ENDPOINT-CALIBRATED PULLBACK INTERPOLATION IN LATENT DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion image interpolation can lose texture and edge definition between otherwise detailed endpoints. Geometric and density-based routing criteria guide intermediate latents, but do not directly target such losses in the rendered images. We propose Endpoint-Calibrated Pullback Interpolation (ECPI), a plug-in correction for existing diffusion interpolation paths. ECPI derives position-dependent detail references from the frequency-band statistics of reconstructed endpoints. Responses below these references define the spectral deficits to correct. We differentiate the rendered responses through denoising and decoding to obtain updates to the noisy intermediate latents, allowing the generative process to realize the correction. On MorphBench and our 100-pair Animals-and-Humans evaluation set (AH100), ECPI improves no-reference image quality and reduces FID against endpoint reference sets across SLERP, Tangent, and GeoDiff routes. These gains are achieved while constraining each correction to have a non-negative first-order change in conditional log density at the correction state. This shows that detail recovery is compatible with locally probability-supported correction directions.

1 INTRODUCTION

Image interpolation aims to produce a coherent visual transition between two inputs while maintaining the quality of the intermediate images. Diffusion methods invert the inputs to noisy latent states at a common noise level, connect those states, and render points along the path by denoising and decoding (Song et al., 2021a; Zhang et al., 2024). Existing approaches guide this construction through different latent-space criteria: SLERP interpolates on a sphere, Tangent penalizes directions with large changes in the model's score (Saito & Matsubara, 2026), and GeoDiff favors paths through higher-density regions (Yu et al., 2025).

These criteria regulate the latent path at the interpolation noise level, but do not directly constrain what visual structure survives the subsequent denoising and decoding process. This distinction matters because diffusion generation is frequency-dependent: coarse, low-frequency structure generally emerges earlier in the reverse process, whereas finer, higher-frequency detail is recovered later (Choi et al., 2022; Rissanen et al., 2023; Falck et al., 2025). In practice, the intermediate images produced by these routes can lose detail relative to otherwise detailed endpoints. This motivates correcting the intermediate latent states according to what survives through the remaining denoising and decoding process.

But what signal should guide such a correction? Image morphing has no ground-truth intermediate image, so a pixelwise reconstruction objective is not available. The endpoints nevertheless provide information about structure at different spatial scales. Frequency-band energy summarizes this information without requiring pixelwise correspondence: lower frequencies primarily capture coarse structure, whereas higher frequencies are more sensitive to fine-scale variations such as edges and textures. We therefore interpolate the reconstructed endpoints' log-band energies to obtain a position-dependent reference. A shortfall relative to this reference defines a *spectral deficit*, indicating which frequency bands fall below the reference and by how much.

We implement this idea as *Endpoint-Calibrated Pullback Interpolation* (ECPI), a plug-in correction for existing diffusion interpolation paths. Differentiating the generated image's log-band energies through the remaining denoising and decoding process pulls the spectral objective back to the intermediate latent state, yielding directions that locally increase the requested responses. ECPI

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.

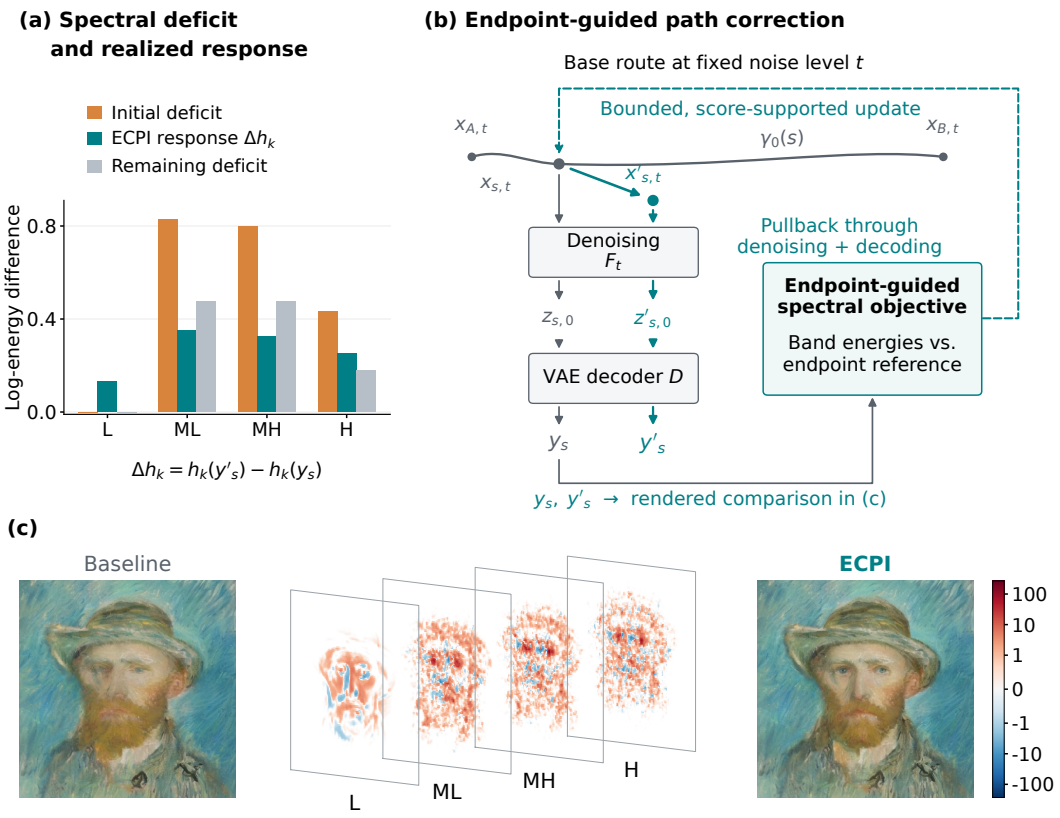

Figure 1: **From spectral requests to rendered changes.** Van Gogh self-portrait interpolation with SLERP at $s = 9/16$. (a) Initial deficits, ECPI responses and remaining deficits. Only the mid-low (ML), mid-high (MH) and high (H) bands have positive deficits in this example. The low (L) band receives no direct correction request but also changes through the joint latent update. (b) Spectral feedback corrects a noisy state on the base route; baseline (gray) and ECPI (teal) share the denoising and decoding map. (c) Baseline and ECPI flank the spatial band-energy changes; red and blue indicate increases and decreases. All bands share the same color scale. Visualization details appear in Appendix C.1.

combines these directions according to the spectral deficits and applies a bounded update to the intermediate noisy state. Figure 1 illustrates the process; details follow below.

Prior work links image detail to sample density and uses density control to adjust it during generation (Karczewski et al., 2025). This raises a question for our correction: does recovering the requested spectral content require moving toward lower model density? We address this by constraining the latent update to have a non-negative first-order change in conditional log density. The endpoint spectra specify the correction request, while the score constrains the direction used to realize it.

Our contributions are:

- We identify a recurring endpoint-relative spectral deficit across three interpolation routes and two datasets, and show through controlled reference interventions that spectral requests shape rendered responses.
- We introduce ECPI, a plug-in method that translates endpoint-derived spectral requests into bounded latent corrections through the denoising-and-decoding map, improving independent image-quality and distributional measures across the three routes.
- We offer a complementary perspective on detail recovery in diffusion interpolation: endpoint-defined spectral correction remains effective even when latent updates are con-

strained not to decrease conditional log density to first order. Matched-displacement experiments show that this constraint retains almost all of the raw task recovery.

2 Related Work

Diffusion interpolation and latent geometry. Training-free diffusion interpolation has been explored through interleaved interpolation and denoising (Wang & Golland, 2023), latent-noise correction (Zheng et al., 2024), and attention-based interpolation (He et al., 2024; Cao et al., 2025). Geometric approaches instead define paths between inverted endpoint latents: SLERP uses spherical interpolation, Tangent (Saito & Matsubara, 2026) derives a metric from the score-function Jacobian, and GeoDiff (Yu et al., 2025) uses a probability-density-induced metric. ECPI complements these routing criteria by measuring endpoint-relative spectral deficits in rendered images and pulling their correction back through the remaining denoising and decoding process.

Differentiable downstream guidance. Universal Guidance (Bansal et al., 2024) evaluates differentiable guidance objectives on predicted clean images during sampling, whereas DOODL (Wallace et al., 2023) differentiates final-image objectives through an invertible diffusion trajectory to optimize its initial latents. ECPI builds on downstream differentiation to formulate bounded corrections of existing interpolation paths. Reconstructed endpoint spectra provide position-dependent references; comparison with the current renderings identifies one-sided spectral deficits, which are addressed through gradients of the full denoising-and-decoding continuation.

Detail control during generation. Karczewski et al. (2025) relate sample likelihood to image detail and introduce Density Guidance, which controls sampling through a prescribed density-evolution policy. Humayun et al. (2025) characterize local generator geometry through local scaling and train a corresponding reward model to guide generation. These studies motivate considering rendered detail alongside model geometry. ECPI addresses a complementary interpolation task: reconstructed endpoints define position-dependent spectral requests for correcting an existing route.

Finite-frame output. IMPUS (Yang et al., 2024) and DiffMorpher (Zhang et al., 2024) observe that uniform latent spacing yields perceptually non-uniform sequences and resample the schedule using rendered LPIPS (Zhang et al., 2018); GeoDiff adopts DiffMorpher's resampling strategy. We apply the same perceptual-resampling protocol to each base route and its ECPI correction.

3 Method

ECPI corrects an existing interpolation path at a fixed diffusion noise level. The base method supplies intermediate noisy states; ECPI evaluates the images obtained through their remaining denoising and decoding. Endpoint-relative spectral deficits specify the correction requests. We pull rendered-response gradients back to the same noisy states and apply bounded updates under local probability support (Figure 1).

3.1 Endpoint-calibrated rendered objective

Let endpoint images y_A, y_B have conditions c_A, c_B and inverted states $x_{A,t}, x_{B,t}$ at noise level t. A base route $\gamma_0 : [0,1] \to \mathcal{X}_t$ connects these states, with $c(s) = (1-s)c_A + sc_B$. Here, s indexes interpolation positions, while t is fixed across the route when applying the correction. Each noisy state is rendered as

$$x_{s,t} = \gamma_0(s) \xrightarrow{F_t(\cdot, c(s))} z_{s,0} \xrightarrow{D} y_s, \tag{1}$$

where F_t is deterministic reverse continuation from t to the clean decoder latent $z_{s,0}$, and D is the VAE decoder. We denote the complete map by $\Psi_t(x, c) = D(F_t(x, c))$. The reconstructed endpoints $\widetilde{y}_A = \Psi_t(x_{A,t}, c_A)$ and $\widetilde{y}_B = \Psi_t(x_{B,t}, c_B)$ calibrate the reference within the same inversion–generation system as the interior frames.

Let $\mathbf{h}(y) \in \mathbb{R}^4$ measure log energy in four fixed radial frequency bands (Appendix A). The endpoints define

$$\mathbf{h}^\star(s) = (1-s)\,\mathbf{h}(\widetilde{y}_A) + s\,\mathbf{h}(\widetilde{y}_B). \tag{2}$$

At sampled path positions $s_i = i/N$, write $x_i = x_{s_i,t}$ and $c_i = c(s_i)$. The signed residual in band k is

$$e_{ik} = h_k^\star(s_i) - h_k(\Psi_t(x_i, c_i)). \tag{3}$$

The endpoint-derived reference is a calibration signal rather than a ground-truth target. We therefore use it as a spectral floor, requesting additional band energy only where the rendered response falls short (Appendix C.2).

Figure 1(a,c) shows localized ML–H energy increases around the eyes and nose, accompanying clearer facial structure. Low-band energy also increases overall despite receiving no direct correction request.

3.2 Task pullback with probability support

We differentiate the rendered band response backward through the chain in Eq. (1). At each interior sample point, holding t and c_i fixed, the gradient of band k with respect to the noisy state is

$$g_{ik} = \nabla_{x_i} h_k(\Psi_t(x_i, c_i)) = J_{F_t}^\top J_D^\top \nabla_y h_k. \tag{4}$$

We retain route-normal motion through an orthogonal projector Q_i that removes radial and local path-tangent components. This directs the update toward route deformation at fixed latent norm. We combine the normalized projected gradients into a task direction u_i, weighting each active band by its deficit capped at its first-order reachable change within trust radius ρ. Appendix A.2 gives the projection and weight definitions.

We seek detail recovery subject to non-negative first-order change in conditional log density. With projected score $v_i = Q_i \nabla_x \log p_t(x_i \mid c_i)$, we retain the closest admissible task direction:

$$\widetilde{d}_i = \operatorname*{arg\,min}_{d:\, \langle v_i, d\rangle \geq 0} \frac{1}{2}\|d - u_i\|_2^2, \qquad d_i = \frac{\widetilde{d}_i}{\|\widetilde{d}_i\|_2}. \tag{5}$$

This half-space projection retains the feasible component of the task direction; its closed form is given in Appendix A.2. Zero directions receive no update. Section 4.4 measures the resulting recovery and finite-step probability change.

Applying the correction. With directions d_i fixed, we select bounded magnitudes $q_i \in [0, 1]$ from one linearization of the band responses, with small neighboring-difference penalties to regularize the correction field. Each interior state moves along a spherical arc of length ρq_i in direction d_i, preserving its latent norm; both endpoints remain fixed. Adjacent corrected states are connected by SLERP. We render the corrected route through the same Ψ_t, using the same rendered-LPIPS 17-frame output protocol as the baseline. Appendix A.3 gives the magnitude objective, solver settings and explicit update.

4 Experiments

The experiments follow the order of the argument. **(Q1)** Do interpolation paths exhibit the spectral deficit our motivation predicts, and does the endpoint-defined request control the rendered response? **(Q2)** Does ECPI improve interpolation quality across routes and datasets, and can output- or decoder-side alternatives reproduce its effect? **(Q3)** Is task-directed detail recovery compatible with local probability support, and what does the constraint cost?

4.1 Setup

Data, model and routes. We evaluate 90 MorphBench pairs (Zhang et al., 2024): 24 Animation and 66 Metamorphosis (Anim./Meta.), and our fixed Animals-and-Humans set (AH100): 50 AFHQ animal pairs (Choi et al., 2020) and 50 CelebA-HQ face pairs (Karras et al., 2018). Using Stable Diffusion v2.1-base (Rombach et al., 2022) and deterministic DDIM inversion (Song et al., 2021a) to $t = 581$, we correct SLERP, GeoDiff (Yu et al., 2025) and Tangent (Saito & Matsubara, 2026) routes. Baseline and ECPI share inversion, conditioning, seeds and official 17-frame perceptual resampling. Spectral deficits are measured at uniform correction knots; output metrics use the resampled frames. Implementation and evaluation details appear in Appendices A and B.

Table 1: Endpoint-relative spectral deficits on uncorrected paths (90 MorphBench and 100 AH100 pairs per route). Entries average positive four-band log-energy residuals within each pair and then across pairs. Central positions use $i = 6, \ldots, 10$; near-boundary positions use $i \in \{1, 2, 3, 13, 14, 15\}$, with $s = i/16$ and endpoints excluded. The last column reports paired central-minus-near-boundary differences and 95% bootstrap intervals.

Dataset	Route	Central	Near-boundary	Difference [95% CI]
MorphBench	SLERP	0.590	0.167	+0.423 [0.354, 0.504]
	Tangent	0.673	0.229	+0.443 [0.371, 0.523]
	GeoDiff	0.330	0.101	+0.229 [0.181, 0.286]
AH100	SLERP	0.284	0.109	+0.176 [0.146, 0.207]
	Tangent	0.332	0.143	+0.188 [0.154, 0.225]
	GeoDiff	0.137	0.070	+0.067 [0.047, 0.088]

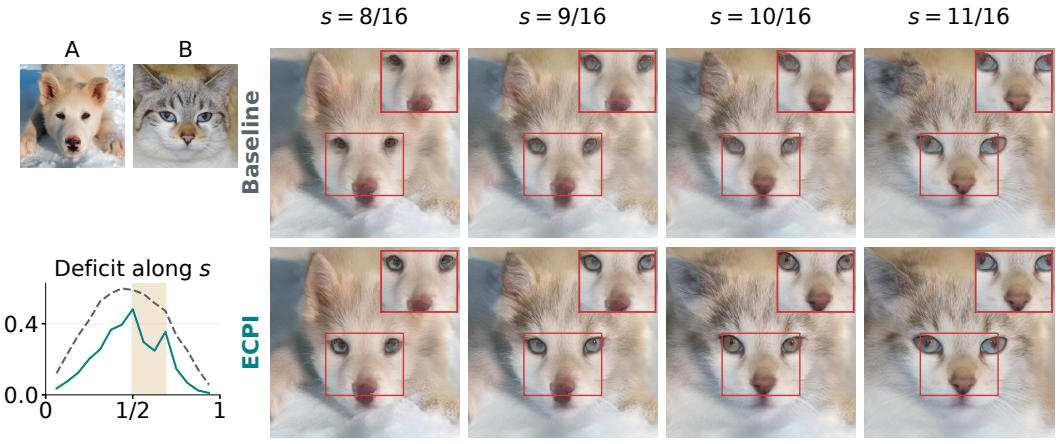

Figure 2: **Spectral deficits and their correction across consecutive intermediate frames.** Baseline and ECPI at $s = 8/16, \ldots, 11/16$ on an animal photograph pair (AH100, Tangent). Small images show reconstructed endpoints; red boxes mark matched regions shown as crops in the insets. Curves show mean positive four-band deficits (gray: baseline; teal: ECPI), with the displayed interval shaded. The complete sequence and a second pair appear in Appendix F.1.

Metrics. The remaining spectral deficit is the mean positive four-band residual against Eq. (2); it measures how much of the request remains unmet. Quality is assessed independently of this objective using TOPIQ (Chen et al., 2024), MUSIQ (Ke et al., 2021) and MANIQA (Yang et al., 2022), with aggregation defined in Appendix B.1. FID (Heusel et al., 2017) measures distributional agreement between the sets of interior images and original endpoints from the same cohort. For transitions, PPL sums adjacent-frame LPIPS-Alex distances, PDV is their standard deviation, and ISTD is the standard deviation of adjacent-frame pixel-RMS distances. Probability support is evaluated at fixed time and conditioning by score line integration (Appendix D).

Statistics. Frame-level quantities are aggregated within pairs before computing pair-bootstrap percentile 95% intervals. FID is a set-level point estimate. Matched-displacement controls use a common movement budget.

4.2 Spectral deficits and endpoint-conditioned response (Q1)

Interior renderings exhibit spectral deficits. We test whether uncorrected interpolation paths exhibit endpoint-relative spectral deficits. Across both datasets and all three routes, central interior positions have larger mean deficits than near-boundary interior positions, with positive paired differences in all six groups (Table 1). Figure 2 illustrates the accompanying attenuation of texture and edge definition in a selected sequence. Full position profiles appear in Appendix C.1.

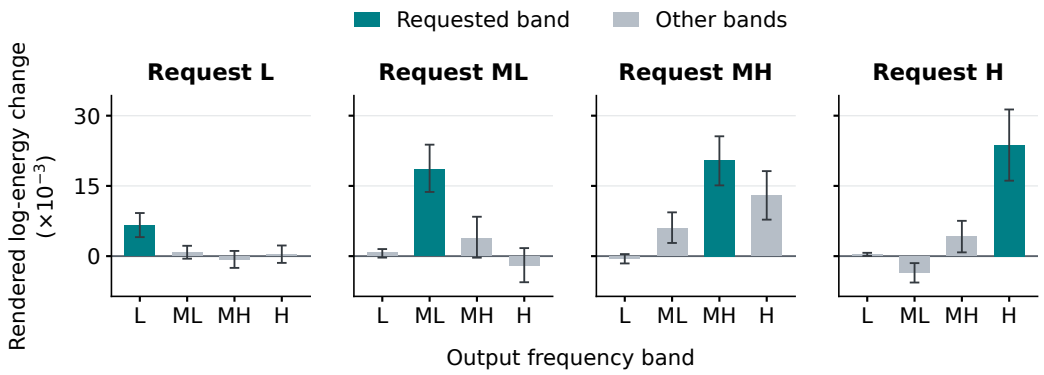

Figure 3: **Strengthening a request preferentially increases the requested band.** At the midpoint of 12 Tangent pairs, we increase one band's reference from 0.05 to 0.15 above its baseline log-energy. Each panel shows the resulting mean changes in all four output bands; teal marks the requested band. Error bars are pointwise 95% pair-bootstrap intervals. Activation tests and complete response matrices appear in Appendix C.1.

Table 2: Image interpolation results under the same official 17-frame output protocol on MorphBench Animation (Anim., 24 pairs), Metamorphosis (Meta., 66 pairs), and Animals-and-Humans (AH100, 100 pairs). All entries are final absolute scores. Rows prefixed by "+" apply a plug-in correction to the preceding base route. Best results within each base-route block are **bold**, with ties at displayed precision sharing the same rank. FID uses interior frames and cohort-matched endpoint references; all three IQA scores use Eq. (12).

	FID↓			TOPIQ↑			MUSIQ↑			MANIQA↑		
Method	Anim.	Meta.	AH100	Anim.	Meta.	AH100	Anim.	Meta.	AH100	Anim.	Meta.	AH100
GeoDiff	38.24	58.13	42.39	0.5246	0.5779	0.6055	66.51	66.10	70.35	0.4009	0.4463	0.5134
+ ECPI	**37.74**	**57.58**	**41.67**	**0.5332**	**0.5954**	**0.6160**	**67.27**	**67.69**	**71.00**	**0.4081**	**0.4611**	**0.5192**
Tangent	41.33	65.13	43.73	0.4946	0.5189	0.5502	62.56	58.05	63.34	0.3666	0.3871	0.4546
+ ECPI	**40.04**	**63.43**	**42.52**	**0.5078**	**0.5501**	**0.5661**	**63.87**	**61.42**	**65.14**	**0.3786**	**0.4144**	**0.4674**
SLERP	41.96	62.16	41.53	0.4964	0.5449	0.5678	63.09	60.52	65.97	0.3726	0.4085	0.4768
+ ECPI	**40.67**	**61.58**	**40.43**	**0.5118**	**0.5706**	**0.5815**	**64.37**	**63.25**	**67.36**	**0.3846**	**0.4316**	**0.4867**

The endpoint request controls the rendered response. We test request dependence by intervening on the reference while holding the underlying state and generation process fixed. On 12 frozen Tangent pairs at $s = 1/2$, we alter a single band's reference value while keeping the latent state, conditioning, endpoint images, band gradients and trust budgets unchanged, then recompute the correction and render with the same continuation and decoder (Appendix C.1). Strengthening an active request increases the rendered energy of the targeted band in every pair and for all four bands, with the targeted band showing the largest mean response (Figure 3). Other bands also change: the request selectively influences the rendered response without isolating one frequency band from the others. Appendix Figure 13 shows how the joint ECPI correction is distributed across frequency bands in the animal example.

At scale, 21,164 of 21,233 initially deficient state–band observations (99.7%) show positive rendered responses. As expected from the one-sided rule, all 772 states with no active request receive zero latent displacement.

4.3 Interpolation quality and trade-offs (Q2)

Rendered quality and transfer. We first check whether ECPI realizes the requested correction. On MorphBench, mean spectral deficits decrease in every band and route, although above-reference

Table 3: Generative and RGB-space correction on the same 50 evaluation pairs per route. Fixed-filter parameters and the adaptive search rule are frozen after selection on a separate development set. Entries use final outputs; FID is a set-level estimate. Best and second-best values within each route are bold and underlined.

Route	Method	TOPIQ ↑	MUSIQ ↑	MANIQA ↑	FID ↓
GeoDiff	Baseline	0.5782	67.70	0.4443	56.67
	ECPI	<u>0.6024</u>	<u>70.22</u>	<u>0.4763</u>	**55.22**
	Fixed RGB	**0.6531**	**72.66**	**0.5277**	58.24
	Endpoint-adaptive RGB	0.5911	68.75	0.4546	<u>56.09</u>
Tangent	Baseline	0.5339	62.64	0.4042	64.21
	ECPI	<u>0.5640</u>	<u>66.43</u>	<u>0.4333</u>	**61.93**
	Fixed RGB	**0.6151**	**68.37**	**0.4843**	67.57
	Endpoint-adaptive RGB	0.5544	64.68	0.4224	<u>62.97</u>

excess increases. Figure 2 shows reduced deficits alongside clearer texture at matched positions; complete sequences appear in Appendix F.1.

We then evaluate image quality independently of the spectral objective. All three no-reference quality measures improve and FID decreases in every route–cohort combination (Table 2). On MorphBench, TOPIQ improves on 86–90 of 90 pairs per route, with larger mean gains on Metamorphosis than Animation. The same settings transfer to AH100 without retuning (Appendix B.2).

Transition trade-off. ECPI primarily targets the rendered quality of intermediate frames and incurs a modest trade-off in transition regularity (Table 6, Appendix B.1). Temporal smoothing of ECPI outputs trades quality for regularity: it lowers PPL, PDV and ISTD while lowering TOPIQ and MUSIQ (Appendix E.3).

Is the correction reducible to filtering? A natural alternative explanation is that ECPI could be an elaborate sharpening filter. We test this in two ways. First, as a practical baseline, we compare ECPI with fixed and endpoint-adaptive unsharp masking on 50 evaluation pairs per route, with fixed-filter parameters and the adaptive search rule frozen after development (Table 3; Appendix E.1). ECPI improves all three no-reference quality scores while reducing FID relative to baseline on both routes, and outperforms endpoint-adaptive RGB filtering on all four metrics. Fixed RGB filtering attains higher IQA scores, but at the cost of higher FID than the unprocessed baseline.

Second, we fit the filter family directly to ECPI's output, optimizing filter strength continuously at three matched positions on all 90 MorphBench pairs and three routes. Averaged over pairs within each route, more than 90% of the squared ECPI-induced RGB change remains unreproduced by the best fit (Appendix E.1, Table 12). Thus, the tested blur/sharpening family cannot reproduce most of ECPI's changes.

Why not correct only at the decoder input? A simpler alternative is to correct the baseline decoder latent z_0 after denoising, bypassing the remaining diffusion process during optimization. We optimize this latent to reduce the same positive four-band deficits, with cumulative movement capped at ECPI's realized net z_0 displacement. On the 14-pair GeoDiff comparison, mean positive deficit is 0.004 versus ECPI's 0.103, but excess energy is significantly higher and there is no clear mean TOPIQ advantage (Appendix E.2). Figure 14 in Appendix F.2 illustrates over-bright eye highlights and distorted eye contours under decoder-only refinement; Figure 4 gives additional selected comparisons. More complete deficit removal therefore does not automatically improve the rendered image. ECPI and decoder-only refinement realize the spectral request through different correction procedures and produce different spatial changes: the reference guides correction rather than fully specifying the desired image.

4.4 Local probability support (Q3)

What the constraint enforces, and what it costs. We test whether endpoint-directed spectral recovery remains effective when correction directions are constrained not to decrease conditional

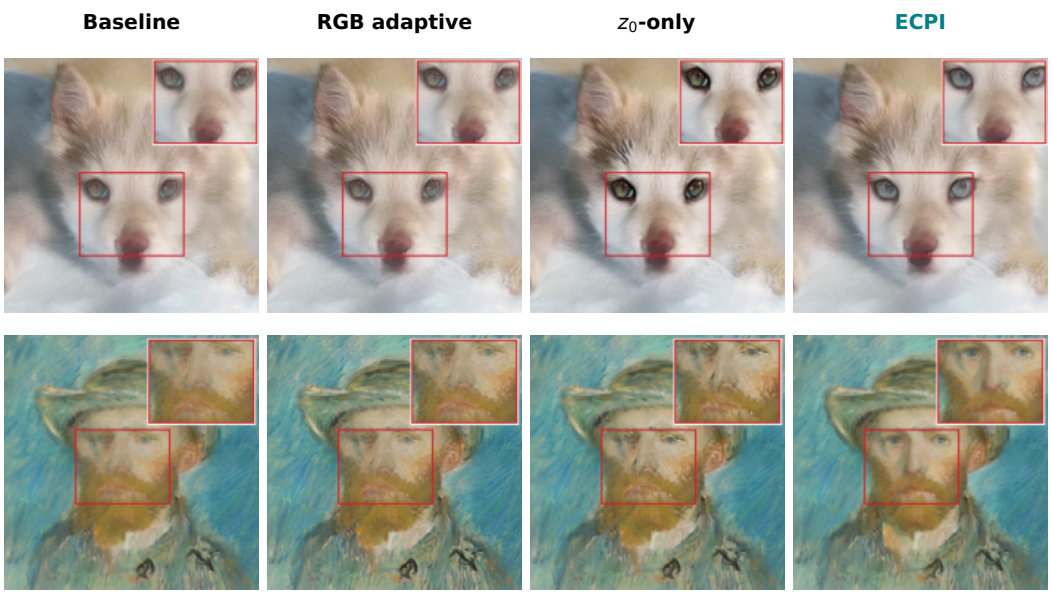

Figure 4: **The correction procedure shapes the rendered result.** Baseline, endpoint-adaptive RGB filtering, decoder-only z_0 refinement and ECPI at $s = 9/16$: animal photographs (AH100, Tangent; top) and Van Gogh self-portraits (MorphBench, SLERP; bottom). Insets show corresponding detail crops, using identical coordinates and display scales across methods within each row. Complete baseline and ECPI sequences appear in Appendix F.1.

Table 4: Recovery cost of local probability support (12 pairs per route). Quality and deficit entries are without-support minus supported ECPI under official-17 sampling, with paired 95% intervals. The last column gives mean relative log-density change per latent dimension, supported → without support, measured at 15 matched uniform interior states per pair.

Route	ΔTOPIQ ($\times 10^{-3}$)	ΔMUSIQ	ΔDeficit ($\times 10^{-3}$)	Mean $\Delta \log p/n_x$ ($\times 10^{-4}$) support → no support
GeoDiff	$+0.34$ $[+0.03, +0.74]$	$+0.050$ $[+0.018, +0.090]$	-2.80 $[-6.15, -0.59]$	$+0.15 \rightarrow -2.86$
Tangent	$+0.25$ $[-0.52, +0.91]$	$+0.067$ $[+0.015, +0.149]$	-3.41 $[-6.41, -1.01]$	$+0.31 \rightarrow -3.80$
SLERP	$+0.19$ $[-0.46, +0.72]$	$+0.051$ $[+0.018, +0.089]$	-2.96 $[-6.11, -0.78]$	$+0.25 \rightarrow -3.39$

log density to first order. On 12 frozen pairs per route, we remove only the score projection, keeping base paths, targets, route-normal projection, trust budgets and solver settings fixed, then recompute directions and doses and apply official-17 sampling (Table 4). Mean measured conditional log-density change is negative without support and positive with it on every route. Removing support slightly increases deficit recovery and some quality scores. The constraint changes the measured local density behavior at a small cost in spectral recovery, without an observed quality advantage in this ablation.

The constraint preserves spectral recovery. At equal small-step displacement on 90 pairs and three routes, the task direction reduces the deficit while a random route-normal direction of the same length produces no measurable reduction (Table 11). The task direction opposes the score in 95.0% of active midpoint states, yet projection retains 98.64% of the raw gain on conflicting states while changing mean measured relative log-density from negative to positive (Figure 5a; Table 11).

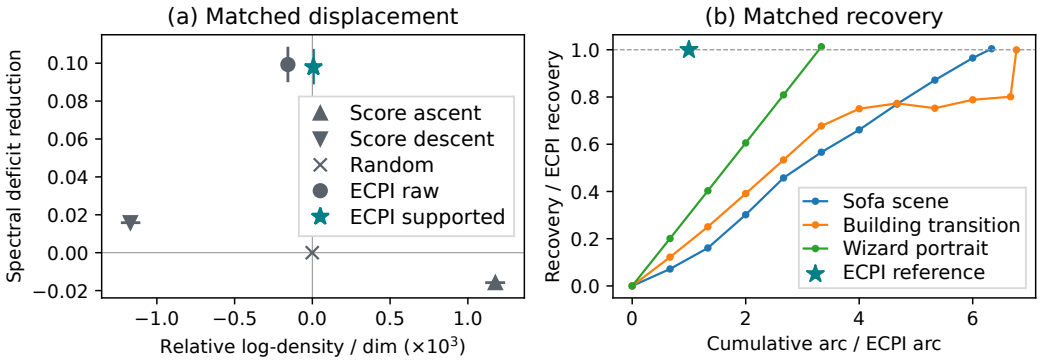

Figure 5: **Fixed-time correction diagnostics.** (a) Equal-displacement task, score and random directions on conflicting states; error bars are pair-bootstrap 95% intervals. (b) Negative-score trajectories on three selected states, normalized by ECPI's arc length and deficit recovery. The star marks ECPI; the horizontal line denotes matched recovery.

Table 5: Fixed-time score descent on 12 pairs per route. Arc ratios use successful matches. PPL and PDV include budget-limited paths after official-17 sampling. Endpoint LPIPS is the score-descent-minus-ECPI change in mean distance to the common reconstructed endpoints at matched uniform positions.

Route	Matched/ searched	Median arc ratio	PPL ECPI → descent	PDV ECPI → descent	Δ endpoint LPIPS
GeoDiff	114/159	5.79×	0.623 → 1.040	0.0269 → 0.0413	+0.023
Tangent	129/180	6.00×	0.646 → 1.137	0.0223 → 0.0381	+0.050
SLERP	128/177	6.33×	0.603 → 0.995	0.0231 → 0.0349	+0.045

Can score motion recover the same deficit? Finally, we ask whether moving along the score itself would recover detail. At equal displacement on conflicting states, score ascent raises density but worsens the deficit (Table 11). Recomputing the negative score after each step at all 15 interior knots of 12 pairs per route, 371 of 516 searched states reach ECPI's recovery. Successful matches use a median 5.79–6.33 times ECPI's cumulative latent arc (Figure 5b; Table 5). After official-17 resampling, the resulting paths have higher PPL and PDV, indicating greater accumulated perceptual change and less uniform transitions. At matched uniform positions, their outputs also lie farther from the shared reconstructed endpoints in mean LPIPS distance. Matching ECPI's spectral recovery therefore requires more latent motion and can introduce structural artifacts rather than simply restore existing detail, as illustrated by the additional floor markings and deformed dome contour in Figure 8. Additional measurements appear in Appendix D.

5 Conclusion

We study endpoint-relative spectral deficit as a rendered-space failure mode in diffusion image interpolation and introduce ECPI to correct it through the remaining generative process. Controlled interventions show that endpoint-defined spectral requests reliably shape the rendered response, while evaluations across SLERP, Tangent, and GeoDiff show consistent improvements in independent image-quality and distributional metrics. These results support endpoint-calibrated spectral deficit as a practical correction signal that complements existing latent-space interpolation criteria. Future work could extend endpoint-calibrated pullback to keyframe-conditioned video generation, using keyframe-derived rendered-space signals to guide detail preservation and temporal coherence.

Reproducibility Statement

We provide the implementation details, hyperparameters, spectral-band definitions, route correction procedure, and evaluation protocols in the appendix. All experiments use fixed pair manifests and saved endpoint inversions, and baseline and ECPI results share the same inversion, conditioning, and rendering protocols.We will publicly release the code upon acceptance.

Ethics statement

We do not identify ethical concerns specific to this work beyond those associated with image generation, including potential misuse to create misleading or manipulated imagery.

AI use Statement

Generative AI tools were used to aid and polish writing, assist with retrieval and discovery of related work, help draft portions of the manuscript, and provide limited feedback on experimental design and interpretation of results. All AI-assisted content was manually reviewed and verified by the authors.

References

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

Yukang Cao, Chenyang Si, Jinghao Wang, and Ziwei Liu. FreeMorph: Tuning-free generalized image morphing with diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18111–18120, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Cao_FreeMorph_Tuning-Free_Generalized_Image_Morphing_with_Diffusion_Model_ICCV_2025_paper.html.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 2024.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://arxiv.org/abs/2204.00227.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8188–8197, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Choi_StarGAN_v2_Diverse_Image_Synthesis_for_Multiple_Domains_CVPR_2020_paper.html.

Fabian Falck, Teodora Pandeva, Kiarash Zahirnia, Rachel Lawrence, Richard E. Turner, Edward Meeds, Javier Zazo, and Sushrut Karmalkar. A Fourier space perspective on diffusion models. *arXiv preprint arXiv:2505.11278*, 2025.

Qiyuan He, Jinghao Wang, Ziwei Liu, and Angela Yao. AID: Attention interpolation of text-to-image diffusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 97766–97799, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b12a1d1014e952e676f5d6931d03241a-Abstract-Conference.html.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Ahmed Imtiaz Humayun, Ibtihel Amara, Cristina Vasconcelos, Deepak Ramachandran, Candice Schumann, Junfeng He, Katherine Heller, Golnoosh Farnadi, Negar Rostamzadeh, and Mohammad Havaei. What secrets do your manifolds hold? understanding the local geometry of generative models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=etif9j1CnG.

Rafał Karczewski, Markus Heinonen, and Vikas K Garg. Devil is in the details: Density guidance for detail-aware generation with flow models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29098–29127, 2025. URL https://proceedings.mlr.press/v267/karczewski25a.html.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=Hk99zCeAb.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2206.13397.

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 (CVPR)*, 2022.

Shinnosuke Saito and Takashi Matsubara. Be tangential to manifold: Discovering riemannian metric for diffusion models. In *European Conference on Computer Vision (ECCV)*, 2026. URL https://arxiv.org/abs/2510.05509.

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

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://arxiv.org/abs/2011.13456.

Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Clinton J. Wang and Polina Golland. Interpolating between images with diffusion models. In *ICML Workshop on Challenges of Deploying Generative AI*, 2023. URL https://arxiv.org/abs/2307.12560.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.

Zhaoyuan Yang, Zhengyang Yu, Zhiwei Xu, Jaskirat Singh, Jing Zhang, Dylan Campbell, Peter Tu, and Richard Hartley. IMPUS: Image morphing with perceptually-uniform sampling using diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/08bd07d567d77d6dd8d82a4474706a5e-Abstract-Conference.html.

Qingtao Yu, Jaskirat Singh, Zhaoyuan Yang, Peter Henry Tu, Jing Zhang, Hongdong Li, Richard Hartley, and Dylan Campbell. Probability density geodesics in image diffusion latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2504.06675.

Kaiwen Zhang, Yifan Zhou, Xudong Xu, Bo Dai, and Xingang Pan. Diffmorpher: Unleashing the capability of diffusion models for image morphing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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 (CVPR)*, 2018.

Pengfei Zheng, Yonggang Zhang, Zhen Fang, Tongliang Liu, Defu Lian, and Bo Han. NoiseDiffusion: Correcting noise for image interpolation with diffusion models beyond spherical linear interpolation. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/31773c0ba7a4a98d729b9fc0d6d0cc13-Abstract-Conference.html.

A Implementation Details

All routes use Stable Diffusion v2.1-base with a 50-step deterministic DDIM schedule and inversion to $t = 581$. Continuation uses classifier-free guidance (CFG) (Ho & Salimans, 2022) at scale 1 (conditional prediction without extrapolation), with $\eta = 0$.

A.1 Global spectral readout

For each RGB channel we subtract its spatial mean and apply a separable Hann window with non-periodic endpoints. For an $H \times W$ image, the channel-mean squared magnitude of the resulting 2-D FFT is divided by $(HW)^2$ to obtain P_y. Radial frequency is normalized by the corner Nyquist frequency, $\omega = \sqrt{f_x^2 + f_y^2}/\sqrt{0.5^2 + 0.5^2}$, where f_x, f_y are measured in cycles per pixel. The four bands are $[0, .10)$, $[.10, .25)$, $[.25, .50)$ and $[.50, 1.01)$. Using floating-point decoder output before clipping or PNG quantization, we compute

$$h_k(y) = \log\Big(\max\{\sum_{\omega \in \mathcal{B}_k} P_y(\omega), \varepsilon\}\Big), \qquad k = 1, \ldots, 4. \tag{6}$$

with $\varepsilon = 10^{-12}$.

A.2 Route-normal direction and band weights

Let $r_i = x_i/\|x_i\|_2$. We estimate the path tangent from $x_{i+1} - x_{i-1}$, remove its radial component, and normalize it to obtain τ_i (zero for a degenerate difference). The orthogonal projector and projected band gradients are

$$Q_i = I - r_i r_i^\top - \tau_i \tau_i^\top, \qquad a_{ik} = Q_i g_{ik}. \tag{7}$$

Within trust radius ρ, band weights are capped by the maximum linearized increase $\rho\|a_{ik}\|_2$:

$$\alpha_{ik} = \min([e_{ik}]_+, \rho\|a_{ik}\|_2), \qquad u_i = \sum_{k=1}^{4} \alpha_{ik}\, \widehat{a}_{ik}, \tag{8}$$

where $\widehat{a}_{ik} = a_{ik}/\|a_{ik}\|_2$ for nonzero gradients and zero otherwise.

The conditional score estimate is $\widehat{s}_t(x, c) = -\epsilon_\theta(x, t, c)/\sqrt{1-\bar{\alpha}_t}$, using the model's conditional noise prediction (Ho et al., 2020; Song et al., 2021b). For $v_i = Q_i \widehat{s}_t(x_i, c_i)$, the half-space projection in Eq. (5) has the closed form

$$\widetilde{d}_i = \begin{cases} u_i, & \langle u_i, v_i \rangle \geq 0, \\ u_i - \dfrac{\langle u_i, v_i \rangle}{\|v_i\|_2^2} v_i, & \text{otherwise.} \end{cases} \tag{9}$$

If $v_i = 0$, the first branch applies. Nonzero projected directions are normalized to d_i; zero directions receive zero dose.

A.3 Route correction and output

We use $N = 16$ intervals, giving 17 knots and 15 interior correction variables. For magnitudes $q_i \in [0, 1]$, the predicted band change is $q_i b_{ik}$, where $b_{ik} = \rho\langle g_{ik}, d_i\rangle$. Let $w_i = q_i d_i$, $w_0 = w_N = 0$, $\Delta w_i = w_{i+1} - w_i$, and $\Delta^2 w_i = w_{i+1} - 2w_i + w_{i-1}$. We minimize

$$\begin{aligned} \min_{q \in [0,1]^{N-1}} \quad & \frac{1}{4(N-1)} \sum_{i=1}^{N-1} \sum_{k=1}^{4} \mathbf{1}[e_{ik} > 0] \, (e_{ik} - q_i b_{ik})^2 \\ & + \frac{\lambda_v}{N} \sum_{i=0}^{N-1} \|\Delta w_i\|_2^2 + \frac{\lambda_a}{N-1} \sum_{i=1}^{N-1} \|\Delta^2 w_i\|_2^2 \end{aligned} \tag{10}$$

Directions and active bands remain fixed after one linearization; only q is optimized. We use $\rho = 0.0075\|x_{B,t} - x_{A,t}\|_2$, $\lambda_v = 0.01$ and $\lambda_a = 0.02$. These values were selected on 12 development pairs by remaining deficit at matched arc length and held fixed across routes and datasets for the full ECPI evaluations.

The dose solver uses L-BFGS-B with bounds $[0, 1]$, at most 500 iterations, function tolerance 10^{-12} and gradient tolerance 10^{-9}. With model weights frozen, we save one continuation trajectory per knot and compute one reverse-mode pullback per band, recomputing each denoising-step graph for its vector–Jacobian product rather than retaining the full graph.

For nonzero d_i, the norm-preserving update is

$$x_i' = \cos\theta_i \, x_i + \sin\theta_i \, \|x_i\|_2 d_i, \qquad \theta_i = \rho q_i / \|x_i\|_2. \tag{11}$$

Zero-dose knots and both endpoints are unchanged.

The corrected route evaluates positions by SLERP between neighboring corrected knots. Output uses the shared official-17 protocol: render at 17 uniform positions, accumulate adjacent LPIPS-VGG distances, invert that cumulative function at uniform quantiles, and render at the resulting positions with endpoints fixed.

Computational workload. Correction requires $15 \times 30 \times 4 = 1800$ denoising-step VJPs (vector-Jacobian products) per pair and route: 15 interior states, 30 remaining DDIM steps and four bands.

B Datasets, Evaluation Protocols and Additional Statistics

B.1 Evaluation protocols and aggregation

TOPIQ (Chen et al., 2024), MUSIQ (Ke et al., 2021) and MANIQA (Yang et al., 2022) assign a no-reference quality score to each image; higher values indicate better predicted quality. We follow GeoDiff's position-weighted TOPIQ aggregation (Yu et al., 2025) and apply the same weights to MUSIQ and MANIQA. For the main results, objective comparisons, score-support tests and smoothing diagnostic, we aggregate each pair's 15 interior frames at their actual resampled positions:

$$\overline{M} = \frac{\sum_{i=1}^{15} \min(s_i, 1 - s_i) M(y_i)}{\sum_{i=1}^{15} \min(s_i, 1 - s_i)}, \tag{12}$$

where $M \in \{\mathrm{TOPIQ}, \mathrm{MUSIQ}, \mathrm{MANIQA}\}$ and s_i is the actual path position. Single-state scores require no position aggregation. Frame-level scores are aggregated within each pair; means and

bootstrap intervals use complete pairs. Bootstrap intervals for the IQA comparisons use $20{,}000$ whole-pair resamples.

Transition inputs are resized to 256×256. PPL sums the adjacent-frame LPIPS-Alex distances, PDV is their standard deviation, and ISTD is the standard deviation of adjacent-frame pixel-RMS distances. Table 6 reports all routes and cohorts: ECPI increases cumulative perceptual change, with route-dependent effects on transition uniformity.

Distributional evaluation. FID compares all 15 official-17 interior frames per pair with the original input endpoints from the same cohort. All methods within a cohort share the same reference set.

Table 6: Transition statistics under the official 17-frame output protocol. Animation (Anim.), Metamorphosis (Meta.), and AH100 contain 24, 66, and 100 pairs per route. Entries are absolute means; "+" denotes ECPI. Bold marks the lowest value within each route, including ties.

	PPL↓			PDV↓			ISTD↓		
Method	Anim.	Meta.	AH100	Anim.	Meta.	AH100	Anim.	Meta.	AH100
GeoDiff	**0.2500**	**0.7455**	**0.6529**	**0.0089**	**0.0282**	**0.01976**	**0.0131**	**0.0253**	**0.02107**
+ ECPI	0.2536	0.7552	0.6561	0.0092	0.0288	0.02020	0.0135	0.0258	0.02132
Tangent	**0.2438**	**0.7504**	**0.6249**	**0.0084**	**0.0227**	**0.01592**	**0.0097**	**0.0235**	**0.01934**
+ ECPI	0.2658	0.7867	0.6448	0.0101	0.0252	0.01691	0.0114	0.0241	0.01978
SLERP	**0.2529**	**0.7118**	**0.6085**	**0.0077**	**0.0211**	**0.01534**	**0.0104**	**0.0225**	**0.02014**
+ ECPI	0.2625	0.7370	0.6206	0.0081	0.0232	0.01616	0.0111	0.0229	0.02032

B.2 AH100 CONSTRUCTION AND PAIRED RESULTS

Protocol. AH100 contains 50 AFHQ animal pairs and 50 CelebA-HQ face pairs, sampled from the validation images distributed with StarGAN v2. The AFHQ portion includes 25 within-category and 25 cross-category pairs. A frozen manifest fixes pair identities, prompts and one seed per pair. All pairs use the same backbone, correction settings and official-17 protocol as MorphBench (Appendix A).

Overall quality and transition changes. Across all 100 pairs, quality gains have positive paired intervals on all three routes, while PPL increases; PDV and ISTD changes vary by route (Table 7).

C SPECTRAL REQUESTS AND TARGET DESIGN

C.1 ENDPOINT REQUESTS AND RENDERED RESPONSES

Deficits along the interpolation path. Figure 6 reports the mean positive four-band log-energy deficit at each of the 15 uniform interior positions, using all 90 MorphBench and 100 AH100 pairs on each route. For each baseline path, we compare central positions $i = 6, \ldots, 10$ with near-boundary interior positions $i \in \{1, 2, 3, 13, 14, 15\}$, where $s_i = i/16$. The mean central-minus-boundary difference is positive in all six dataset–route groups, with pair-bootstrap 95% intervals above zero. The endpoints themselves are excluded from this comparison.

Observed responses. We analyze all 90 MorphBench and 100 AH100 pairs on three routes at the 15 uniform interior positions, giving 8,550 states. At each state and band, the initial deficit is $[h_k^\star - h_k(y_i^0)]_+$ and the measured response is $\Delta h_{ik} = h_k(y_i^{\mathrm{ECPI}}) - h_k(y_i^0)$. Both use matched floating-point rendered readouts. Of 21,233 initially deficient state–band observations, 21,164 have positive responses. No-update counts use the controller's recorded requests and actual latent displacements: all 772 states with no active requests have zero displacement. Non-requested bands can change through the combined update, as illustrated in Figure 1.

Table 7: AH100 overall paired changes, ECPI minus baseline. Intervals are pointwise 95% whole-pair bootstrap percentile intervals. IQA scores use position weighting. "Improved" counts positive quality changes and negative transition changes, using each metric's preferred direction.

Route	Metric	ECPI − baseline	95% CI	Improved / 100
GeoDiff	TOPIQ	+0.01045	[+0.00840, +0.01252]	92
	MUSIQ	+0.64902	[+0.50294, +0.80616]	91
	MANIQA	+0.00579	[+0.00459, +0.00702]	93
	PPL	+0.00315	[+0.00147, +0.00481]	31
	PDV	+0.00044	[−0.00011, +0.00108]	42
	ISTD	+0.00024	[−0.00005, +0.00056]	41
Tangent	TOPIQ	+0.01586	[+0.01332, +0.01850]	94
	MUSIQ	+1.79925	[+1.49049, +2.12454]	96
	MANIQA	+0.01287	[+0.01081, +0.01501]	94
	PPL	+0.01990	[+0.01584, +0.02432]	13
	PDV	+0.00100	[+0.00032, +0.00169]	45
	ISTD	+0.00044	[+0.00013, +0.00077]	32
SLERP	TOPIQ	+0.01370	[+0.01126, +0.01614]	92
	MUSIQ	+1.39292	[+1.14768, +1.64708]	92
	MANIQA	+0.00990	[+0.00807, +0.01174]	88
	PPL	+0.01210	[+0.00893, +0.01536]	23
	PDV	+0.00083	[+0.00024, +0.00146]	42
	ISTD	+0.00018	[−0.00021, +0.00058]	48

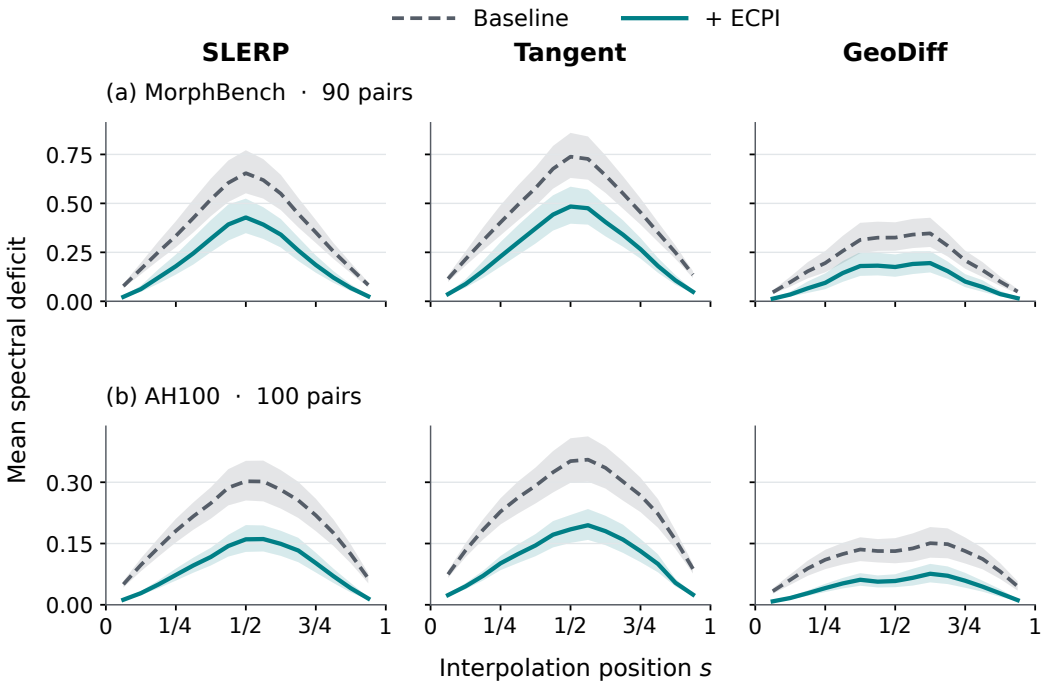

Figure 6: **Spectral deficits along interpolation paths.** Mean positive four-band log-energy deficits before and after ECPI at 15 matched uniform interior positions. Rows show 90 MorphBench and 100 AH100 pairs per route. Shading denotes pointwise 95% pair-bootstrap intervals; vertical scales are shared within each dataset.

Spatial response visualization. Spatial maps in Figures 1 and 13, and the bars in Figure 1, are computed from saved display RGB images. Using the same centering, Hann window and band masks as Appendix A, let $E_k^a(u)$ be the channel-mean squared magnitude of the inverse FFT of

masked band k for arm a. The spatial field is

$$H_k(u) = \frac{E_k^{\mathrm{ECPI}}(u) - E_k^{\mathrm{base}}(u)}{\langle E_k^{\mathrm{base}} \rangle_u}.$$

For display, we apply a Gaussian with $\sigma = 2$ pixels and periodic boundaries, then average 4×4 blocks. Within each example, all bands share a symmetric logarithmic color scale, linear within $[-1, 1]$; near-zero responses are transparent. The maps separate the observed joint output change by frequency. Exact display settings are saved with the figure sources.

Reference-input intervention. We use 12 frozen development pairs on Tangent. At $s = 1/2$ and $t = 581$, we replace one reference component by $h_k^{\star,(\ell)} = h_{0,k}^{\mathrm{ctrl}} + \ell$, with $\ell \in \{-0.05, 0.05, 0.15\}$ and $h_{0,k}^{\mathrm{ctrl}}$ the cached baseline readout. All other reference components, endpoint images, latent states, conditions, band gradients and trust budgets remain fixed. Each arm recomputes the combined direction, score projection and joint doses, then renders all 15 interior states with the unchanged continuation and decoder. The sweep contains 144 intervention paths; the original-reference and uncorrected paths are retained. This isolates the controller's response to its reference input while the images defining the original pair stay fixed.

Figure 7 compares actual output log-band energies between request levels, rather than errors against the changing reference. Intervals use 10,000 paired bootstrap draws. Increasing an active request strengthens its corresponding band in every tested pair. Requests share a constrained latent update, so cross-band changes are present. Activating a previously inactive band changes the combined correction direction and the optimized update magnitude. Its marginal response therefore need not be positive, unlike the consistently positive responses observed when strengthening an already active request in these tests. These measurements characterize control behavior; independent rendered-quality results are reported in Table 2.

C.2 Objective comparisons

The signed four-band ablation applies the same endpoint-derived spectral reference as ECPI and corrects both positive and negative residuals. All variants share the official-17 output protocol and position-weighted IQA aggregation.

One-sided floor versus signed setpoint. We test whether above-reference energy should also be suppressed by comparing ECPI with the signed variant. With the reference and solver unchanged, correcting only shortfalls yields higher pooled TOPIQ on every route and reduces the number of GeoDiff pairs whose interior-mean TOPIQ falls below the baseline from 29 to 2 (Table 8). The signed variant retains slightly lower FID in several cohorts. Full quality and transition comparisons follow below.

Table 8: Floor minus signed setpoint on 90 MorphBench pairs per route under official-17 sampling. Degraded pairs have lower interior-mean TOPIQ than the base route; the last column reports the floor's gain over baseline on the subset degraded by signed correction.

Route	ΔTOPIQ	Degraded pairs signed → floor	ECPI gain on degraded subset
GeoDiff	+0.0062	29 → 2	+0.0045
Tangent	+0.0033	12 → 2	+0.0055
SLERP	+0.0042	12 → 0	+0.0088

D Fixed-Time Latent Correction Diagnostics

Equal displacement. The ECPI audit uses saved baseline paths and reconstructed-endpoint band targets for all 90 pairs on GeoDiff, Tangent and SLERP. At $s = 0.5$ and $t = 581$, raw ECPI is exactly the reachable-weight four-band positive direction, with weights computed at the deployment trust radius 0.0075 times endpoint chord. Interventions use a common arc length of 0.0025 times that

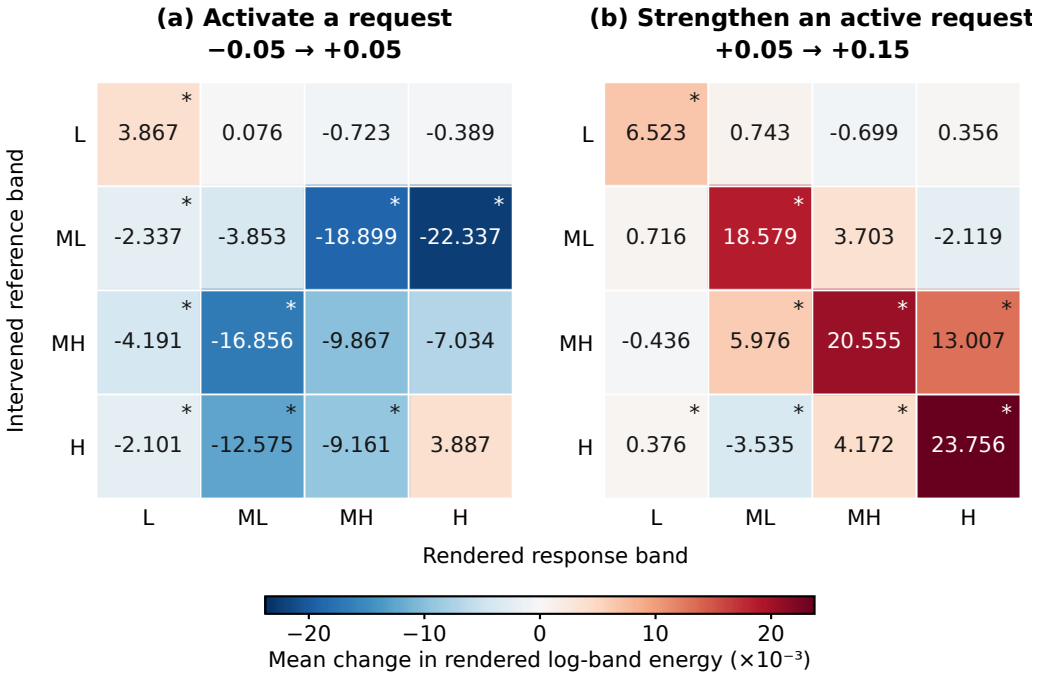

Figure 7: **Rendered responses to reference-input interventions.** Rows identify the intervened band and columns the measured band at the midpoint of 12 Tangent pairs. Values are paired mean log-energy changes ($\times 10^{-3}$) between the two intervention arms, with a shared color scale. Panels compare activating a request and strengthening an active request. Asterisks mark pointwise 95% pair-bootstrap intervals excluding zero.

Table 9: Full objective comparisons, with absolute Animation (Anim.) and Metamorphosis (Meta.) scores. Bold and underline indicate the best and second-best values within each base route.

	FID↓		TOPIQ↑		MUSIQ↑		MANIQA↑	
Method	Anim.	Meta.	Anim.	Meta.	Anim.	Meta.	Anim.	Meta.
GeoDiff	38.24	58.13	0.5246	0.5779	66.51	66.10	0.4009	0.4463
+ Signed four-band	**37.68**	57.59	0.5286	0.5886	67.00	67.41	0.4044	0.4548
+ ECPI	37.74	**57.58**	**0.5332**	**0.5954**	**67.27**	**67.69**	**0.4081**	**0.4611**
Tangent	41.33	65.13	0.4946	0.5189	62.56	58.05	0.3666	0.3871
+ Signed four-band	**39.59**	63.49	0.5058	0.5463	63.74	61.25	0.3769	0.4114
+ ECPI	40.04	**63.43**	**0.5078**	**0.5501**	**63.87**	**61.42**	**0.3786**	**0.4144**
SLERP	41.96	62.16	0.4964	0.5449	63.09	60.52	0.3726	0.4085
+ Signed four-band	**40.25**	61.69	0.5089	0.5659	64.20	63.08	0.3824	0.4275
+ ECPI	40.67	**61.58**	**0.5118**	**0.5706**	**64.37**	**63.25**	**0.3846**	**0.4316**

chord. The five arms are normalized raw ECPI, score-supported ECPI, positive Q-score, negative Q-score, and a seeded isotropic random direction projected through Q. Each uses the same condition and deterministic continuation. Task improvement is the reduction in mean positive log-band residual, recomputed over all four bands after intervention.

Zero ECPI directions are retained as no-op states and excluded from conflict rates and gain ratios. Overall effects average routes within each pair before bootstrapping pairs (10,000 resamples). Retention is the ratio of supported to raw pair-mean deficit improvement; its interval bootstraps the ratio. Relative log-density changes in these diagnostics are learned-score line-integral estimates at

Table 10: Transition statistics for the objective baselines and ECPI on Animation (Anim.) and Metamorphosis (Meta.).

	PPL↓		PDV↓		ISTD↓	
Method	Anim.	Meta.	Anim.	Meta.	Anim.	Meta.
GeoDiff	**0.2500**	**0.7455**	**0.0089**	**0.0282**	**0.0131**	**0.0253**
+ Signed four-band	0.2522	0.7536	0.0094	0.0288	0.0136	0.0257
+ ECPI	0.2536	0.7552	0.0092	0.0288	0.0135	0.0258
Tangent	**0.2438**	**0.7504**	**0.0084**	**0.0227**	**0.0097**	**0.0235**
+ Signed four-band	0.2646	0.7862	0.0101	0.0253	0.0116	0.0242
+ ECPI	0.2658	0.7867	0.0101	0.0252	0.0114	0.0241
SLERP	**0.2529**	**0.7118**	**0.0077**	**0.0211**	**0.0104**	**0.0225**
+ Signed four-band	0.2618	0.7341	0.0083	0.0232	0.0114	0.0229
+ ECPI	0.2625	0.7370	0.0081	0.0232	0.0111	0.0229

fixed t, c. For connector points $\xi_0, \ldots, \xi_J$, we compute

$$\Delta\widehat{\ell} = \sum_{j=0}^{J-1} \frac{\widehat{s}_t(\xi_j, c) + \widehat{s}_t(\xi_{j+1}, c)^\top}{2} (\xi_{j+1} - \xi_j), \qquad n_x = 4 \cdot 64 \cdot 64. \tag{13}$$

We report $\Delta\widehat{\ell}/n_x$, where n_x is latent dimensionality. Eight-segment linear and spherical connectors are checked against four segments. Both connectors give positive changes at all 258 active supported states; the median eight-versus-four discrepancy is 3.97×10^{-9} per dimension, and the median connector discrepancy is 2.10×10^{-9}. Table 11 uses the spherical connector.

Across the 270 midpoints, 258 have nonzero task directions and 245 (95.0% of active states) oppose the score; mean cosine is -0.134. Table 11 reports the conflicting-state results, aggregated within 88 pairs. Projection retains 98.64% of the raw task gain (95% CI $[98.44, 98.83]\%$).

Table 11: Equal-displacement ECPI mechanism audit on 245 conflicting states, averaged within 88 endpoint pairs. Positive deficit reduction means improvement. All directions use arc length 0.0025 times endpoint chord.

Direction	Deficit reduction ↑	$10^3\Delta\hat{\ell}/n_x$
Score ascent	-0.01580	$+1.1768$
Score descent	$+0.01583$	-1.1705
Random	-0.00000	-0.0007
ECPI raw	$+0.09930$	-0.1573
ECPI supported	$+0.09795$	$+0.0070$

Fixed-time score-descent protocol. We match ECPI's mean positive four-band deficit at each knot, keeping interpolation position, diffusion time, conditioning and continuation fixed. Each step refreshes $d^- = -Q\nabla \log p_t / \|Q\nabla \log p_t\|_2$, recomputing the radial projection and reprojecting the initial route tangent. Steps use 0.5% of the endpoint chord, with a 12% cumulative arc cap. States already satisfying the matching tolerance remain unchanged. Otherwise, we stop at the first match, using the larger of 10^{-4} and 1% of ECPI's mean positive deficit as tolerance, with bisection at the first crossing. Cumulative arc measures latent path length, not net displacement.

Selected trajectories. The GeoDiff examples show a sofa scene and a building transition at $s = 9/16$, and a wizard portrait at $s = 5/16$. ECPI uses 0.75% endpoint chord in each case; all three score-descent searches reach tolerance, although their per-band deficits remain different.

Full-path matched recovery. We apply the same search to 15 interior knots of 12 frozen pairs on GeoDiff, Tangent and SLERP, giving 540 states. Of these, 24 require no update. Of the 516 searched states, 371 reach tolerance, 142 stop at the cap and 3 stop at the refinement limit. All states, including budget-limited ones, are assembled into complete paths before official rendered-LPIPS resampling.

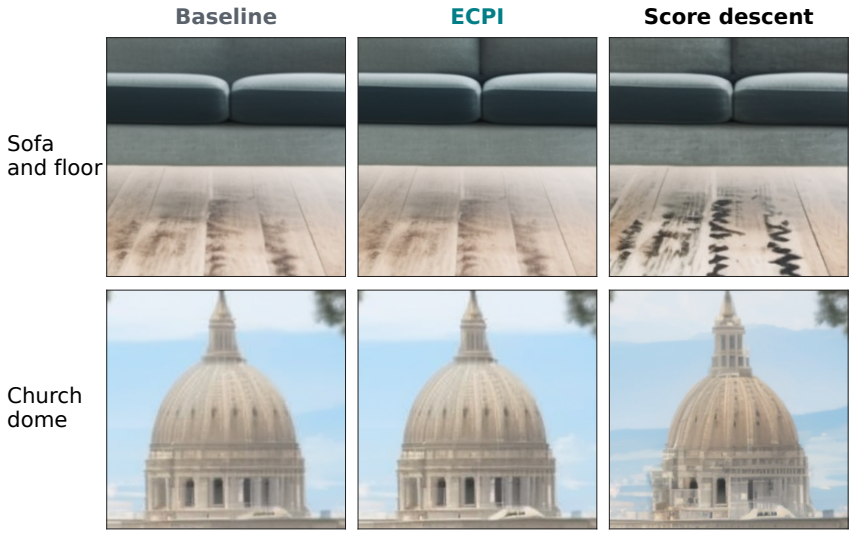

Figure 8: **Matched spectral recovery produces different structural changes.** Two selected GeoDiff states at fixed interpolation positions, with identical crop coordinates and enlargement. Relative to ECPI, score descent strengthens floor markings (top) and changes the dome outline (bottom).

Table 5 reports the complete-path comparison. PPL sums adjacent-frame LPIPS distances after each method's official-17 resampling. Endpoint proximity instead uses matched uniform positions $s_i = i/16$ and the shared baseline reconstructions $\widetilde{y}_A, \widetilde{y}_B$:

$$L_{\text{end}} = \frac{1}{30} \sum_{i=1}^{15} \big[\text{LPIPS}(y_{s_i}, \widetilde{y}_A) + \text{LPIPS}(y_{s_i}, \widetilde{y}_B)\big].$$

Distances use LPIPS-Alex at 256×256. The endpoint-LPIPS column reports score descent minus ECPI; positive values indicate greater mean distance from the shared endpoints. The PPL and PDV increases have positive paired intervals on all three routes. For endpoint LPIPS at identical uniform positions, the score-descent-minus-ECPI intervals are $[0.0068, 0.0407]$ for GeoDiff, $[0.0137, 0.0964]$ for Tangent and $[0.0100, 0.0872]$ for SLERP. Negative score also produces higher spectral excess and significantly higher MUSIQ on all three routes, while its TOPIQ difference from ECPI spans zero.

E Output-Side and Decoder-Latent Controls

We compare correction through the remaining generator with RGB filtering and decoder-latent refinement, then evaluate the quality–regularity trade-off of temporal smoothing on corrected sequences.

E.1 RGB filtering

On 50 evaluation pairs per route (GeoDiff and Tangent), both RGB controls use

$$y' = \text{clip}_{[0,1]}\big(y + a[y - G_\sigma * y]\big), \tag{14}$$

where G_σ is a Gaussian kernel with scale σ in pixels. Fixed filtering uses $(\sigma, a) = (0.7, 2)$, selected by mean development TOPIQ. Endpoint-adaptive filtering selects each frame's parameters from $\sigma \in \{0.7, 1.5, 3\}$ and $a \in \{-1, -0.5, -0.25, 0, 0.25, 0.5, 1, 2\}$ to minimize four-band log-energy MSE against its endpoint-derived reference at the recorded s. Selection evaluates clipped, 8-bit-quantized outputs. The fixed setting and adaptive grid and rule are established on a separate development set and frozen for evaluation.

RGB filtering preserves the baseline's official-17 positions and endpoints, without further resampling; ECPI uses its own official-17 outputs. TOPIQ uses actual path positions; FID uses cohort-matched original endpoints.

Figure 4 applies the adaptive rule at matched uniform positions, using log-linearly interpolated reconstructed endpoint PNG spectra as references. Selected settings are $(3, 0.25)$ for the animal photographs and $(3, 0.5)$ for the Van Gogh self-portraits.

Best-fit reconstruction of ECPI updates. We fit the same filter family directly to ECPI outputs on all 90 MorphBench pairs and three routes, at matched positions $s \in \{1/4, 1/2, 3/4\}$. For each $\sigma \in \{0.7, 1.5, 3\}$, we minimize RGB squared error over continuous $a \in [-1, 2]$ by solving the clipped, piecewise-quadratic objective exactly and retain the σ with the lowest error. Fitting compares floating-point filtered RGB with the ECPI PNG; fits are quantized only for viewing and LPIPS.

Let $y^0_{p,s}$, $y^E_{p,s}$ and $\widehat{y}_{p,s}$ denote baseline, ECPI and best-fit outputs. Within pair p, we pool the active sampled states:

$$\eta_p^2 = \frac{\sum_s \|\widehat{y}_{p,s} - y^E_{p,s}\|_2^2}{\sum_s \|y^0_{p,s} - y^E_{p,s}\|_2^2}, \qquad c_p = 1 - \eta_p^2. \tag{15}$$

Excluding 36 no-update states leaves 774 active states. We average c_p over pairs within each route and bootstrap whole pairs 10,000 times. The secondary LPIPS ratio is the pair-wise sum of fit–ECPI distances divided by the sum of baseline–ECPI distances, evaluated on 256×256 PNGs without optimizing LPIPS. Excluding changes below $1/255$ RGB RMS gives similar explained fractions (6.4%–10.3%).

Table 12: **Best-fit RGB reconstruction of ECPI changes.** All 90 MorphBench pairs are evaluated per route; active pair/state counts exclude unchanged cases. Entries are pair means with 95% bootstrap intervals. Explained RGB change is $100c_p$; the LPIPS ratio measures the remaining perceptual difference.

Route	Active pairs / states	Explained RGB change (%) mean [95% CI]	Remaining LPIPS ratio mean [95% CI]
GeoDiff	89 / 241	9.5 [8.0, 11.1]	0.890 [0.867, 0.912]
Tangent	90 / 267	6.5 [5.4, 7.6]	0.912 [0.889, 0.933]
SLERP	90 / 266	8.0 [6.8, 9.3]	0.899 [0.875, 0.922]

Figure 9 illustrates two $s = 9/16$ states outside the quantitative anchors. This diagnostic tests reconstruction by the chosen filter family; Table 3 compares rendered quality.

E.2 Decoder-latent refinement

Pairs and protocol. The GeoDiff comparison pools the frozen random-12 cohort with two additional illustrative pairs, cinematic portraits and Van Gogh self-portraits, giving 14 distinct pairs. Selected-case SLERP and Tangent runs supply qualitative comparisons only. The animal illustration in Figure 4 applies the same twenty-step rule to the saved Tangent state of the AH100 animal photograph pair at $s = 9/16$.

The quantitative comparison uses the same 17 uniform path positions and reconstructed endpoint targets. For each frame, direct z_0 refinement receives a maximum cumulative movement budget equal to the net z_0 displacement induced by ECPI. Twenty updates split this allowance evenly, recomputing the gradient of the mean squared positive log-band deficit after each step and projecting into the same-radius ball around the initial z_0. Endpoints remain fixed; zero budget yields no update. All variants are decoded with the same FP32 VAE. The decoder-only control operates directly in z_0, without the x_t score projection or coupled dose solver.

Metrics. Deficit and excess use floating-point RGB before clipping. TOPIQ and transition metrics use the saved PNGs, at the same uniform positions without output resampling. Interior-frame

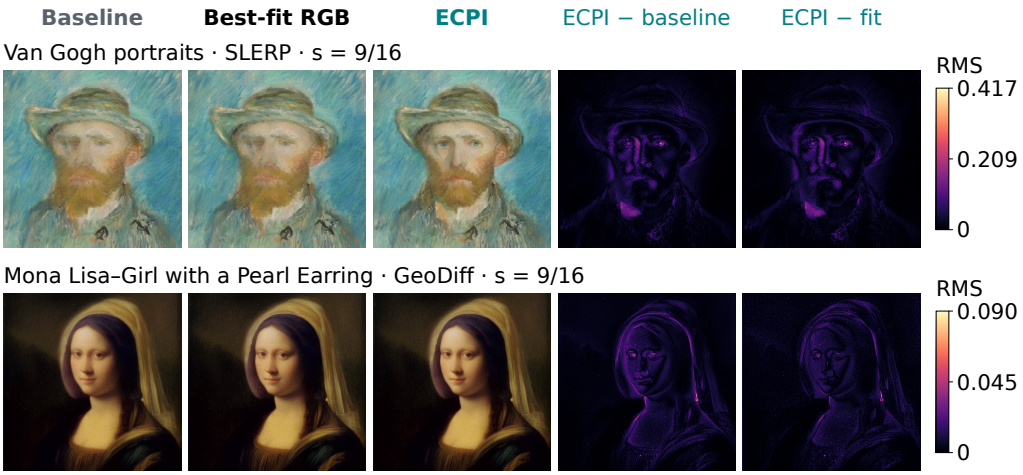

Figure 9: **Changes remaining after best-fit RGB reconstruction.** Van Gogh self-portraits with SLERP (top) and Mona Lisa to Girl with a Pearl Earring with GeoDiff (bottom), both at $s = 9/16$. Each row compares baseline, fitted RGB filtering and ECPI, followed by the original RGB-change magnitude and the residual after fitting. The two heatmaps share a linear scale within each row.

statistics are averaged within each pair and then equally across the 14 pairs. Reported difference intervals bootstrap pairs 20,000 times.

Table 13: Pooled fourteen-pair GeoDiff decoder-latent comparison. Deficit is $\mathrm{mean}([e]_+)$ and excess is $\mathrm{mean}([-e]_+)$; both are nonnegative. TOPIQ is the unweighted interior-frame mean.

Method	Deficit↓	Excess	TOPIQ↑	PPL	ISTD
Baseline	0.2012	0.0590	0.5947	0.6085	0.0291
ECPI	0.1033	0.0844	0.6074	0.6188	0.0311
z_0 refinement (20 steps)	0.0036	0.1110	0.6064	0.6199	0.0320

Relative to ECPI, direct refinement reduces deficit by 0.0997 (95% CI $[-0.1445, -0.0619]$ for the direct-minus-ECPI difference), with 0.0265 more excess ($[0.0109, 0.0416]$). Mean TOPIQ differs by -0.0010 (95% CI $[-0.0048, 0.0034]$, direct refinement minus ECPI), showing no resolved mean TOPIQ advantage despite more complete deficit removal.

Figure 4 compares RGB filtering, decoder refinement and ECPI at matched positions; Appendix F.2 provides additional artifact examples.

E.3 Temporal RGB smoothing

Protocol. We test temporal smoothing on the saved ECPI outputs for all 90 pairs on GeoDiff and Tangent after official-17 resampling. Each strength acts simultaneously on the same original sequence:

$$\widehat{y}_i = (1-\alpha)y_i + \frac{\alpha}{2}(y_{i-1} + y_{i+1}), \qquad i = 1, \ldots, 15, \quad \alpha \in \{0.10, 0.25, 0.50\}. \tag{16}$$

Endpoints and the 17 output positions are fixed. Mixing uses stored RGB values, rounded once to 8-bit PNG, with no subsequent resampling or diffusion inference. Transition metrics follow Appendix B.1. TOPIQ and MUSIQ use Eq. (12), with each smoothed sequence inheriting the actual positions of its ECPI input, followed by an equal mean over pairs; HFE is the channel-pooled variance of the Laplacian response on saved RGB images. All compared arms use the same weighting rule, with 20,000 whole-pair bootstrap draws.

Table 14: Temporal RGB smoothing of ECPI official-17 outputs, averaged over 90 pairs per route. Each smoothing strength is compared with the same unsmoothed ECPI sequence. IQA scores are position-weighted; HFE is reported as a frequency diagnostic.

Output	PPL↓	PDV↓	ISTD↓	TOPIQ↑	MUSIQ↑	HFE$\times 10^3$
GeoDiff (90 pairs)						
Base route	0.6134	0.02302	0.02206	0.5637	66.21	3.064
Base route + ECPI	0.6214	0.02361	0.02256	0.5788	67.58	3.377
+ smoothing ($\alpha = 0.10$)	0.5919	0.02188	0.02075	0.5740	67.26	3.240
+ smoothing ($\alpha = 0.25$)	0.5433	0.01842	0.01795	0.5659	66.70	3.042
+ smoothing ($\alpha = 0.50$)	0.4829	0.01228	0.01426	0.5541	65.80	2.803
Tangent (90 pairs)						
Base route	0.6153	0.01891	0.01983	0.5124	59.26	2.263
Base route + ECPI	0.6478	0.02120	0.02069	0.5388	62.08	2.682
+ smoothing ($\alpha = 0.10$)	0.6134	0.01954	0.01962	0.5342	61.75	2.574
+ smoothing ($\alpha = 0.25$)	0.5544	0.01639	0.01803	0.5264	61.21	2.414
+ smoothing ($\alpha = 0.50$)	0.4812	0.01090	0.01595	0.5155	60.39	2.225

Measured trade-off. At all three tested strengths, smoothing reduces mean PPL, PDV and ISTD alongside TOPIQ, MUSIQ and HFE, with paired intervals excluding zero on both routes (Table 14). Smoothing therefore adjusts ECPI's quality–regularity trade-off rather than eliminating it.

F Complete Qualitative Sequences

F.1 Main-figure examples

Examples in Figure 2 were selected by visual inspection of paired outputs. The sequences below provide complete context for the animal and Van Gogh examples. Both variants use the same 17 uniform positions $s = i/16$, before output-stage perceptual resampling; reconstructed endpoints are unchanged. Each full sequence is split into positions 0–8 and 9–16, with baseline above ECPI in each block. Gray and teal borders mark $s = 8/16, \ldots, 11/16$. The main-figure curves use saved floating-point readouts at the 15 interior knots without smoothing. Main-table metrics use the common official-17 output protocol described in Section 4.1.

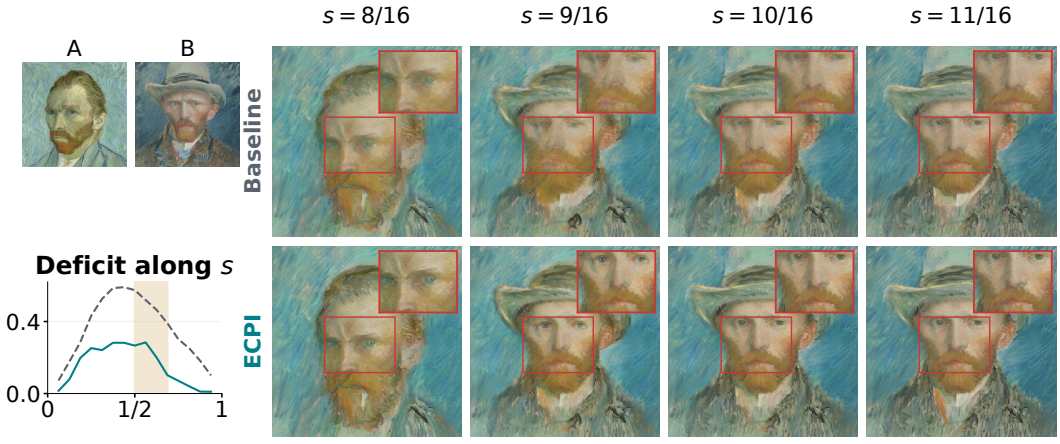

Figure 10: **Spectral deficits and their correction on the Van Gogh self-portrait pair.** Baseline and ECPI at $s = 8/16, \ldots, 11/16$ (MorphBench, SLERP), in the same format as Figure 2: small images show reconstructed endpoints, red boxes mark matched regions enlarged in the insets, and curves show mean positive four-band deficits (gray: baseline; teal: ECPI) with the displayed interval shaded. This is the pair used for the method overview in Figure 1.

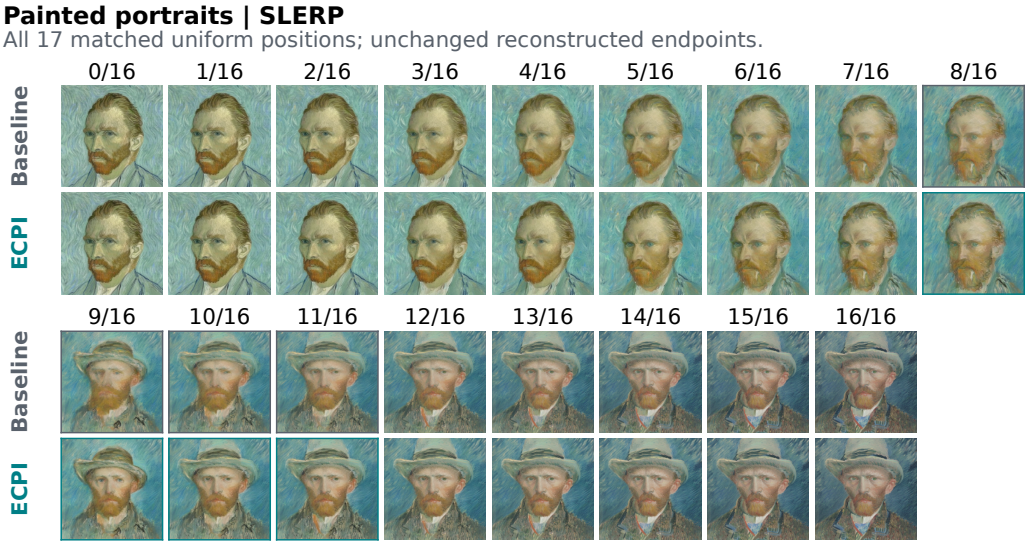

Figure 11: Complete SLERP and ECPI paths for the Van Gogh self-portrait pair (MorphBench). Borders mark the four consecutive positions shown in Figure 10.

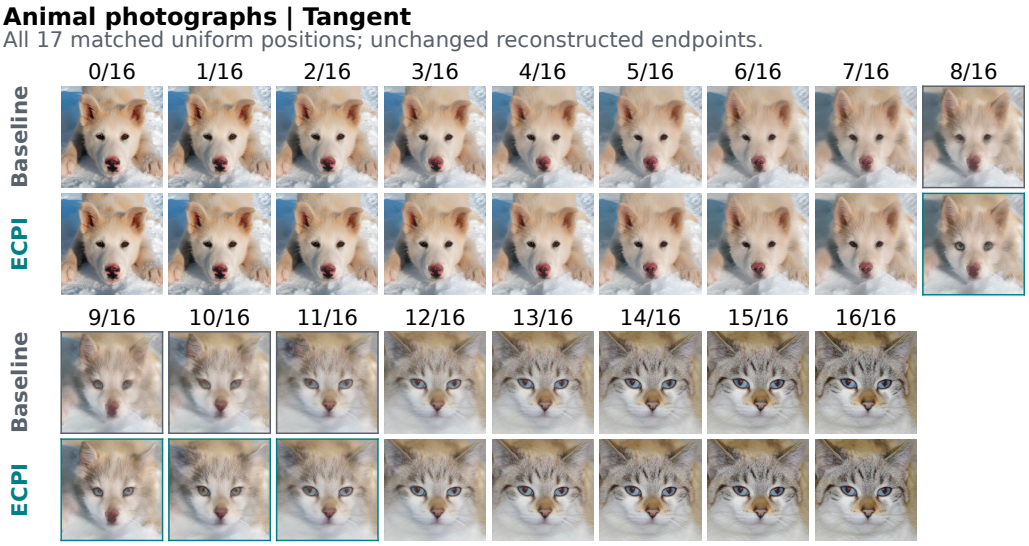

Figure 12: Complete Tangent and ECPI paths for the animal example (AH100). Borders mark the four consecutive positions shown in Figure 2.

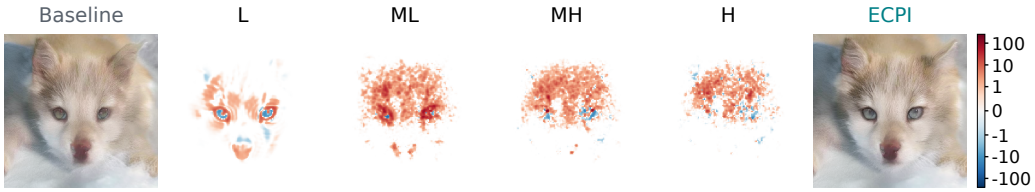

Figure 13: **Spatial band-energy changes on the animal example.** Animal photographs (AH100, Tangent) at $s = 9/16$: baseline (left), L/ML/MH/H spatial responses (middle), and ECPI (right). Each map shows ECPI-minus-baseline local band energy, normalized by that band's global mean baseline energy. Red indicates increases and blue decreases, with a shared color scale across bands. In this example, low-band changes are spatially broad around the eyes and nose, while higher-band changes reveal finer patterns across the face and forehead. These changes accompany clearer eye definition and fur texture in the ECPI image.

F.2 DECODER-REFINEMENT ARTIFACTS

Figure 14 illustrates artifacts from decoder-only refinement: over-bright eye highlights in the cinematic portrait (top) and darkened, distorted eye contours in the photographic portrait (bottom). These examples complement the quantitative comparison in Appendix E.2.

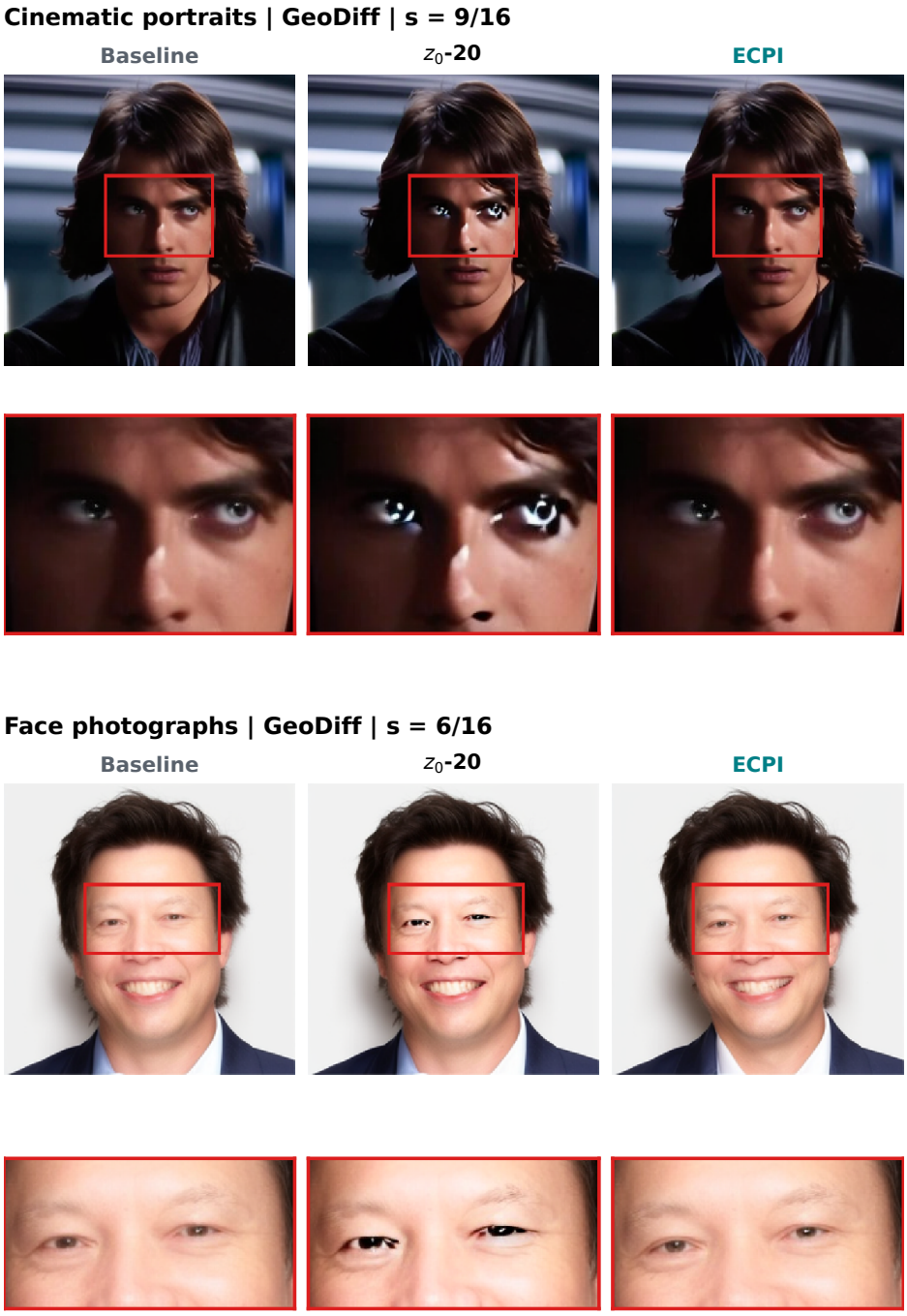

Figure 14: **Artifacts from decoder-only refinement.** Baseline, decoder-only z_0 refinement and ECPI on GeoDiff: cinematic portraits at $s = 9/16$ (top) and face photographs at $s = 6/16$ (bottom). Red boxes mark identical crop regions, enlarged below.