; all generated outputs are then evaluated by one common paired evaluator. These comparisons therefore measure practical benchmark operating points, while the isolated contribution of direction adaptation is evaluated separately with matched ablations in Sec. 4.6.

4.2 Comparison with Existing Methods

Table 1 evaluates appearance-changing edits. SPECTRA has the lowest DINO discrepancy and highest Edge similarity on both Color40 and ImageNet-R RGB-fixed, indicating stronger preservation of source geometry and boundaries. The trade-off is visible in CLIP-T: several baselines obtain higher target-text similarity, especially on ImageNet-R RGB-fixed. The result should therefore be read as a preservation–editability operating point rather than uniform dominance over every objective.

Table 2 evaluates the harder preservation objective with six complementary metrics. On both Style-9 and ImageNet-R Official, SPECTRA obtains the strongest DINO, Edge, CSD, ΔE_{00}, and LPIPS values among the evaluated methods. Its CLIP-T is lower than several baselines. This pattern is consistent across the two datasets: SPECTRA moves the operating point toward source structure, color, and perceptual fidelity while retaining non-trivial target alignment. The qualitative examples in Fig. 2 examine whether this preservation comes from visually ignoring the requested edit.

4.3 Qualitative Comparison

The appearance-changing rows in Fig. 2 illustrate the central failure mode behind the quantitative trade-off: methods can obtain stronger prompt alignment by changing broader image content than the requested attribute. Conversely, overly strong source locking can make the edit weak. In the appearance-preservation rows, SPECTRA retains local geometry and source appearance while changing the requested rendering style. We therefore use the full task-specific metric set rather than a single aggregate score.

4.4 Transfer of the Preservation Interface Across Backbones

Table 3 tests whether the clean-prediction preservation interface is tied to a U-Net diffusion architecture. SD1.5/SD2.1 use Self Attn-Conv in addition to the shared controller; SD3 and FLUX.1-dev instead use native joint attention and FlowMatch without emulating the U-Net injection path. Across all four backbones, SPECTRA moves the operating point toward lower DINO and higher Edge; for appearance preservation, CSD, ΔE_{00}, and LPIPS also improve while CLIP-T decreases. These comparisons are not intended to attribute the full Vanilla-to-SPECTRA difference to the scale–direction controller: SD3/FLUX.1-dev use a selected strength of 0.4 for SPECTRA and 0.6 for Vanilla, and the U-Net variants additionally contain Self Attn-Conv. Their narrower purpose is to establish that the same clean-prediction preservation interface can be instantiated across substantially different native pipelines without replacing their denoisers or schedulers. Controlled component effects are evaluated separately below.

Table 2: Source-guided appearance preservation. Higher Edge, CLIP-T, and CSD are better; lower DINO, ΔE_{00}, and LPIPS are better. Bold denotes the best value in each panel.

(a) PIE-Bench Style-9 ($n = 80$)

Method	DINO ↓	Edge ↑	CLIP-T ↑	CSD ↑	ΔE_{00} ↓	LPIPS ↓
PnP (Tumanyan et al., 2023)	0.4650	0.931850	**0.339857**	0.585640	11.558568	0.364959
FSI-Edit (Yang et al., 2025)	0.2739	0.953318	0.313164	0.785124	4.414129	0.190941
FreeDiff (Wu et al., 2024)	0.4677	0.917802	0.313238	0.710960	8.046738	0.325889
FIA-Edit (Yang et al., 2026)	0.2154	0.953988	0.323922	0.771782	4.867489	0.170106
Ctrl-X (Lin et al., 2024)	0.4685	0.935669	0.314789	0.651415	12.728090	0.406545
FBSDiff (Gao & Liu, 2024)	0.5637	0.914867	0.327249	0.514609	24.627120	0.527509
FBSDiff++ (Gao & Jia, 2026)	0.5813	0.916741	0.328798	0.499506	25.948383	0.544150
DDPM-Inversion (Huberman-Spiegelglas et al., 2024)	0.3356	0.938427	0.325395	0.711654	7.714913	0.262373
LEDITS++ (Brack et al., 2024)	0.4467	0.920429	0.328842	0.664181	10.832525	0.333747
SPECTRA (ours)	**0.1535**	**0.971377**	0.303636	**0.830143**	**4.306591**	**0.161643**

(b) ImageNet-R Official ($n = 87$)

Method	DINO ↓	Edge ↑	CLIP-T ↑	CSD ↑	ΔE_{00} ↓	LPIPS ↓
PnP (Tumanyan et al., 2023)	0.4711	0.926492	0.306057	0.481649	10.966128	0.366987
FSI-Edit (Yang et al., 2025)	0.3028	0.945274	0.296027	0.677226	4.964035	0.214145
FreeDiff (Wu et al., 2024)	0.5765	0.906282	0.308439	0.567379	9.305818	0.380600
FIA-Edit (Yang et al., 2026)	0.2708	0.940538	0.302776	0.694631	5.842714	0.201335
Ctrl-X (Lin et al., 2024)	0.4783	0.930707	0.291945	0.566347	12.892351	0.421791
FBSDiff (Gao & Liu, 2024)	0.6058	0.914581	0.311865	0.390922	28.879362	0.571614
FBSDiff++ (Gao & Jia, 2026)	0.6359	0.911147	0.313732	0.381843	30.209286	0.599476
DDPM-Inversion (Huberman-Spiegelglas et al., 2024)	0.3989	0.918887	0.313668	0.581791	9.179937	0.311198
LEDITS++ (Brack et al., 2024)	0.5742	0.897639	**0.317290**	0.517529	13.180767	0.416248
SPECTRA (ours)	**0.1744**	**0.968612**	0.288178	**0.753732**	**4.578329**	**0.192274**

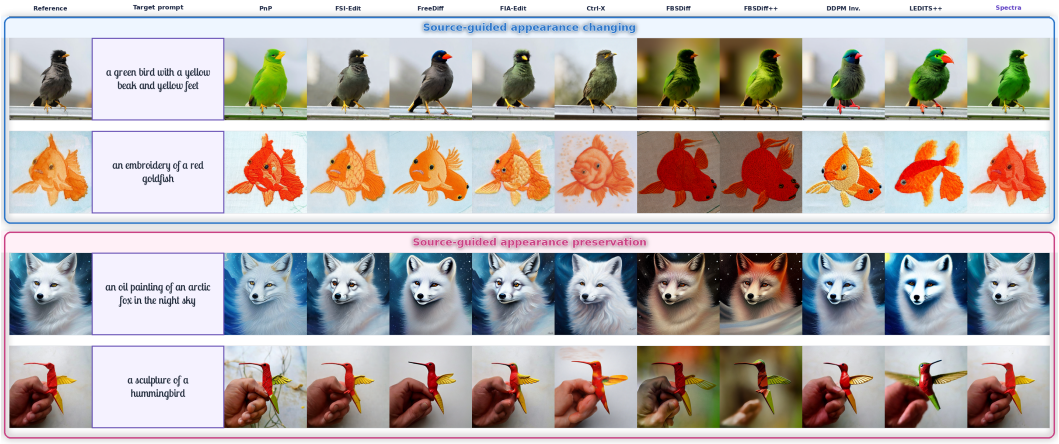


Figure 2: Qualitative comparison against the evaluated baselines for source-guided appearance changing (top block) and appearance preservation (bottom block). The examples are selected to make both sides of the objective visible: the target attribute/style should change, while source geometry and non-target content should remain stable. Across these cases, SPECTRA exhibits less geometric drift and stronger source-local preservation; the comparison also makes clear that preservation should be judged jointly with whether the requested edit is actually expressed.

4.5 IS PRESERVATION COMPLEMENTARY ACROSS SCALES?

If preservation evidence is genuinely distributed across scales, progressively exposing additional transform scales should improve source retention even when directional adaptation is held fixed. Table 4 varies the complete NSST decomposition depth under the final preservation setting. Increasing J from 1 to 3 monotonically improves DINO, Edge, CSD, ΔE_{00}, and LPIPS, while CLIP-T changes modestly in the opposite direction and is highest at $J = 1$. The ablation therefore supports cross-scale complementarity while making the operating-point trade-off explicit: additional scales move the result toward source fidelity at a small target-alignment cost.

Table 3: Cross-backbone operating-point comparison on PIE-Bench. Bold marks the better Vanilla/SPECTRA value within each backbone pair and metric. SD3 and FLUX.1-dev use strength 0.4 for SPECTRA and 0.6 for Vanilla, so those pairs establish transfer rather than a controlled component effect.

	Appearance changing: Color40 ($n = 40$)			Appearance preservation: Style-9 ($n = 80$)					
Model	DINO ↓	Edge ↑	CLIP-T ↑	DINO ↓	Edge ↑	CLIP-T ↑	CSD ↑	ΔE_{00} ↓	LPIPS ↓
Vanilla (SD1.5)	0.7415	0.879225	**0.326803**	0.8310	0.884778	**0.333407**	0.477852	16.598175	0.542218
SPECTRA (SD1.5)	**0.1321**	**0.971661**	0.310350	**0.1271**	**0.972858**	0.298909	**0.853918**	**4.319736**	**0.148353**
Vanilla (SD2.1)	0.7345	0.886537	**0.333083**	0.8463	0.879751	**0.347094**	0.450109	16.896389	0.548978
SPECTRA (SD2.1)	**0.1483**	**0.970599**	0.310811	**0.1535**	**0.971377**	0.303636	**0.830143**	**4.306591**	**0.161643**
Vanilla (SD3)	0.4193	0.929038	**0.308297**	0.4594	0.928791	**0.325140**	0.673228	7.751022	0.366294
SPECTRA (SD3)	**0.1845**	**0.963370**	0.299517	**0.1646**	**0.969802**	0.296483	**0.873107**	**3.054036**	**0.159466**
Vanilla (FLUX.1-dev)	0.3263	0.935293	**0.296042**	0.3470	0.938004	**0.295440**	0.825216	6.758716	0.322700
SPECTRA (FLUX.1-dev)	**0.1870**	**0.965861**	0.293674	**0.1586**	**0.970501**	0.292088	**0.923750**	**3.243154**	**0.170879**

Table 4: NSST scale-depth ablation on PIE-Bench Style-9 ($n = 80$), using the final SD2.1 preservation setting. J is cumulative transform depth; direction adaptation remains enabled in every row.

J	Included directional scales	DINO ↓	Edge ↑	CLIP-T ↑	CSD ↑	ΔE_{00} ↓	LPIPS ↓
1	{Scale 1}	0.2029	0.967400	**0.310059**	0.783478	5.067538	0.196868
2	{Scale 1, Scale 2}	0.1802	0.970127	0.307306	0.799610	4.826044	0.180084
3	{Scale 1, Scale 2, Scale 3}	**0.1535**	**0.971377**	0.303636	**0.830143**	**4.306591**	**0.161643**

4.6 DOES DIRECTION-AWARE ALLOCATION HELP BEYOND SCALE ADAPTATION?

Cross-scale heterogeneity alone does not establish that directions within the same scale should be treated differently. We therefore isolate within-scale direction adaptation with Self Attn-Conv, AdaIN, and DCRCI disabled while SSMO and the LF scheduler remain fixed. Scale+Direction reduces DINO for SD1.5 ($0.6084 \rightarrow 0.5869$), SD2.1 ($0.6413 \rightarrow 0.6274$), SD3 ($0.3805 \rightarrow 0.3710$), and FLUX.1-dev ($0.3072 \rightarrow 0.3048$), while Edge, CSD, and CLIP-T are more mixed. Thus direction-aware allocation is a conditional refinement rather than a universal improvement knob. A stricter assignment control in Appendix H preserves the same non-uniform direction-weight multiset but cyclically misassigns it to different bands; the correct assignment gives the best DINO and Edge preservation on Color40. The complete cross-architecture metric table is reported in Appendix F.

4.7 MECHANISTIC ANALYSIS: WHEN DOES DIRECTION ADAPTATION HELP?

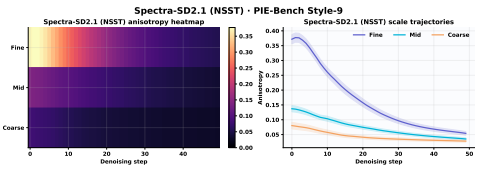
If directional adaptation responds to a meaningful source property rather than merely adding flexibility, its advantage should be larger when the source itself is more directionally anisotropic. Ablation averages do not establish this link, so we test three questions under the same matched SD2.1 NSST isolation setting. For scale s and timestep t, directional anisotropy is

$$\mathcal{A}_{s,t} = 1 - \frac{-\sum_{d=1}^{D_s} q_{s,d,t} \log q_{s,d,t}}{\log D_s} \in [0, 1], \tag{13}$$

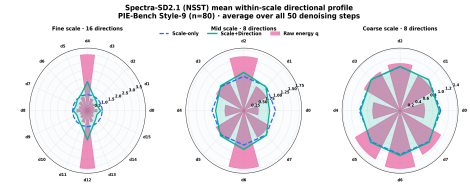
where zero denotes a uniform within-scale directional-energy distribution.

Figure 3 audits the mechanism. Panel (a) establishes a scale-dependent, time-varying anisotropy signal; panel (b) verifies that Scale+Direction, unlike the uniform Scale-only control, tracks the source directional profile. Panel (c) relates per-image mean anisotropy to an equal-weight composite of signed, standardized gains over DINO, Edge, CLIP-T, CSD, ΔE_{00}, and LPIPS. The positive Pearson/Spearman association indicates that more anisotropic sources tend to obtain larger paired directional benefit; we do not interpret it causally.

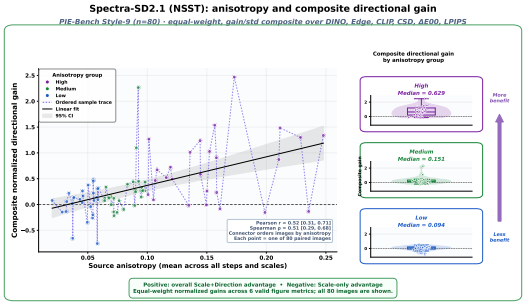
Limitations. Directional refinement is modest when source anisotropy is weak, and stronger preservation can reduce CLIP-T; SPECTRA therefore structures rather than removes the preservation–



(a) Scale- and time-dependent source anisotropy.



(b) Realized within-scale directional treatment.



(c) Source anisotropy versus paired Scale+Direction gain.

Figure 3: Directional-heterogeneity mechanism audit for SPECTRA (SD2.1, NSST) on Style-9. (a) Source anisotropy varies across scale and denoising time. (b) Scale+Direction responds with non-uniform within-scale allocation, unlike Scale-only. (c) Larger source anisotropy is associated with larger paired directional gain; the reported correlation is descriptive, not causal.

editability trade-off. The isolated DCRCI ablation likewise shows targeted gains in color/style fidelity rather than uniform improvement across all metrics, and appearance-preservation mode adds a VAE decode–reencode cycle. Finally, our experiments do not establish that the same policy is optimal when large non-rigid geometry changes require source structure itself to be discarded.

5 CONCLUSION

We studied source preservation in text-guided image editing as a structured allocation problem rather than a fixed frequency-selection problem. The evidence reveals a hierarchy: complementary information across scales provides the primary source-retention gain, while within-scale direction adaptation becomes useful when the source representation is sufficiently anisotropic. SPECTRA converts these source statistics into threshold-free scale–direction weights, couples them with target-semantic spatial evidence through SSMO, and returns the fused clean prediction to the backbone's native scheduler. The shuffled-direction control further shows that the structural benefit depends on assigning source-derived weights to the corresponding directional bands, while the isolated DCRCI study confirms its narrower role in shifting appearance-preservation outputs toward source color/style fidelity. Across diffusion and flow backbones, these results support treating source preservation as a dynamic Scale $\times$ Direction $\times$ Space $\times$ Time allocation problem rather than assigning fixed semantic roles to frequency bands.

REPRODUCIBILITY STATEMENT

The submission is accompanied by code and frozen experiment manifests. Section 4.1 states the main evaluation protocol and task-dependent metrics; Appendix B specifies benchmark subsets and prompt construction; Appendix D records backbone-specific generation settings; Appendix C gives metric definitions and preprocessing; and Appendix A provides the complete per-step pseudocode and scheduler conversions. Every paired comparison fixes source IDs, target prompts, seeds, checkpoints, schedules, and the common evaluator, and incomplete sample sets are rejected rather than silently averaged.

AI USE STATEMENT

Generative AI tools were used to assist with literature search and summarization, methodological and experimental framing, code debugging, organization and interpretation of already-computed results, figure/table presentation, and manuscript drafting and language editing. All experimental

outputs, numerical results, dataset manifests, and implementation artifacts reported in the paper were produced and checked by the authors. The authors manually reviewed the cited literature and verified the final claims, equations, references, tables, figures, and AI-assisted text, and take responsibility for the complete content of the submission.

REFERENCES

Black Forest Labs. FLUX. `https://github.com/black-forest-labs/flux`, 2024.

Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. LEDITS++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8861–8870, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Glenn R. Easley, Demetrio Labate, and Wang-Q Lim. Sparse directional image representations using the discrete shearlet transform. *Applied and Computational Harmonic Analysis*, 25(1):25–46, 2008. doi: 10.1016/j.acha.2007.09.003.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633, 2024.

Xiang Gao and Yunpeng Jia. FBSDiff++: Improved frequency band substitution of diffusion features for efficient and highly controllable text-driven image-to-image translation, 2026.

Xiang Gao and Jiaying Liu. FBSDiff: Plug-and-play frequency band substitution of diffusion features for highly controllable text-driven image translation. In *Proceedings of the ACM International Conference on Multimedia*, pp. 4101–4109, 2024. doi: 10.1145/3664647.3681380.

Xiang Gao, Zhengbo Xu, Junhan Zhao, and Jiaying Liu. Frequency-controlled diffusion model for versatile text-guided image-to-image translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1824–1832, 2024. doi: 10.1609/aaai.v38i3.27951.

Zheng Gao, Jifei Song, Zhensong Zhang, Jiankang Deng, and Ioannis Patras. Frequency-guided diffusion for training-free text-driven image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19195–19205, 2025.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510, 2017.

Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly DDPM noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12469–12478, 2024.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. PnP inversion: Boosting diffusion-based editing with 3 lines of code. In *International Conference on Learning Representations*, 2024.

Gwanhyeong Koo, Sunjae Yoon, Ji Woo Hong, and Chang D. Yoo. Flexiedit: Frequency-aware latent refinement for enhanced non-rigid editing. In *European Conference on Computer Vision*, pp. 363–379, 2024. doi: 10.1007/978-3-031-73036-8_21.

Kuan Heng Lin, Sicheng Mo, Ben Klingher, Fangzhou Mu, and Bolei Zhou. Ctrl-X: Controlling structure and appearance for text-to-image generation without guidance. In *Advances in Neural Information Processing Systems*, volume 37, pp. 128911–128939, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7465–7475, 2024.

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*, pp. 6038–6047, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

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*, pp. 10684–10695, 2022.

Gaurav Sharma, Wencheng Wu, and Edul N. Dalal. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application*, 30(1):21–30, 2005. doi: 10.1002/col.20070.

Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. In *European Conference on Computer Vision*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

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*, pp. 1921–1930, 2023.

Wei Wu, Qingnan Fan, Shuai Qin, Hong Gu, Ruoyu Zhao, and Antoni B. Chan. Freediff: Progressive frequency truncation for image editing with diffusion models. In *European Conference on Computer Vision*, pp. 194–209, 2024. doi: 10.1007/978-3-031-72652-1_12.

Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1395–1403, 2015.

Kaixiang Yang, Xin Li, Yuxi Li, Qiang Li, and Zhiwei Wang. FSI-Edit: Frequency and stochasticity injection for flexible diffusion-based image editing. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4434.

Kaixiang Yang, Boyang Shen, Xin Li, Yuchen Dai, Yuxuan Luo, Yueran Ma, Wei Fang, Qiang Li, and Zhiwei Wang. FIA-Edit: Frequency-interactive attention for efficient and high-fidelity inversion-free text-guided image editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11613–11621, 2026. doi: 10.1609/aaai.v40i14.38145.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Algorithm 1 One native SPECTRA sampling step at timestep t.

Require: source state z_t^{src}, edit state z_t^{edit}, target prompt p^{tgt}
predict $\hat{z}_{0,t}^{\mathrm{src}}$ under empty text and $\hat{z}_{0,t}^{\mathrm{edit}}$ under p^{tgt}
record target-semantic probability mass A_t from native cross/joint attention
$(L_t^{\mathrm{src}}, \{S_{s,d,t}^{\mathrm{src}}\}) \leftarrow \mathcal{T}(\hat{z}_{0,t}^{\mathrm{src}})$; $(L_t^{\mathrm{edit}}, \{S_{s,d,t}^{\mathrm{edit}}\}) \leftarrow \mathcal{T}(\hat{z}_{0,t}^{\mathrm{edit}})$
compute source scale energies, $w_{s,t}^{\mathrm{LF}}$, directional energies $E_{s,d,t}$, and $w_{s,d,t}$
$G_{s,d,t} \leftarrow \mathrm{SSMO}(A_t, w_{s,d,t})$ and fuse directional coefficients by Eq. equation 10
set $\widetilde{L}_t \leftarrow L_t^{\mathrm{edit}}$ and reconstruct $\widetilde{z}_{0,t} \leftarrow \mathcal{T}^{-1}(\widetilde{L}_t, \{\widetilde{S}_{s,d,t}\})$
if appearance preservation **then**
    apply AdaIN, then DCRCI, to obtain $\hat{z}_{0,t}^{\mathrm{merged}}$
else
    $\hat{z}_{0,t}^{\mathrm{merged}} \leftarrow \widetilde{z}_{0,t}$
end if
convert $\hat{z}_{0,t}^{\mathrm{merged}}$ to the native DDIM noise or FlowMatch velocity
execute exactly one unmodified native scheduler step to obtain z_{t-1}^{edit}

A COMPLETE SPECTRA PROCEDURE

A.1 PER-STEP PSEUDOCODE

Algorithm 1 makes the separation explicit: backbone-specific prediction, attention capture, and scheduler conversion stay native, whereas the transform controller, SSMO fusion, and optional appearance alignment act on clean latent predictions. No SPECTRA component changes the backbone weights.

A.2 DIFFUSION SOURCE TRAJECTORY AND DDIM CONVERSION

SD1.5 and SD2.1 encode the source image with the frozen VAE and construct a deterministic empty-text DDIM inversion trajectory using 999 inversion steps. The latent states corresponding to the 50 generation steps are cached. For epsilon prediction, the usual clean estimate is

$$\hat{z}_0 = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_t}{\sqrt{\bar{\alpha}_t}}. \tag{14}$$

After SPECTRA fusion, the epsilon implied by the desired clean prediction is recomputed as

$$\epsilon_t^{\mathrm{merged}} = \frac{z_t^{\mathrm{edit}} - \sqrt{\bar{\alpha}_t}\,\hat{z}_{0,t}^{\mathrm{merged}}}{\sqrt{1-\bar{\alpha}_t}}, \tag{15}$$

and passed to the unmodified scheduler. The reported DDIM configuration is deterministic ($\eta = 0$).

A.3 FLOW SOURCE STATE AND NATIVE CONVERSION

For SD3 and FLUX.1-dev, a single seed-fixed noise realization ϵ is shared by source-state reconstruction and target initialization. At a selected native FlowMatch step with scheduler coordinate σ_t, the source state is generated through the installed scheduler's native noising function; under the standard linear interpretation,

$$z_t^{\mathrm{src}} = (1-\sigma_t) z_0^{\mathrm{src}} + \sigma_t \epsilon. \tag{16}$$

A predicted velocity v_t corresponds to a clean prediction $\hat{z}_0 = z_t - \sigma_t v_t$. After fusion, the compatible velocity is

$$v_t^{\mathrm{merged}} = \frac{z_t^{\mathrm{edit}} - \hat{z}_{0,t}^{\mathrm{merged}}}{\sigma_t}, \tag{17}$$

and the installed FlowMatch Euler scheduler advances the sample. The strength parameter selects a suffix of the official native image-to-image schedule; SPECTRA uses the sigma aligned with that selected step rather than the first sigma of the untruncated schedule.

A.4 NATIVE SEMANTIC PROBABILITY MASS

For Stable Diffusion, let $P_t^{(\ell,h)}(p,k)$ be cross-attention probability from spatial position p to token k at layer ℓ and head h. With $\mathcal{T}_{\text{text}}$ the non-special target-token set,

$$A_t(p) = \frac{1}{|\mathcal{L}||\mathcal{H}|} \sum_{\ell\in\mathcal{L}} \sum_{h\in\mathcal{H}} \sum_{k\in\mathcal{T}_{\text{text}}} P_t^{(\ell,h)}(p,k). \tag{18}$$

Maps from different native resolutions are bilinearly aligned before layer aggregation. For SD3/FLUX.1-dev, the recorder uses image queries over the full native joint key sequence, sums attention probability assigned to target-text positions, then averages over heads and eligible blocks. The recorder is active only during the target-conditioned edit pass.

A.5 DCRCI DETAILS

Let x^{base} be the RGB image decoded from the AdaIN-aligned latent and let $(L^{\text{base}}, a^{\text{base}}, b^{\text{base}})$ be its D65 CIELAB representation. Let $(a^{\text{src}}, b^{\text{src}})$ be source chroma at the same image coordinates. SPECTRA forms a gamut-safe color target

$$x^{\text{color}} = \text{Lab}^{-1}_{\text{gamut}-\text{safe}}\left(L^{\text{base}}, a^{\text{src}}, b^{\text{src}}\right), \tag{19}$$

using 12 deterministic gamut-mapping iterations. The active frozen VAE deterministically encodes x^{base} and x^{color} to z^{base} and z^{color}, then injects only the cycle residual,

$$\hat{z}_{0,t}^{\text{merged}} = z_{0,t}^{\text{AdaIN}} + \lambda_C\left(z_{0,t}^{\text{color}} - z_{0,t}^{\text{base}}\right), \qquad \lambda_C = 1. \tag{20}$$

Thus DCRCI preserves the current lightness and geometry while correcting source-relative chroma drift.

B DATASETS AND BENCHMARK GOVERNANCE

Table 5: Frozen task-specific evaluation subsets. "Cases" denotes source–prompt pairs; ImageNet-R Official and RGB-fixed share the same 29 source images and differ only in target text.

Family	Subset	Task	Cases / sources	Prompt protocol
PIE-Bench	Color40	Appearance changing	40 / 40	Official `editing_prompt`
PIE-Bench	Style-9	Appearance preservation	80 / 80	Official `editing_prompt`
ImageNet-R-TI2I	Official	Appearance preservation	87 / 29	First three official target prompts
ImageNet-R-TI2I	RGB-fixed	Appearance changing	87 / 29	Same prompts with deterministic red/blue/green class-color insertion

PIE-Bench contains 700 annotated images over ten editing types (Ju et al., 2024). We use its official `6_change_attribute_color_40` and `9_change_style_80` subsets. Square-bracket annotation markers are removed without changing semantic words. The ImageNet-R-TI2I mapping distributed with PIE-Bench/PnP Inversion supplies five target prompts for each of 29 mapped ImageNet-R sources; we retain the first three. For RGB-fixed, if $p_{i,k}$ is the k-th retained official target prompt and c_i the class noun, we construct

$$\tilde{p}_{i,k} = \text{insert}(p_{i,k}, c_i, \gamma_k), \qquad \gamma_0 = \texttt{red},\ \gamma_1 = \texttt{blue},\ \gamma_2 = \texttt{green}. \tag{21}$$

The builder fails if the class noun is absent, and stores both the original and edited prompt. No image selection is based on model outputs. Every frozen manifest stores a relative source path, source identifier, target prompt, and verified image size; all source images used here are 512×512.

C EVALUATION METRICS

All source-output metrics use the same paired source x^{src}, generated image x^{gen}, and target prompt p^{tgt}. If a method generates at a different native resolution, x^{gen} is bicubically resized to the source resolution only for the temporary source-output metric pair; the saved generation is not overwritten.

DINO structure distance. We use a pinned DINO ViT-B/8 (Caron et al., 2021). From block 11 attention keys, including the CLS token, each token key is ℓ_2-normalized and used to form a cosine self-similarity matrix. DINO is the MSE between source and output self-similarity matrices. All paper tables multiply the raw MSE by 100.

Edge similarity. HED (Xie & Tu, 2015) produces edge maps for source and output. We embed the edge maps with the same CLIP image encoder (Radford et al., 2021) and report cosine similarity.

Target-text alignment. CLIP-T is the cosine similarity between the generated-image CLIP embedding and the embedding of p^{tgt}. The source prompt is never used for this metric.

Appearance-preservation metrics. CSD uses the official ViT-L style descriptor and reports source-output cosine similarity (Somepalli et al., 2024). ΔE_{00} is mean pixel-wise CIEDE2000 distance after a common sRGB-to-CIELab D65 conversion with unit parametric factors (Sharma et al., 2005). LPIPS uses the official AlexNet v0.1 model (Zhang et al., 2018). These three metrics are reported only when retaining source appearance is part of the task objective.

D Backbone and Task Configuration

Table 6: Backbone-specific generation settings used by SPECTRA. The SD3/FLUX.1-dev strength shown here is the operating point used in the main cross-backbone tables; the complete strength sweep is in Appendix E.

Backbone	Resolution	Native steps	Guidance	Strength	Source-state construction
SD1.5	512^2	50 DDIM	CFG 7.5	–	Empty-text DDIM inversion; 999 inversion steps; matched source states cached
SD2.1	512^2	50 DDIM	CFG 7.5	–	Empty-text DDIM inversion; 999 inversion steps; matched source states cached
SD3-Medium	1024^2	50 FlowMatch	CFG 7.0	0.4	Native img2img schedule, seed-fixed flow noise, scheduler-native source noising
FLUX.1-dev	1024^2	28 FlowMatch	guidance 3.5	0.4	Native img2img schedule, seed-fixed flow noise, scheduler-native source noising

For SD1.5/SD2.1, Self Attn-Conv denotes source self-attention Q/K injection over the first 80% of generation steps together with source convolutional feature injection over all steps. This U-Net-specific path is not emulated for SD3 or FLUX.1-dev. Across all backbones, the SPECTRA controller uses three-scale NSST, Scale+Direction adaptation, SSMO at every selected native step, LF scheduling on, HF fusion off, and empty-text source conditioning. AdaIN and DCRCI are off for Color40/RGB-fixed and on with unit strength for Style-9/ImageNet-R Official.

The directional-heterogeneity isolation studies intentionally differ from this final configuration: Self Attn-Conv, AdaIN, DCRCI, and HF fusion are off; SSMO and LF scheduling stay on. This distinction prevents appearance alignment or U-Net-specific injection from being attributed to the directional controller.

E SD3/FLUX.1-dev Strength Sensitivity

The native image-to-image strength changes the amount of the native schedule traversed and therefore moves the preservation–editability operating point. Tables 7 and 8 sweep only this parameter.

The monotonic preservation trend and the weaker/non-monotonic CLIP-T trend are why we treat 0.4 as a selected preservation-oriented operating point rather than a universally optimal strength.

F Full Cross-Architecture Directional Ablation

The main text reports the DINO trend and the mechanistic analysis. Table 9 gives the complete matched Scale-only/Scale+Direction comparison. Self Attn-Conv, AdaIN, and DCRCI are off;

Table 7: Strength sensitivity for appearance changing on PIE-Bench Color40 ($n = 40$). SSMO and LF scheduling are on; Self Attn-Conv, AdaIN, and DCRCI are off.

Model	Strength	DINO ↓	Edge ↑	CLIP-T ↑
SD3	0.4	**0.1845**	**0.963370**	0.299517
SD3	0.5	0.2606	0.953579	0.302705
SD3	0.6	0.3555	0.943482	**0.303805**
FLUX.1-dev	0.4	**0.1870**	**0.965861**	0.293674
FLUX.1-dev	0.5	0.2418	0.954120	0.292217
FLUX.1-dev	0.6	0.2858	0.944901	**0.295297**

Table 8: Strength sensitivity for appearance preservation on PIE-Bench Style-9 ($n = 80$). AdaIN and DCRCI are enabled; Self Attn-Conv is off.

Model	Strength	DINO ↓	Edge ↑	CLIP-T ↑	CSD ↑	ΔE_{00} ↓	LPIPS ↓
SD3	0.4	**0.1646**	**0.969802**	0.296483	**0.873107**	**3.054036**	**0.159466**
SD3	0.5	0.2233	0.958931	0.302507	0.830049	3.619108	0.203665
SD3	0.6	0.2961	0.949617	**0.309107**	0.782773	4.329398	0.249465
FLUX.1-dev	0.4	**0.1586**	**0.970501**	0.292088	**0.923750**	**3.243154**	**0.170879**
FLUX.1-dev	0.5	0.2181	0.959972	0.294987	0.893549	3.867091	0.222740
FLUX.1-dev	0.6	0.2610	0.953944	**0.295888**	0.877777	4.392754	0.258915

SSMO and the LF scheduler are on, so the two rows within each backbone differ only in the direction-aware odds modulation.

Table 9: Matched Scale-only versus Scale+Direction ablation on PIE-Bench Style-9 ($n = 80$) with NSST. SSMO and the LF scheduler are on; Self Attn-Conv, AdaIN, and DCRCI are off. Bold marks the better value within each backbone pair.

Backbone	Adaptation	DINO ↓	Edge ↑	CLIP-T ↑	CSD ↑	ΔE_{00} ↓
SD1.5	Scale-only	0.6084	0.901989	0.329210	0.591019	12.561252
	Scale+Direction	**0.5869**	**0.906085**	**0.329934**	**0.596218**	**12.355466**
SD2.1	Scale-only	0.6413	0.902645	**0.342469**	**0.558335**	12.360673
	Scale+Direction	**0.6274**	**0.904151**	0.341949	0.557640	**12.097185**
SD3	Scale-only	0.3805	**0.939675**	**0.317157**	0.719125	6.073878
	Scale+Direction	**0.3710**	0.939357	0.316005	**0.724979**	**5.968601**
FLUX.1-dev	Scale-only	0.3072	0.945525	**0.297217**	0.849005	5.889809
	Scale+Direction	**0.3048**	**0.945923**	0.297040	**0.849686**	**5.849706**

G TRANSFORM-GENERAL DIRECTIONAL ABLATIONS

The default NSST is useful because it exposes multiple native directions at each scale, but the controller is defined over generic directional subbands. We therefore repeat the Scale-only/Scale+Direction isolation with DTCWT, DWT, and NSCT. Each transform keeps its native analysis/synthesis and number of directions; the scale scheduler, source-only energy normalization, odds rule, SSMO gate, prompts, seeds, and evaluation protocol are fixed.

Table 10: Transform-general directional-heterogeneity ablation on PIE-Bench Style-9 ($n = 80$). Self Attn-Conv, AdaIN, and DCRCI are off; SSMO and LF scheduling are on. Bold marks the better value within each backbone–transform Scale-only/Scale+Direction pair.

Backbone	Transform	Adaptation	DINO ↓	Edge ↑	CLIP-T ↑	CSD ↑	ΔE_{00} ↓
SD1.5	DTCWT	Scale-only	0.630900	0.900893	**0.331566**	0.570083	13.786419
SD1.5	DTCWT	Scale+Direction	**0.619300**	**0.903570**	0.330081	**0.574700**	**13.649394**
SD1.5	DWT	Scale-only	0.619900	0.904396	**0.331286**	0.571097	13.831777
SD1.5	DWT	Scale+Direction	**0.613300**	**0.905741**	0.330079	**0.574846**	**13.727056**
SD1.5	NSCT	Scale-only	0.623900	**0.898967**	**0.330926**	0.583754	12.009912
SD1.5	NSCT	Scale+Direction	**0.623600**	0.898511	0.330653	**0.585887**	**12.007414**
SD2.1	DTCWT	Scale-only	0.667400	0.899972	0.343685	0.539182	13.564509
SD2.1	DTCWT	Scale+Direction	**0.653600**	**0.903597**	**0.344153**	**0.542858**	**13.387058**
SD2.1	DWT	Scale-only	0.651700	0.902969	**0.342219**	0.541375	13.661466
SD2.1	DWT	Scale+Direction	**0.649500**	**0.903659**	0.342041	**0.545559**	**13.518173**
SD2.1	NSCT	Scale-only	0.655900	**0.898018**	**0.343129**	0.555116	11.812372
SD2.1	NSCT	Scale+Direction	**0.655400**	0.897570	0.343015	**0.555544**	**11.810691**
SD3	DTCWT	Scale-only	0.381400	**0.939024**	**0.316698**	0.721651	6.133696
SD3	DTCWT	Scale+Direction	**0.375000**	0.938867	0.316216	**0.724296**	**6.063314**
SD3	DWT	Scale-only	0.381700	**0.940139**	**0.316670**	0.723406	6.245104
SD3	DWT	Scale+Direction	**0.377000**	0.939685	0.316208	**0.725869**	**6.200093**
SD3	NSCT	Scale-only	0.387900	0.938528	**0.316396**	0.720356	5.918791
SD3	NSCT	Scale+Direction	**0.387400**	**0.938987**	0.316342	**0.720784**	**5.911797**
FLUX.1-dev	DTCWT	Scale-only	0.307800	**0.945696**	**0.296936**	0.848719	5.904492
FLUX.1-dev	DTCWT	Scale+Direction	**0.305500**	0.945530	0.296872	**0.849322**	**5.865041**
FLUX.1-dev	DWT	Scale-only	0.306200	**0.945432**	0.296956	0.849560	5.900141
FLUX.1-dev	DWT	Scale+Direction	**0.304100**	0.945084	**0.297109**	**0.850499**	**5.886910**
FLUX.1-dev	NSCT	Scale-only	0.310800	0.944567	0.297045	0.847266	5.921114
FLUX.1-dev	NSCT	Scale+Direction	**0.310400**	**0.944809**	**0.297069**	**0.847692**	**5.917606**

For SD2.1, Tables 11 and 12 additionally include NSST and LPIPS. The transform study does not establish that every transform benefits on every metric. Its narrower conclusion is that the source-only direction rule can be instantiated on different native directional decompositions without transform-specific tuning.

Table 11: SD2.1 transform ablation for appearance preservation on Style-9 ($n = 80$), all-off isolation setting.

Transform	Adaptation	DINO ↓	Edge ↑	CLIP-T ↑	CSD ↑	ΔE_{00} ↓	LPIPS ↓
NSST	Scale-only	0.6413	0.902645	**0.342469**	**0.558335**	12.360673	0.434451
NSST	Scale+Direction	**0.6274**	**0.904151**	0.341949	0.557640	**12.097185**	**0.425276**
DTCWT	Scale-only	0.6674	0.899972	0.343685	0.539182	13.564509	0.454775
DTCWT	Scale+Direction	**0.6536**	**0.903597**	**0.344153**	**0.542858**	**13.387058**	**0.449578**
DWT	Scale-only	0.6517	0.902969	**0.342219**	0.541375	13.661466	0.450149
DWT	Scale+Direction	**0.6495**	**0.903659**	0.342041	**0.545559**	**13.518173**	**0.447336**
NSCT	Scale-only	0.6559	**0.898018**	**0.343129**	0.555116	11.812372	0.440765
NSCT	Scale+Direction	**0.6554**	0.897570	0.343015	**0.555544**	**11.810691**	**0.440481**

Table 12: SD2.1 transform ablation for appearance changing on Color40 ($n = 40$), all-off isolation setting.

Transform	Adaptation	DINO ↓	Edge ↑	CLIP-T ↑
NSST	Scale-only	0.5905	0.901333	**0.329739**
NSST	Scale+Direction	**0.5810**	**0.903617**	0.329178
DTCWT	Scale-only	0.6038	0.897254	**0.332482**
DTCWT	Scale+Direction	**0.6021**	**0.901915**	0.332171
DWT	Scale-only	0.6001	0.897642	0.332141
DWT	Scale+Direction	**0.5977**	**0.900713**	**0.332188**
NSCT	Scale-only	0.5971	0.897179	0.327959
NSCT	Scale+Direction	**0.5947**	**0.898957**	**0.328705**

H TARGETED ASSIGNMENT AND APPEARANCE-ALIGNMENT CONTROLS

The two controls in this section isolate specific mechanisms that are not answered by the broader architecture and transform sweeps. The first asks whether the directional controller benefits from assigning source-derived weights to the *correct* bands, rather than from non-uniform weighting alone. The second isolates the task-specific contribution of DCRCI after AdaIN in the final appearance-preservation configuration.

Correct directional assignment. Table 13 compares Scale-only, the proposed Scale+Direction assignment, and a Scale+Shuffled-Direction control on PIE-Bench Color40. All three use the same final SPECTRA-SD2.1 appearance-changing configuration: three-scale NSST ($J = 3$), SSMO, LF scheduling, and Self Attn-Conv are enabled, while AdaIN and DCRCI are disabled. Scale+Shuffled-Direction applies a fixed non-identity cyclic permutation to the direction-specific fusion weights within each scale, preserving the weight multiset while breaking its correspondence to the source directional bands. Correct assignment gives the best DINO and Edge preservation, whereas the shuffled control yields slightly higher CLIP-T. The result supports the narrower mechanistic conclusion that the structural benefit depends on matching source-derived weights to their corresponding directions, rather than on introducing non-uniform direction weights alone.

Table 13: Correct directional assignment on PIE-Bench Color40 ($n = 40$). Scale+Shuffled-Direction preserves the direction-weight multiset within each scale but cyclically reassigns the weights to different bands. All other settings are identical. Bold denotes the best value per metric.

Adaptation	DINO ↓	Edge ↑	CLIP-T ↑
Scale-only	0.1485	0.970097	0.310990
Scale+Direction (correct)	**0.1483**	**0.970599**	0.310811
Scale+Shuffled-Direction	0.1496	0.969944	**0.311387**

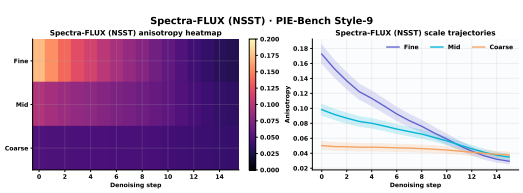
Effect of DCRCI after AdaIN. Table 14 isolates DCRCI on PIE-Bench Style-9. Both variants use the final SPECTRA-SD2.1 preservation configuration with three-scale NSST ($J = 3$), Scale+Direction adaptation, SSMO, LF scheduling, Self Attn-Conv, and AdaIN; source–target pairs, cached DDIM trajectories, and seed are identical. The only difference is whether DCRCI is applied after AdaIN. Adding DCRCI improves CSD, reduces ΔE_{00} from 5.016290 to 4.306591, and slightly improves Edge similarity, consistent with its intended role as an internal source-relative chroma-correction stage. AdaIN alone retains better DINO, CLIP-T, and LPIPS, showing that DCRCI shifts the operating point toward source color/style fidelity rather than uniformly improving every objective.

Table 14: Appearance-alignment ablation on PIE-Bench Style-9 ($n = 80$). Both variants use identical SPECTRA-SD2.1 settings and differ only in whether DCRCI is applied after AdaIN. Bold denotes the better value per metric.

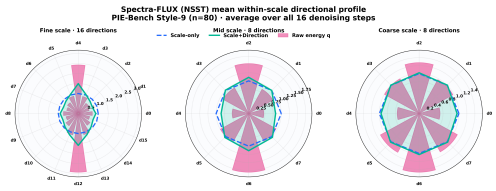
Appearance alignment	DINO ↓	Edge ↑	CLIP-T ↑	CSD ↑	ΔE_{00} ↓	LPIPS ↓
AdaIN only	**0.1509**	0.970445	**0.313372**	0.813432	5.016290	**0.140420**
AdaIN + DCRCI	0.1535	**0.971377**	0.303636	**0.830143**	**4.306591**	0.161643

I ADDITIONAL DIRECTIONAL-HETEROGENEITY EVIDENCE

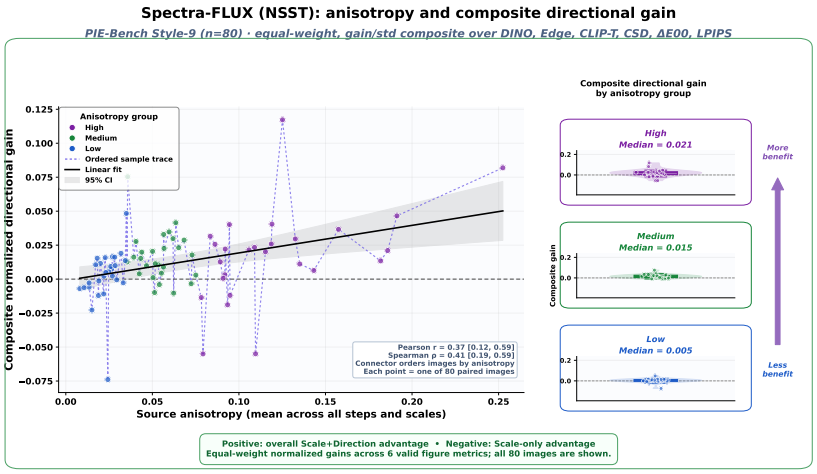
The main paper reports the SD2.1 mechanistic sequence. We repeat the same measurements for FLUX.1-dev and keep the two backbones separate. Figure 4 mirrors the three main-paper diagnostics: the source signal changes across scale and time, the controller responds to the native directional-energy profile, and higher anisotropy is associated with larger paired directional gain. Figure 5 complements average metrics with case-wise win rates and uncertainty.



(a) Anisotropy over the native FlowMatch trajectory.

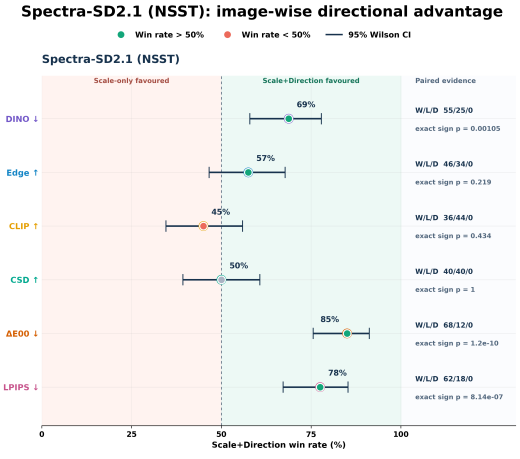


(b) Mean within-scale directional treatment.

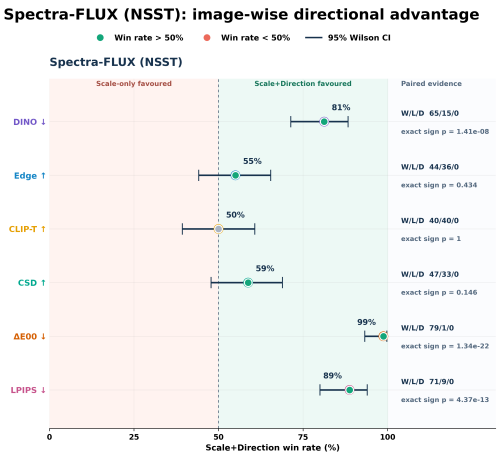


(c) Source anisotropy versus paired directional gain.

Figure 4: Mechanistic analysis for SPECTRA (FLUX.1-dev, NSST) on Style-9. (a) Source anisotropy remains scale-dependent and time-varying under the selected native FlowMatch trajectory. (b) Scale+Direction produces non-uniform source-dependent allocation while Scale-only is direction-invariant. (c) The anisotropy–gain association uses the same signed, standardized composite as Fig. 3(c); it is descriptive rather than causal.



(a) SD2.1.



(b) FLUX.1-dev.

Figure 5: Image-wise directional advantage over Scale-only on Style-9. Each point is the Scale+Direction win fraction for one metric; horizontal bars are 95% Wilson intervals, and exact two-sided sign tests compare wins with losses while excluding draws. The case-wise view shows that directional refinement is strong for several preservation metrics but mixed or neutral for others, matching the qualified claim in the main text.

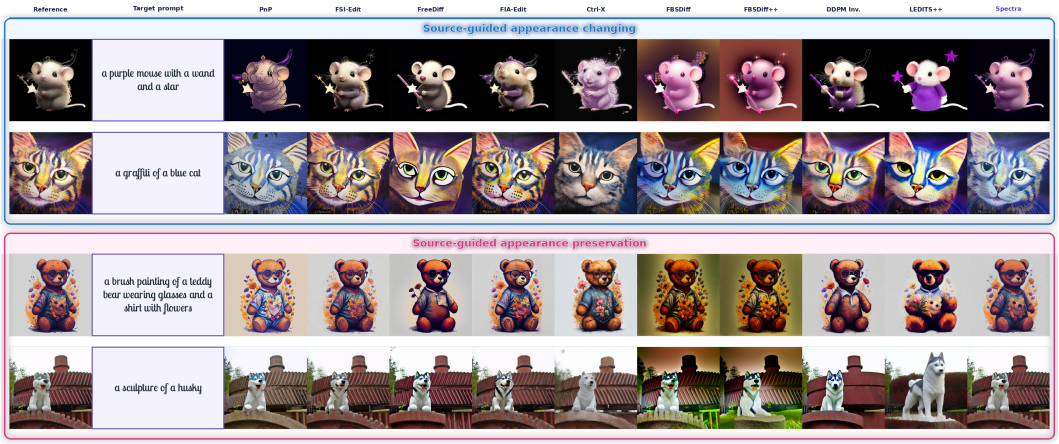


Figure 6: Additional baseline comparisons for appearance-changing and appearance-preserving edits. These examples complement Fig. 2 with different sources and target prompts.

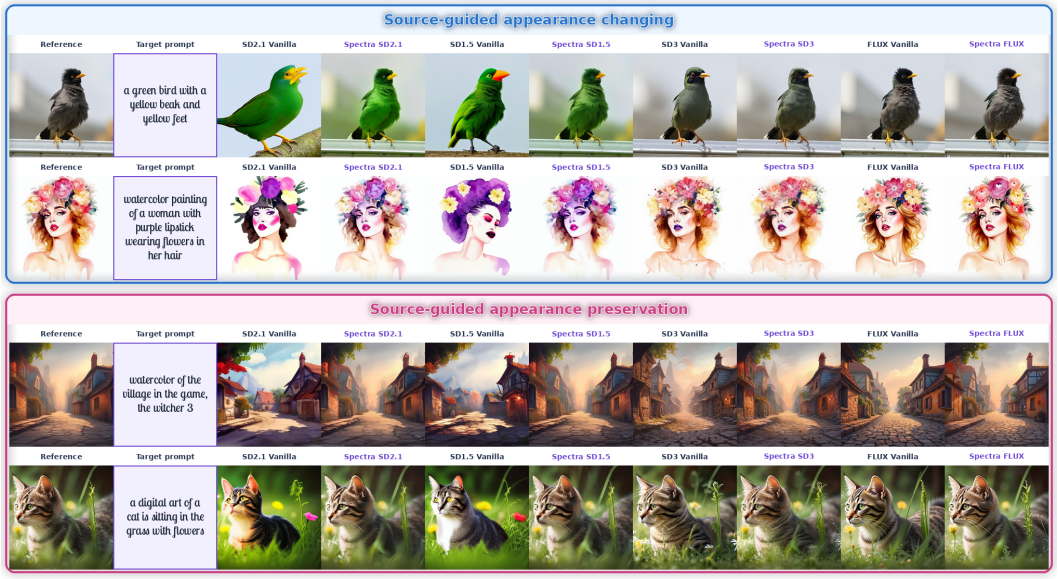


Figure 7: Additional Vanilla-versus-SPECTRA comparisons across diffusion and flow backbones. The figure is intended as a cross-backbone qualitative check rather than a controlled component ablation because backbone-native source construction and the selected operating points differ.

J ADDITIONAL QUALITATIVE RESULTS

K IMPLEMENTATION AND REPRODUCIBILITY DETAILS

All experiments were run on a Linux server with four NVIDIA GeForce RTX 3090 GPUs, each with 24 GB VRAM, using NVIDIA driver 580.173.02. A generation or evaluation worker uses one GPU. Multi-GPU execution only shards independent images or experiment conditions; there is no cross-device activation exchange or gradient aggregation. Within every reported paired comparison, the source image, target prompt, seed, backbone checkpoint, inference schedule, and resolution are fixed. Generation and evaluation artifacts are stored per sample with explicit identifiers, and an experiment is considered complete only when every expected sample is present. Missing samples are not silently skipped.

The common evaluator pins the DINO checkpoint and repository state, the official CSD ViT-L checkpoint, CLIP/HED preprocessing, CIEDE2000 validation data, and official LPIPS-AlexNet v0.1. Each run logs transform metadata, scale and directional energies, attention statistics, SSMO gate statistics, reconstruction diagnostics, model revision, scheduler settings, and appearance-alignment diagnostics. These records are intended to make the scale/direction mechanism auditable independently of the rendered image.