CONFIDENCE IS NOT EVIDENCE: PROBING CLASSIFIER SENSITIVITY WITH STEERABLE LENS

Anonymous authors
Paper under double-blind review

ABSTRACT

Moving an image along a classifier's gradient is a common way to probe what change would carry it across a decision boundary, but crossing that boundary does not by itself establish a recognizable change toward the target class. We introduce *Steerable Lens*, a framework for investigating classifier sensitivity through complex steerable-pyramid (CSP) interventions, with pixel and Fourier-phase edits as comparators. At matched image-change budgets, we measure three quantities separately: the optimized classifier's confidence, transfer to independent classifiers, and compatibility with a smooth image transformation. These measurements yield different method rankings. On MNIST at $L_2 = 2$, pixel edits reach a median target probability of 96.2%, yet joint independent-classifier agreement is infrequent and smooth reconstruction explains less edit energy than for CSP. At this budget, CSP edits have greater warp-plus-gain compatibility while median driver confidence remains low. On CelebA the structural ordering between representations depends on the reconstruction model. In a pilot human study, CSP edits receive more ratings of clear attribute change and no visible corruption than pixel or Fourier-phase edits. Frozen band probes further reveal fine-scale sensitivity in the examined driver classifiers, while larger budgets reduce common attainment across bands and controls. These results position CSP as a controlled diagnostic representation and motivate evaluating optimized confidence, classifier transfer, and smooth reconstruction separately, without treating any one as certification of recognizable target-class change.

1 INTRODUCTION

A classifier gradient specifies the direction in image space along which an infinitesimal change most increases a class score. It does not say what that change is in image terms, whether the model prefers a displaced contour, a strengthened texture, or a shift in intensity. Saliency maps describe local input sensitivity (Simonyan et al., 2013), while feature visualization synthesizes inputs that increase selected activations (Olah et al., 2017). We study source-conditioned edits in constrained multiscale coordinates and evaluate their trajectories using measurements excluded from optimization.

The complex steerable pyramid (CSP) is a useful coordinate system for such trajectories. Its scale- and orientation-selective complex coefficients factor local structure into magnitude and phase (Simoncelli & Freeman, 1995; Portilla & Simoncelli, 2000), and it is in phase that local spatial displacement is encoded. Phase-based video processing exploits exactly this, amplifying measured temporal phase variations to reveal otherwise invisible motion (Wadhwa et al., 2013). For a classifier, however, the phase change is *selected* by a decision function rather than *observed* in a physical motion: the analogy motivates a parameterization, but it does not guarantee an interpretable trajectory.

Interpreting such a trajectory requires separating two things the naive construction conflates. The first is when the gradient is evaluated. A gradient taken once at the source measures the model's initial preference; recomputing it at each step follows an optimization path, and an edit that succeeds under repeated optimization need not amplify that initial preference. The second is who evaluates the result. The classifier being optimized is not an independent judge of its own edit; agreement from a second classifier reports shared sensitivity, not human-perceived attribute change. This gap

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.

matters because visually small perturbations can change recognition (Ilyas et al., 2019) and robust classifiers can themselves drive image synthesis (Santurkar et al., 2019).

We develop *Steerable Lens*, a protocol for studying these distinctions through controlled image edits. Its aim is not a new pyramid transform or a generally superior counterfactual generator, but an operational question: which properties of a gradient-driven edit follow from its coordinates, from how the gradient is optimized, and from how the result is judged. At matched image-change budgets, the protocol reports optimized-classifier response, transfer to independent classifiers, and compatibility with a restricted smooth warp with intensity gain. Separate frozen probes measure finite per-band phase sensitivity. Human ratings provide an additional assessment of the CelebA edits.

Across our measurements the methods do not rank consistently, and that inconsistency is the central observation. On MNIST, CSP edits are more compatible with the tested smooth-reconstruction family than pixel or Fourier-phase edits, even though independent judges confirm the target class for few edits of any method. On CelebA the structural ordering reverses: Fourier phase is the most compatible under warp-plus-gain reconstruction, yet a limited human study rates the CSP edits as clearer and less corrupted, so the automated and human orderings need not coincide. Band probes at small budgets show strong variation across scale, while larger budgets remove the common support needed to compare bands. These results argue for a diagnostic use of CSP: constrain a family of changes explicitly, measure the response, and test the interpretation with evidence beyond the optimized score.

Our contributions are: (i) a separation of three intervention regimes (frozen sensitivity, persistent-coordinate optimization, and step-local steering) with explicit magnitude and residual handling; (ii) a matched-distance comparison of pixel, Fourier, and CSP edits that reports both a positive structural result and its reversal across datasets; (iii) a signed, random-controlled band probe that reports coverage instead of scoring unreachable edits as zero response; and (iv) a human pilot indicating that CSP edits are more often judged clear attribute changes, and less often corrupted, than pixel or Fourier-phase edits, while classifier transfer and smooth-reconstruction compatibility need not track those judgments. We treat the qualitative natural-image trajectories as illustrations rather than a population-level evaluation.

2 Related work

Saliency, attribution, and their audits. Attribution methods characterize a model's response by assigning importance to input features. Examples include vanilla and guided gradients (Simonyan et al., 2013; Springenberg et al., 2015), SmoothGrad (Smilkov et al., 2017), Integrated Gradients (Sundararajan et al., 2017), and Grad-CAM (Selvaraju et al., 2017). Feature visualization instead synthesizes inputs that increase a unit's activation (Olah et al., 2017). Audits have shown that some attribution methods can be insensitive to model parameters or training labels (Adebayo et al., 2018), and that explanations can change substantially under small input perturbations (Ghorbani et al., 2019). We extend this emphasis on evaluation from attribution maps to gradient-driven image edits. Alongside the optimized classifier's response, we measure transfer to classifiers excluded from the optimization and compatibility with a specified smooth transformation family.

Adversarial examples and counterfactuals. Adversarial examples demonstrate that visually small perturbations can change a classifier's prediction (Szegedy et al., 2014; Goodfellow et al., 2015). Such perturbations can exploit predictive features that are not robust to small input changes (Ilyas et al., 2019). Counterfactual explanations seek changes that produce an alternative prediction, with different formulations imposing different notions of proximity and plausibility (Wachter et al., 2018; Goyal et al., 2019). Robust classifiers can also support image synthesis and manipulation (Santurkar et al., 2019). Steerable Lens studies target-directed edits without assuming that they successfully cross a decision boundary or constitute plausible counterfactuals. Its contribution is a controlled comparison of intervention families that reports optimized confidence, independent-classifier transfer, and smooth-reconstruction compatibility separately. These measurements characterize different properties of an edit; none alone establishes recognizable target-class change.

Multiscale representations and phase. Our intervention coordinates draw on classical signal processing. Spatial translation corresponds to a phase shift in the Fourier domain (Oppenheim et al.,

1999). Steerable pyramids and wavelet representations organize image information across scales and, in suitable constructions, orientations (Simoncelli & Freeman, 1995; Mallat, 1999). Complex steerable-pyramid coefficients provide local magnitude and phase coordinates (Portilla & Simoncelli, 2000). Phase-based video magnification amplifies measured temporal phase variations to reveal small motions (Wadhwa et al., 2013). We instead obtain phase directions from a classifier objective. This analogy motivates the representation, but does not make arbitrary classifier-driven phase changes a coherent displacement field. We therefore treat smooth reconstruction as an empirical diagnostic of the resulting edits, rather than a property guaranteed by their coordinates.

3 Intervention coordinates and optimization

We edit an image through a fixed, redundant multiscale representation with analysis transform $\mathcal{F}$ and synthesis operator $\mathcal{F}^{-1}$; the synthesis reconstructs an image from any edited coefficients. We denote the source coefficient in band b (scale and orientation) as $\mathbf{c}_b^0 = \mathbf{A}_b^0 \exp(j\mathbf{\Phi}_b^0)$, and the low- and high-pass residuals are separate real tensors $\mathbf{r}_L^0, \mathbf{r}_H^0$. A phase increment $\mathbf{u}$ and a log-magnitude increment $\mathbf{v}$ define the rendered image

$$\mathbf{s}(\mathbf{u}, \mathbf{v}) = \mathrm{clip}_{[0,1]}\left[\mathcal{F}^{-1}\left(\left\{\mathbf{A}_b^0 \exp(\mathbf{v}_b) \exp(j(\mathbf{\Phi}_b^0 + \mathbf{u}_b))\right\}_b, \mathbf{r}_L^0, \mathbf{r}_H^0\right)\right]. \tag{1}$$

A phase-only intervention sets $\mathbf{v} = \mathbf{0}$ and a joint intervention updates both coordinates; the source decomposition stays the reference, and both residuals stay fixed, while clipping to $[0, 1]$ is part of the evaluated operator rather than a display step. As in phase-based motion magnification (Wadhwa et al., 2013) we use local phase as an intervention coordinate, but derive increments from a classifier objective rather than observed motion. Our experiments update the phase and log-magnitude coordinates directly, which differs from the alternative Wirtinger-derived complex-coefficient update with phase extraction (Appendix B).

We compare CSP interventions with pixel-space and Fourier edits: the Fourier comparator preserves Fourier magnitudes and keeps phase increments antisymmetric across conjugate frequency pairs, preserving conjugate symmetry of the Fourier coefficients and giving a real image before clipping. It is deliberately phase-restricted rather than an unconstrained Fourier reparameterization of pixel optimization. The regimes below differ in when gradients are evaluated and whether coordinates stay anchored to the source or are recomputed.

Frozen sensitivity. For a target score q we compute the phase-only source gradient $\mathbf{g}_0 = \nabla_{\mathbf{u}} q(s(\mathbf{u}, \mathbf{0}))|_{\mathbf{u}=0}$ once and render $s(\alpha\, \mathbf{g}_0, \mathbf{0})$ subject to the coordinate bounds. A single-band probe restricts the update to one band; joint frozen interventions use the source gradients in both $\mathbf{u}$ and $\mathbf{v}$ at the specified relative scaling. This regime extrapolates a fixed direction and does not follow the changing gradient of the edited image, so increasing α need not keep raising the target score: the linear approximation can fail, bounds can activate, and clipping can alter the trajectory.

Persistent-coordinate optimization. This regime recomputes gradients at the current edited image while retaining the source decomposition in Eq. 1. Starting from $\mathbf{u}_0 = \mathbf{v}_0 = \mathbf{0}$, a direct-coordinate ascent update has the form

$$\begin{pmatrix}\mathbf{u}_{k+1} \\ \mathbf{v}_{k+1}\end{pmatrix} = \Pi_{\mathcal{B}}\left[\begin{pmatrix}\mathbf{u}_k \\ \mathbf{v}_k\end{pmatrix} + \eta_k \mathbf{D} \nabla_{(\mathbf{u},\mathbf{v})} q\big(\mathbf{s}(\mathbf{u}_k, \mathbf{v}_k)\big)\right], \tag{2}$$

where $\mathbf{D}$ specifies the phase and log-magnitude block weights, $\Pi_{\mathcal{B}}$ enforces the coordinate bounds, and η_k is the accepted step size; phase-only optimization fixes $\mathbf{v}_k = \mathbf{0}$. Although the gradients change each iteration, the coefficients are always constructed relative to $\mathbf{A}_b^0$ and $\mathbf{\Phi}_b^0$, so the current rendered image is not substituted as a new anchor. On MNIST we minimize target cross-entropy with adaptive per-step image-change control and loss-based backtracking; on CelebA we maximize a target-minus-source margin using development-calibrated optimizer settings. With different objectives, block weights, and search limits, the MNIST and CelebA runs are distinct optimizers, and we report their settings separately.

Step-local steering. For the natural-image illustrations, we also reanalyze the current image at every iteration, $(\{\mathbf{c}_b^k\}, \mathbf{r}_L^k, \mathbf{r}_H^k) = \mathcal{F}(\mathbf{s}_k)$, and take the next phase update from this fresh decomposition. Magnitudes and residuals are held fixed within a step but drift over the trajectory as synthesis,

clipping, and reanalysis change the recovered coefficients; this distinguishes step-local steering from persistent-coordinate optimization, though both recompute gradients. These trajectories illustrate repeated steering and are not matched-budget evidence of superiority. In every regime the driver's confidence is a response the intervention pursues, not independent validation of the change.

Sign, scale, and numerical checks. A target-directed update ascends q or descends target cross-entropy; we verify update signs by finite differences or small rendered-loss changes, confirm reconstruction of the unedited source, and check that coefficient blocks meant to stay fixed do, distinguishing the coefficients supplied to synthesis from those recovered by reanalyzing the clipped image. Because an amplification constant α or step η_k is not comparable across coordinate systems, we measure image change after reconstruction and clipping at the evaluation resolution, $d(\mathbf{u}, \mathbf{v}) = \|\mathbf{s}(\mathbf{u}, \mathbf{v}) - \mathbf{s}(\mathbf{0}, \mathbf{0})\|_2$, and match rendered-image budgets. A trajectory that does not attain its requested budget is recorded as non-attainment rather than treated as budget-matched, and we report attainment with each comparison, restricting analyses to sources every method attains. Independent classifiers and smooth-reconstruction scores are excluded from edit optimization.

4 Evaluation protocol

For quantitative comparisons we evaluate edits at matched rendered-image budgets along three axes: the optimized classifier's response, transfer to classifiers held out of edit optimization, and compatibility with a restricted family of smooth image transformations. The first is the quantity being optimized; the other two are excluded from optimization. These axes need not agree, and none alone establishes a recognizable target-class change.

Response and transfer. For the optimized classifier $\mathcal{G}$, we report the target probability and the target-minus-source logit margin. Two additional classifiers, O and O_2, are excluded from edit-gradient computation and optimization-based step selection. Their target predictions measure classifier transfer, not ground truth. On CelebA, we additionally evaluate a frozen CLIP encoder with a trained linear head for each attribute (Radford et al., 2021); this is a supervised attribute probe, not zero-shot text matching.

We distinguish an endpoint target prediction (the edited image receives the target label) from a source-to-target switch (which also requires the source to be labeled correctly first). Two-judge agreement requires both O and O_2 to predict the target; its failure means at least one does not, not that both reject. Each reported transfer statistic states which it is.

Matched image distance. Since an amplification constant is not comparable across representations (Sec. 3), we compare endpoints with $\|\mathbf{s}' - \mathbf{s}\|_2 \approx \epsilon$ within tolerance, searching each method's scalar amplification or step length and reconstructing through its editing operator rather than interpolating displayed pixels; distances are measured after reconstruction and clipping at the evaluation resolution. Sources are selected without regard to edit success, and failure to attain a budget is recorded. Comparisons use the sources attained by both methods (pairwise) or all methods (jointly), and we report the denominators, since the estimates are conditional on common attainment. Separately selected qualitative examples are labeled as illustrations, not representative samples.

Smooth-reconstruction diagnostic. To measure compatibility with a smooth image transformation, we fit the source $\mathbf{s}$ to the edited image $\mathbf{s}'$ using a smooth inverse warp W and, optionally, a spatially varying gain $\mathbf{g}$. We report

$$\widehat{\mathbf{s}} = \mathrm{clip}_{[0,1]}\big(\mathbf{g} \odot W(\mathbf{s})\big), \qquad E = 1 - \frac{\|\mathbf{s}' - \widehat{\mathbf{s}}\|_2^2}{\|\mathbf{s}' - \mathbf{s}\|_2^2}. \tag{3}$$

For a nonzero edit, E measures the reduction in squared reconstruction error relative to leaving the source unchanged. Warp-only fits set $\mathbf{g} = 1$. The fit uses no classifier, and its optimization budget and initialization procedure are held fixed across editing methods. We compare 4×4 and 8×8 control grids, with bounded displacement and, when enabled, gain in $[0.5, 1.5]$.

This diagnostic measures approximation by the specified family, not physical motion or semantic change; spatial gain in particular can explain intensity changes without displacement. Comparing warp-only and warp-plus-gain fits tests dependence on the family but does not uniquely split an edit into motion and intensity, since the warp is reoptimized when gain is added. A low E can reflect

the family's limits or the nonconvex fit; negative E is worse than the unchanged source, and E is undefined for a zero-energy edit.

Signed band controls. The band probe measures classifier sensitivity to individual scales and orientations. For each development face, we compute the source-margin gradient once and intervene on one phase band at a time. We evaluate both gradient signs and five independent Gaussian phase fields per band as controls. Each random field is reused with both signs, and each intervention is separately scaled to the requested rendered-image distance.

For band b and sign $\sigma \in \{-1, +1\}$ the signed response is $R_{b,\sigma} = \sigma[q(\mathbf{s}'_{b,\sigma}) - q(\mathbf{s})]$. We subtract the mean signed response of the matching random controls within each face, average across signs, then aggregate across faces, using the common attainment set and reporting how its size changes with budget. The controls match band dimensionality and rendered-image distance but not the gradient's spatial correlation, and the driver's alignment with its own gradient already favors a positive infinitesimal response. The probe therefore measures the size of finite responses, their variation across scale and orientation, their transfer, and the attainable budget range; it does not establish that a sensitive band is a semantic feature.

5 EXPERIMENTS

Figure 1 shows selected matched-budget edits that preview the separations the following experiments quantify.

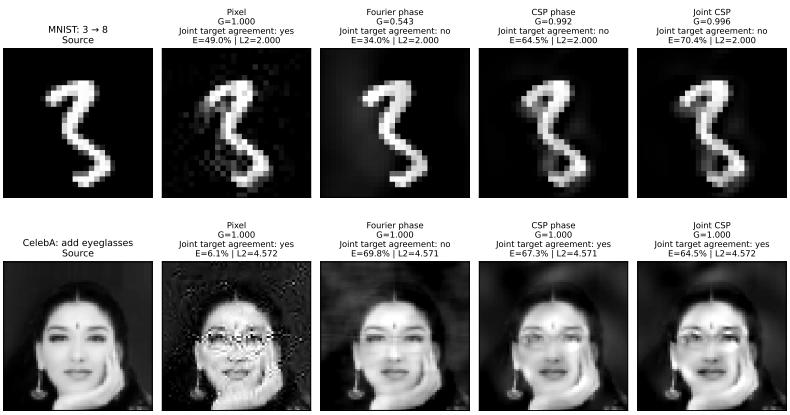


Figure 1: Confidence, transfer, and smooth compatibility assess different properties of an edit. Columns show the source and actual scored persistent-coordinate endpoints for pixel, Fourier-phase, CSP-phase, and joint-CSP editing. Within each row, endpoints have matched image distance: $L_2 = 2$ for MNIST and $128/28$ for CelebA, corresponding to equal per-pixel RMS across resolutions. G is driver target probability; joint target agreement requires both independent classifiers to predict the target ("no" means at least one does not). E is explained edit energy under the 8×8 warp-plus-gain fit. On the MNIST source, pixel and both CSP variants attain target probability above 0.99, but only pixel achieves joint target agreement, whereas CSP has higher E. On the CelebA source, all methods attain near-unit driver probability, yet transfer and reconstruction scores differ. Cases were deliberately selected after inspecting appearance and outcomes; they illustrate measurements rather than aggregate success rates or validated human recognition.

5.1 MNIST: SMOOTH COMPATIBILITY AND CLASSIFIER RESPONSE SEPARATE

We first compare pixel, Fourier-phase, CSP-phase, and joint-CSP edits on MNIST (LeCun et al., 1998). Three transitions, $1 \to 7$, $3 \to 8$, and $5 \to 6$, contribute 100 official-test sources each, and both the sampling and the optimization settings were fixed before the run. Because these test digits had been used in earlier experiments on this project, we read the comparison as a fixed-protocol replication rather than held-out confirmation. Disabling confidence-based early stopping lets every

method reach both image budgets ($L_2 = 1$ and 2) on all 300 sources, so the four methods are compared on an identical set of sources with no exclusions for non-attainment.

On the structural axis the CSP edits stand apart. At the larger budget the 8×8 warp-plus-gain model reproduces 71.9% of CSP-phase and 72.2% of joint-CSP edit energy, against 45.7% for pixel and 41.0% for Fourier-phase edits (Fig. 2). The median within-source advantage over pixel is 24.4 percentage points for CSP phase (95% source-bootstrap interval: 23.2–25.8) and 24.7 points for joint CSP (23.8–25.6). These are medians of paired differences rather than differences between marginal medians. Both CSP variants remain more reconstructible than both comparators at $L_2 = 1$ and under the coarser reconstruction grid.

Greater smooth-reconstruction compatibility does not translate into greater classifier transfer. Median driver target probability is 96.2% for pixel, 0.35% for CSP phase, and 0.95% for joint CSP. Both independent judges predict the target on 10, 3, 6, and 8 of 300 edits for pixel, Fourier phase, CSP phase, and joint CSP, respectively. Pixel edits produce the largest median driver response, whereas CSP edits are more compatible with the tested smooth transformation family. Two-judge agreement is infrequent for every method. These results separate driver response, transfer, and smooth compatibility without establishing recognizable target-class change.

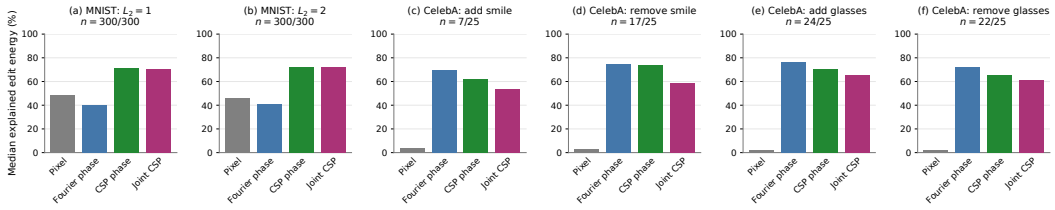


Figure 2: Smooth reconstruction at matched image distance under the 8×8 warp-plus-gain fit. Bars show median explained edit energy. (a,b) MNIST at $L_2 = 1$ and 2; all methods attain each budget on the same 300 sources. (c–f) CelebA at RMS change $2/28$ ($L_2 = 128/28$); comparisons use sources on which all four methods attain the budget. Each panel reports common attainment out of the number requested. Fourier phase exceeds CSP under this fitting family on CelebA; smile addition is especially conditional on limited attainment. Scores measure compatibility with the fitted family, not human-perceived quality or target-class validity.

5.2 CelebA: The Reconstruction Ordering is not Universal

We repeat the structural comparison on 100 grayscale 64×64 faces, 25 per direction for Smiling and Eyeglasses, reusing the pipeline's binary classifiers, preprocessing, and attribute labels (Liu et al., 2015). Image L_2 budgets of $64/28$ and $128/28$ match the MNIST per-pixel RMS changes of $1/28$ and $2/28$, and the reconstruction settings are adjusted for the higher resolution. Step sizes and the global CSP phase and magnitude weights are calibrated on held-out development faces, again without reference to judge outcomes.

Unlike MNIST, not every face reaches the budget: of 800 requested endpoints, 681 are attained, and attainment varies sharply by direction. At the larger budget all four methods attain the budget on most sources for eyeglasses (24/25 addition, 22/25 removal) and for smile removal (17/25), but on only 7 of 25 for smile addition (Fig. 2). The comparisons that follow are therefore conditional on the sources where all four methods attain the budget, and smile addition, with the thinnest coverage, is the most tentative.

Under the same 8×8 warp-plus-gain diagnostic the ordering reverses. Fourier-phase edits are now the most reconstructible in every attribute–direction group and at both budgets, with median explained energy in the low-to-mid seventies (69.9–76.3% at the larger budget), just ahead of CSP phase (62.1–73.6%) and far ahead of pixel edits (2.0–3.5%).

The ranking even depends on the fitting family. At the larger budget a warp-only fit favors CSP phase over Fourier phase in all four directions, and adding spatial gain flips that comparison. Since the warp is reoptimized once gain is allowed, the reversal cannot be read as separating motion from

intensity; it simply shows that smooth compatibility is a property of an edit relative to a chosen family, not a fixed ranking of the representations.

Classifier transfer gives yet another ordering. For smile removal, both independent judges move from the source label to the target on 16 of 17 common sources for each CSP variant, against 6 and 7 of 17 for pixel and Fourier phase. So within a single group Fourier phase is the most reconstructible while CSP produces the most two-judge switches. A CLIP-linear probe responds even to pixel edits, a reminder that adding a classifier supplies one more model response, not a verdict on perceptual quality. None of these signals, alone or in agreement, shows that the intended attribute became visible or that the rest of the face was preserved; we take up those questions with human raters (Sec. 5.2), whose sampling and denominators differ from this transfer comparison.

An informal human pilot on CelebA edits

Classifier agreement does not directly measure human perception. We collected voluntary ratings from friends and family on 90 source–edit pairs spanning 24 source–attribute cases, retaining two complete response sets (Appendix D gives the export provenance). The pilot covers smile and eyeglasses addition and removal, with 23, 23, 22, and 22 edited images for pixel, Fourier phase, CSP phase, and joint CSP. Images came from cases with common budget attainment at $L_2 = 128/28$, and method identities and classifier scores were hidden in the rating interface.

Raters judged target-attribute change, preservation of unrelated content, and visible corruption as separate questions rather than a single quality score: target change as clear, partial, unchanged, opposite, or unclear; preservation as intact, minor, or major unrelated change; and corruption as none, mild, or severe. Keeping these apart lets an edit be recognizable and yet corrupted, or clean and yet uninformative.

The two rating sets give 180 ratings across the 90 items (Table 1). We report descriptive counts rather than majority-vote labels, since the repeated ratings of an image are repeated observations, not independent samples. The two response sets assigned the same category to 57 of 90 items for target change, 62 for preservation, and 66 for corruption.

Recognizability and corruption. The comparators drew few clear target changes, 2 of 46 pixel ratings and 9 of 46 for Fourier phase, whereas each CSP variant drew many more (20 of 44 for CSP phase, 22 of 44 for joint CSP); counting partial changes as well, the CSP variants again lead, 38/44 and 41/44 against 28/46 for each comparator. Corruption separates the methods in the same direction: no pixel edit was rated free of corruption, against 15 of 46 Fourier-phase, 30 of 44 CSP-phase, and 35 of 44 joint-CSP ratings. Low corruption did not guarantee preservation, though, as major unrelated changes still appeared in a few CSP ratings (5 of 44 CSP phase, 7 of 44 joint CSP). Intended change, unrelated change, and corruption are thus best scored separately.

Table 1: Human judgments from two complete rating sets on 90 edited images. N denotes images and R denotes ratings. Clear and partial refer to target-attribute change; major unrelated change and no corruption are responses to separate questions. Counts are descriptive and do not treat repeated ratings as independent images.

Method	N	R	Clear	Partial	Major unrelated	No corruption
Pixel	23	46	2	26	45	0
Fourier phase	23	46	9	19	3	15
CSP phase	22	44	20	18	5	30
Joint CSP	22	44	22	19	7	35

Interpretation and scope. Within the pilot, then, CSP edits collected more clear-change and no-corruption responses than either comparator, adding perceptual information that classifier predictions and reconstruction scores do not capture. This does not show that every automated measurement fails to predict human judgment; that stronger claim would require matching automated scores and human labels item by item. The 16-of-17 smile-removal transfer result above, for instance, rests on a different denominator and cannot be lined up against these counts.

The study reaches only so far: two rating sets, two attributes, and a small set of faces conditional on common attainment, with shared source cases and non-independent ratings. We read the results as

descriptive evidence about these particular edits rather than a population-level ranking of methods, and they do not extend to the MNIST transitions or the natural-image trajectories.

5.3 Frozen Band Probes Localize Small-budget Sensitivity

The frozen band probe runs on 80 development face–attribute cases, 20 per direction. Each case is edited along all 16 pyramid bands, in both signs, at five budgets ($L_2 \in \{0.01, 0.025, 0.05, 0.1, 0.25\}$), and the source-gradient direction is set against five Gaussian random controls per band, with every scalar search capped at a phase increment of $\pi/2$. Counting the source direction together with its five controls, this requests $80 \times 16 \times 2 \times 6 \times 5 = 76{,}800$ endpoints; non-attainment is retained, and the independent classifiers and the CLIP-linear probe are read only on attained endpoints. For band b and sign σ, we report the signed response $R_{b,\sigma} = \sigma[q(\mathbf{s}'_{b,\sigma}) - q(\mathbf{s})]$ minus the within-face mean of its random controls, averaged over signs before aggregating across faces.

A clean cross-band comparison needs every band, sign, and control to reach the budget on the same face, and that common-attainment set thins rapidly as the budget grows. All 20 cases per group qualify at the two smallest budgets and nearly all at 0.05, but by $L_2 = 0.1$ coverage falls unevenly, as low as 3 of 20 for eyeglasses addition, and at 0.25 no group is fully comparable. Larger requested edits thus leave less to compare, not more.

At $L_2 = 0.025$, the budget where every case stays comparable, the signed driver response above random controls falls from the finest to the coarsest scale in all four groups (Fig. 3), from finest-scale advantages of up to roughly 1.8 logit units down to a few hundredths at the coarsest scale. Fine-scale phase is where these drivers are most sensitive, but a larger response marks neither an anatomical feature nor a more meaningful change, and the independent-model logits are read on their own scale rather than compared across classifiers.

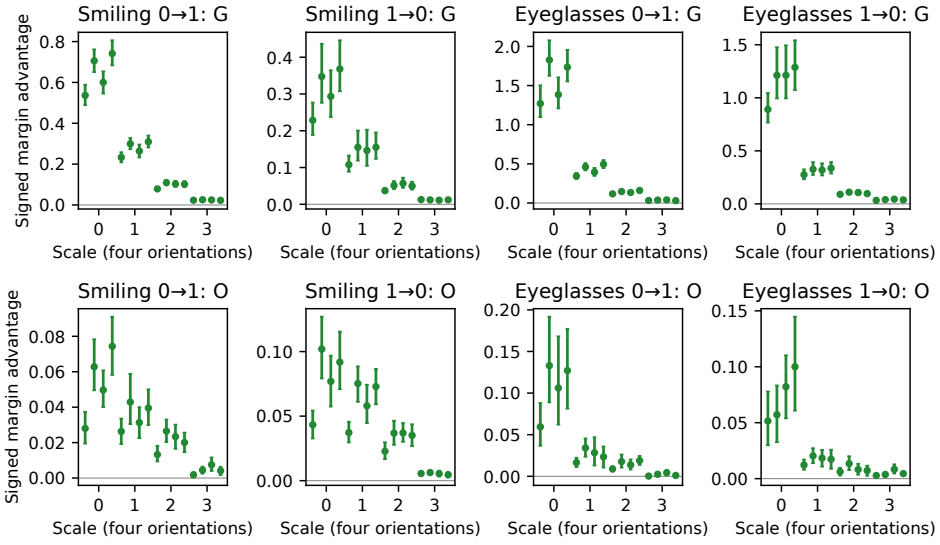


Figure 3: Frozen phase-band probes on development faces at image $L_2 = 0.025$. Each group contains the same 20 cases across all bands and controls. Points show mean signed margin change above the within-image random-control mean. Bars are pointwise 95% image-bootstrap intervals from 2,000 draws, with signs and random controls averaged within each image. Top: driver G; bottom: independent classifier O. Four orientations are shown per scale; scale 0 is finest. Intervals are exploratory and are neither multiplicity-adjusted nor estimates of variation across independently trained classifiers.

5.4 NATURAL-IMAGE ILLUSTRATIONS

We illustrate frozen single-band sensitivity probes and all-band step-local steering on three local cat photographs with ImageNet-pretrained classifiers (Deng et al., 2009; He et al., 2016). The probes use a gradient evaluated at the source, whereas steering recomputes gradients after reanalyzing each edited image. Appendix F gives the setup and all six selected reference-backend trajectories. Additional cat–dog close-ups and within-cat trajectories from a local-pyramid implementation appear separately in Appendices G and H. These qualitative illustrations are not an ImageNet benchmark or a matched-budget comparison of the regimes; their classifier responses alone do not establish recognizable target-class change.

6 DISCUSSION AND LIMITATIONS

Across our measurements no single ordering is shared by response, transfer, and smooth compatibility. On MNIST, CSP edits are the most compatible with the smooth warp-plus-gain family at matched distance, yet pixel edits produce far higher driver confidence and two-judge agreement is uncommon for every method; on CelebA the reconstruction ranking depends on the fitting family while some CSP interventions transfer frequently. The human pilot adds a qualified positive result, with raters reporting clear attribute changes and no corruption more often for the CSP variants, but it does not establish general superiority, guarantee preservation of unrelated content, or show that every automated measurement fails to predict human judgments. Frozen band probes localize stronger fine-scale sensitivity at small budgets, with rankings stable to leaving out images, while common attainment declines at larger budgets; they measure sensitivity in specified coordinates, not a semantic feature.

Scope. The experiments cover three MNIST transitions, two CelebA attributes, and a small natural-image illustration; the natural-image trajectories are a demonstration, not a semantic-editing benchmark. Historical reuse of the evaluated images and test-based selection of the legacy CelebA judge limit claims of fresh held-out confirmation.

Measurement and statistical limitations. Smooth reconstruction uses a restricted family and nonconvex fitting, so its score depends on both the family and the fit, and spatial gain can explain intensity without motion. Bootstrap intervals quantify resampling within the evaluated cases, not reproducibility across trained classifiers, and attainment-restricted comparisons are conditional on that selection. The band analysis has 80 images from 79 identities; image-bootstrap intervals are supplemented by identity-cluster resampling and leave-one-image-out checks, which address within-collection dependence but not variation across training runs, and Gaussian controls match band dimensionality and distance, not the gradient's spatial correlation.

Limits of validation. Independent classifiers supply transfer, not ground truth about realism, attribute change, or identity preservation. The human pilot gives direct judgments but from only two complete rating sets on 90 edits across 24 source-attribute cases, with repeated ratings and shared sources that are not independent, and it does not validate the MNIST or natural-image results. Finally, the three regimes answer different questions; where their objectives, models, budgets, or constraints differ, their outcomes cannot be pooled into a controlled comparison of gradient refresh alone.

7 CONCLUSION

Steerable Lens makes classifier sensitivity measurable at specified scales, orientations, and image-change budgets, and shows that no single axis certifies an edit: on MNIST, CSP edits are the most warp-plus-gain compatible despite low driver confidence and infrequent two-judge agreement; on CelebA the structural ordering depends on the fitting family while a small human pilot favors CSP; and frozen band probes reveal fine-scale sensitivity at small budgets, while common attainment declines at larger budgets. These findings support CSP as a controlled diagnostic representation rather than a universally superior editor. Optimized confidence, classifier transfer, smooth-reconstruction compatibility, and human judgment answer different questions, and reporting them separately makes the interpretation of an edit explicit: a classifier response is evidence about that classifier, while a claim of recognizable change requires validation beyond the score being optimized.

REPRODUCIBILITY STATEMENT

Appendix A documents models, objectives, coordinate constraints, and fitting settings; the figure manifest records figure inputs and checksums. Aggregate figures are generated from completed records. Qualitative examples follow the selection procedures stated in their captions and appendices, including outcome-based selection. Replication scripts preserve nonattainment and distinguish displayed-image measurements from coefficient constraints. Access to the original datasets and pretrained checkpoints is required.

ETHICS STATEMENT

Face edits may change appearance or support misleading interpretations of attributes. We use them to examine classifier responses without claims about personal identity or sensitive traits. Human ratings were provided voluntarily by a convenience sample of friends and family. The small sample and familiarity with the authors limit generalizability and may influence judgments. Classifier agreement is not treated as human judgment (Sec. 5.2).

AI USE STATEMENT

Generative AI tools assisted with manuscript editing, reference verification, and preparation of figures. The authors are responsible for verifying the text, code, numerical results, citations, and final claims, including material prepared with AI assistance.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL https://arxiv.org/abs/1810.03292.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009. URL https://www.image-net.org/about.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *AAAI Conference on Artificial Intelligence*, volume 33, pp. 3681–3688, 2019. doi: 10.1609/aaai.v33i01.33013681. URL https://ojs.aaai.org/index.php/AAAI/article/view/4252.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6572.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning (ICML)*, volume 97, pp. 2376–2384. PMLR, PMLR, 2019. URL https://proceedings.mlr.press/v97/goyal19a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html?_hsmi=2.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://gradientscience.org/adv/.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. URL https://bottou.org/papers/lecun-98h.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2015. URL https://arxiv.org/abs/1411.7766.

Stéphane Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 2 edition, 1999. URL https://www.di.ens.fr/~mallat/book.html.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017. doi: 10.23915/distill.00007. URL https://distill.pub/2017/feature-visualization/.

Alan V. Oppenheim, Ronald W. Schafer, and John R. Buck. *Discrete-Time Signal Processing*. Prentice Hall, 2 edition, 1999. URL https://ocw.mit.edu/courses/6-341-discrete-time-signal-processing-fall-2005/pages/readings/.

Javier Portilla and Eero P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–71, October 2000. doi: 10.1023/A:1026553619983. URL https://www.cns.nyu.edu/pub/lcv/portilla99-reprint.pdf.

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 *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Shibani Santurkar, Dimitris Tsipras, Brandon Tran, Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Image synthesis with a single (robust) classifier. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://arxiv.org/abs/1906.09453.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. URL https://arxiv.org/abs/1610.02391.

Eero P. Simoncelli and William T. Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *IEEE International Conference on Image Processing (ICIP)*, volume 3, pp. 444–447, Washington, DC, October 1995. IEEE. URL https://www.cns.nyu.edu/pub/eero/simoncelli95b.pdf.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. URL https://arxiv.org/abs/1312.6034.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. URL https://arxiv.org/abs/1706.03825.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations (Workshop)*, 2015. URL https://arxiv.org/abs/1412.6806.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, pp. 3319–3328. PMLR, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. URL https://arxiv.org/abs/1312.6199.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31 (2):841–887, 2018. URL https://jolt.law.harvard.edu/volumes/volume-31.

Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T. Freeman. Phase-based video motion processing. *ACM Transactions on Graphics*, 32(4):80:1–80:10, 2013. URL https://people.csail.mit.edu/billf/publications/Phase-Based_Video.pdf.

A Implementation and experimental provenance

The main quantitative experiments use direct gradients in source-anchored phase and log-magnitude coordinates, as in Eq. 1. Oriented coefficients are edited locally, without an additional Fourier transform of each subband. Both source residuals remain fixed. A joint intervention changes oriented-band magnitudes through the log-magnitude coordinates; it does not free the residuals. The alternative complex-gradient rule in Appendix B is a separate experiment.

The reference complex pyramid uses order 3, automatic height, transition width 1, downsampling, and non-tight-frame synthesis. The grayscale 64×64 band experiment contains 16 oriented bands. Clipping is included in the rendered operator. Constraints on coefficients supplied to synthesis do not imply identical constraints on coefficients recovered after clipping and reanalysis.

A.1 Models, objectives, and budgets

MNIST uses a convolutional driver G, a width-32 residual-network judge O, and a width-512 MLP judge O_2, trained with seeds 0, 1009, and 2018. Their validation accuracies are 98.74%, 99.20%, and 97.94%, respectively. These clean-image accuracies are not guarantees on edited images. The structural replication includes 100 sources for each of $1 \to 7$, $3 \to 8$, and $5 \to 6$. All official-test images had historical project exposure; the replication is not an untouched test evaluation.

The MNIST persistent optimizer minimizes target cross-entropy, uses at most 300 updates, and has calibrated local image-step settings 0.25 for pixel, 0.10 for Fourier phase, and 0.25 for both CSP variants. Budget matching uses tolerance 5×10^{-4}. Confidence-based early stopping is disabled. The reconstruction diagnostic uses 400 fitting steps from each of two starts, control grids 4×4 and 8×8, displacement bounded by three pixels per component, and optional gain in $[0.5, 1.5]$.

CelebA uses historical convolutional, residual-network, and MLP checkpoints on grayscale 64×64 images. The additional CLIP probe uses a frozen ViT-B/32 encoder and an attribute-specific linear head. CelebA calibration uses development faces and a binary target-minus-source margin objective; it is not an identical optimizer replication of MNIST. Image and displacement budgets scale by $64/28$. The legacy O checkpoint was selected using test performance, which limits evaluation independence even though current edit optimization excludes judge outcomes.

A.2 Why intervention coordinates matter

For a real linear synthesis map $\mathbf{s} = \mathbf{B}\mathbf{c}$, a Euclidean coefficient-gradient step induces the image direction $\mathbf{B}\mathbf{B}^{\mathsf{T}}\nabla_{\mathbf{s}} q$. An orthonormal reparameterization therefore reproduces the pixel direction. Phase and log-magnitude coordinates instead define a nonlinear, source-dependent map. With Jacobian $\mathbf{J}$ and coordinate weighting $\mathbf{D}$, the local image direction is $\mathbf{J}\mathbf{D}\mathbf{J}^{\mathsf{T}}\nabla_{\mathbf{s}} q$ before active constraints and clipping are accounted for. Coordinate restrictions and weighting can change the direction without establishing semantic validity.

B Alternative Wirtinger-derived update

This section describes an alternative to the direct-coordinate rule used in the headline experiments. For a real objective q to be maximized and a complex coefficient $c = Ae^{j\Phi}$ with $A > 0$, define $q_A = \partial q/\partial A$ and $q_\Phi = \partial q/\partial \Phi$. Wirtinger differentiation gives

$$\frac{\partial q}{\partial c^*} = \frac{1}{2} e^{j\Phi} \left(q_A + j \frac{q_\Phi}{A} \right), \qquad \widetilde{c} = c + \eta \frac{\partial q}{\partial c^*}. \tag{4}$$

Consequently,

$$\Delta\Phi = \arg \left[1 + \frac{\eta}{2} \left(\frac{q_A}{A} + j \frac{q_\Phi}{A^2} \right) \right]. \tag{5}$$

Joint editing uses $|\widetilde{c}|$ as the new magnitude; phase-only editing restores the prescribed magnitude while retaining the updated angle. To first order, the phase increment is $\eta q_\Phi/(2A^2)$, not a uniform multiple of the direct phase gradient across coefficients. Thus the two rules use different coordinate metrics. This construction follows complex differentiation; phase-based temporal motion magnification motivates local phase coordinates but does not supply this classifier-gradient rule.

In the reported comparison, q is negative target cross-entropy. The separate implementation regularizes denominators with 10^{-12}, retains source residuals and persistent coordinates, and applies total phase and log-magnitude bounds. Its phase-only angle also uses the magnitude derivative. Development calibration searches image-step settings 0.05, 0.10, and 0.25 using attained endpoints and driver response, without judges. The same 300 historically reused MNIST sources and reconstruction procedure are evaluated. Because calibration and preconditioning differ, this comparison is an alternative optimizer, not a change of notation alone.

Table 2: Separate MNIST update-rule comparison at $L_2 = 2$. G is median target probability; agreement requires both judges to predict the target; E is median 8×8 warp-plus-gain explained energy. All methods attain the budget on 300 sources.

Update	Intervention	G (%)	Agreement	E (%)
Direct	Pixel	96.24	10/300	45.70
Direct	Fourier phase	0.002	3/300	41.03
Direct	CSP phase	0.35	6/300	71.89
Direct	Joint CSP	0.95	8/300	72.25
Wirtinger-derived	CSP phase	26.80	4/300	73.24
Wirtinger-derived	Joint CSP	68.86	5/300	64.34

The alternative increases driver response for these CSP settings without increasing joint target agreement. The comparison does not establish semantic editing quality or generalize to CelebA without separate evaluation.

C Band coverage, identity dependence, and visual differences

Frozen probes use one source-margin gradient, 80 logarithmic search points, phase cap $\pi/2$, and image-distance tolerance 10^{-5}. Cross-band comparisons require attainment by all 16 bands, both signs, and the gradient and five random directions. The 80 images represent 79 identities; one identity occurs twice, within eyeglasses removal. Identity-cluster resampling supplements the main figure's 2,000-draw image bootstrap. It changes the resampling unit, not the original point estimates or historical selection.

At $L_2 = 0.025$, orientation-averaged scale rankings for G and O remain finest-to-coarsest under every leave-one-image-out analysis in all four groups. O_2 exhibits different scale rankings, also stable under these deletions. For the CLIP probe, eyeglasses-removal ordering changes in two of 20 deletions; the top scale changes in one. These checks concern the evaluated images, not independent training runs. A checkpoint audit found no compatible, independently documented second driver suitable for a clean seed replication.

D Human-rating protocol and export provenance

The blinded interface presents a source, an edit, and the requested attribute direction, without method identity or model scores. It separates visible change, target-attribute change, preservation of unrelated content, and corruption; free-text descriptions are optional. The 90 items span 24 source–attribute cases selected from common-attainment cases at $L_2 = 128/28$. Methods have 23, 23, 22, and 22 items, respectively. Repeated ratings and methods sharing sources are not independent image samples.

Five supplied export files contain two complete response sets, two earlier partial exports, and one exact duplicate of a complete export. The complete sets contain 90 items each; the partial files contain 42 and 86. Duplicate export files do not establish additional participants. The deduplicated descriptive analysis therefore uses 180 ratings. Exact category agreement is 57/90 for target change, 62/90 for preservation, and 66/90 for corruption.

Pooling all five files instead produces 398 entries with duplicate weighting (103 pixel, 106 Fourier phase, 93 CSP phase, and 96 joint CSP). Such pooling is a file-level calculation, not a five-independent-rater analysis, and is not used to estimate inter-rater reliability. Human ratings were

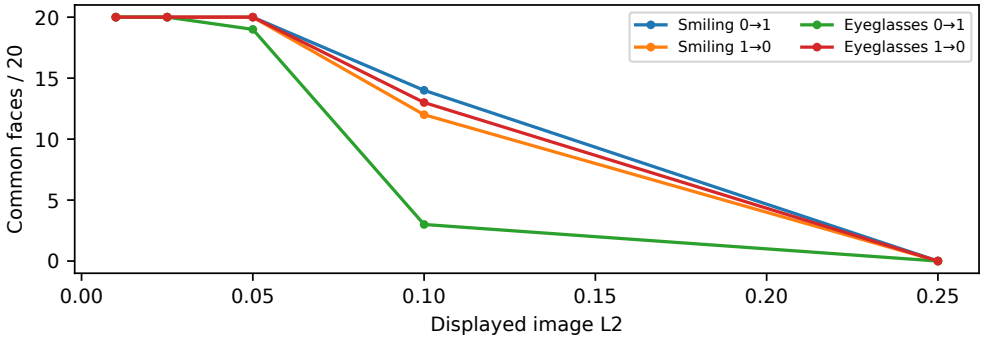


Figure 4: Common attainment across all bands, signs, and controls. Zero support at the largest budget means no face supports the full comparison, not that all individual edits fail.

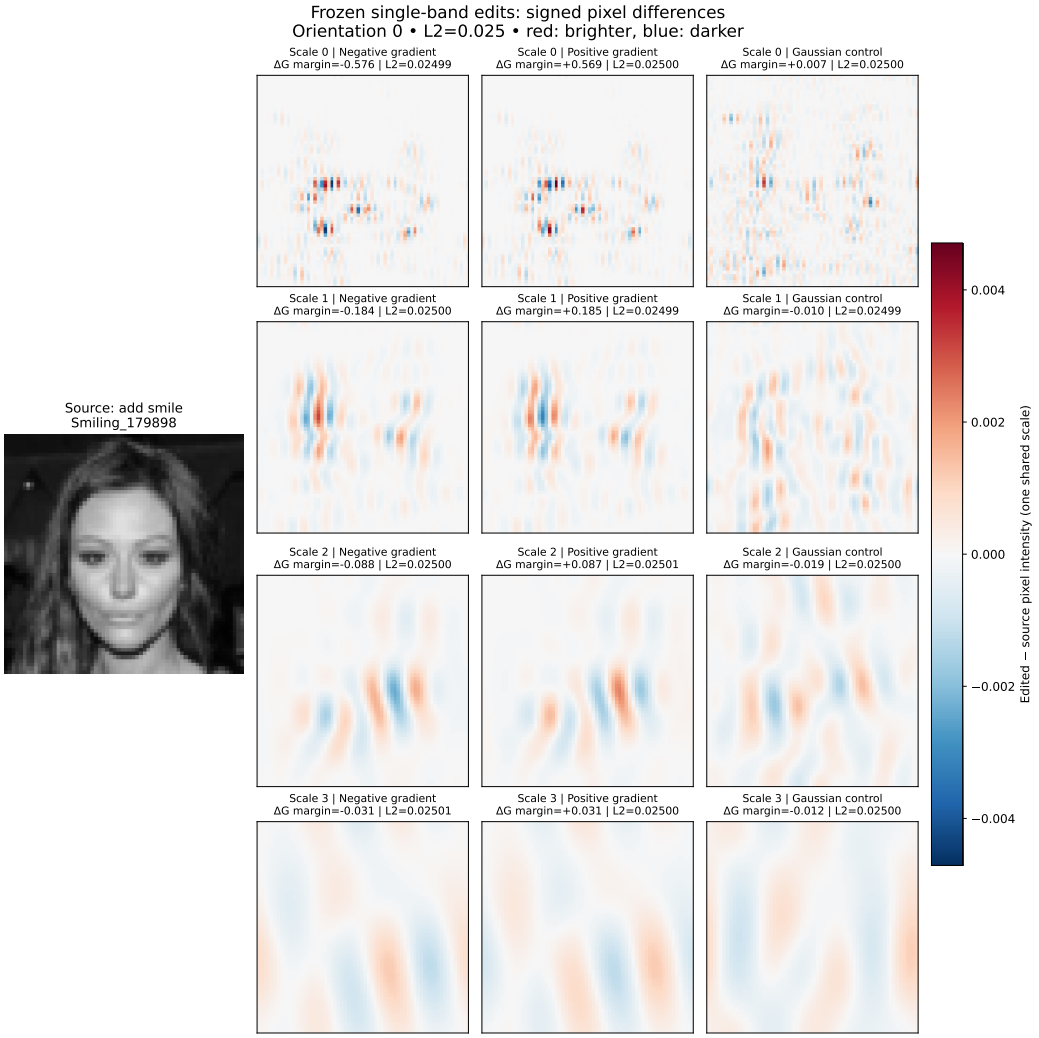


Figure 5: Small-budget band interventions on development source Smiling_179898 at $L_2 = 0.025$, orientation 0, and scales 0–3. Signed differences use floating-point edited images and a shared symmetric color range of approximately ± 0.004711 in $[0, 1]$ image units. They are not separately normalized per panel. The annotations describe driver logit-margin changes, not probabilities. Difference visualization exposes changes that may be imperceptible in ordinary image displays; it is not an enlarged-budget edit.

provided voluntarily by a convenience sample of friends and family. We retained two complete response sets for descriptive analysis. Recruitment through personal contacts and the small sample limit generalizability, and familiarity with the authors may influence judgments.

E ADDITIONAL MATCHED-BUDGET EXAMPLES

The following panels use actual scored persistent-coordinate endpoints from the main structural runs. Within a row, all methods share the source and rendered-image budget. MNIST uses $L_2 = 2$ and CelebA uses $L_2 = 128/28$, equal in per-pixel RMS across resolutions. G denotes target probability, and E the 8×8 warp-plus-gain explained energy. In the existing panel labels, "both judges: no" means that at least one judge does not predict the target, not that both reject it. The supplementary panels take the first two common-attainment sources in saved order for each group. They are illustrative cases, not human-validated target conversions. The main-text figure was selected separately after inspection of appearance and outcomes.

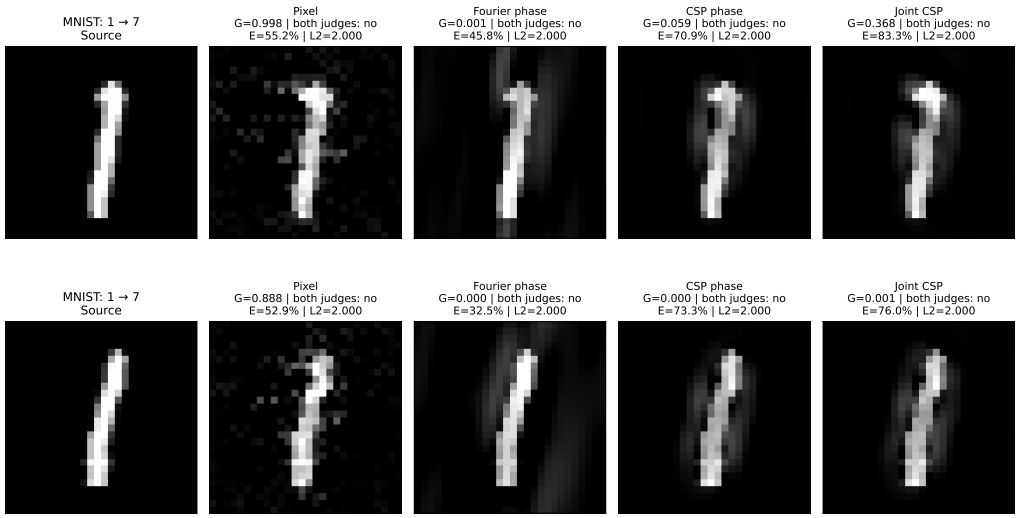


Figure 6: MNIST $1 \rightarrow 7$: two common-attainment sources with pixel, Fourier-phase, CSP-phase, and joint-CSP endpoints. Each row is matched in rendered image distance. Confidence, joint target agreement, and smooth reconstruction describe distinct properties; none certifies human recognition.

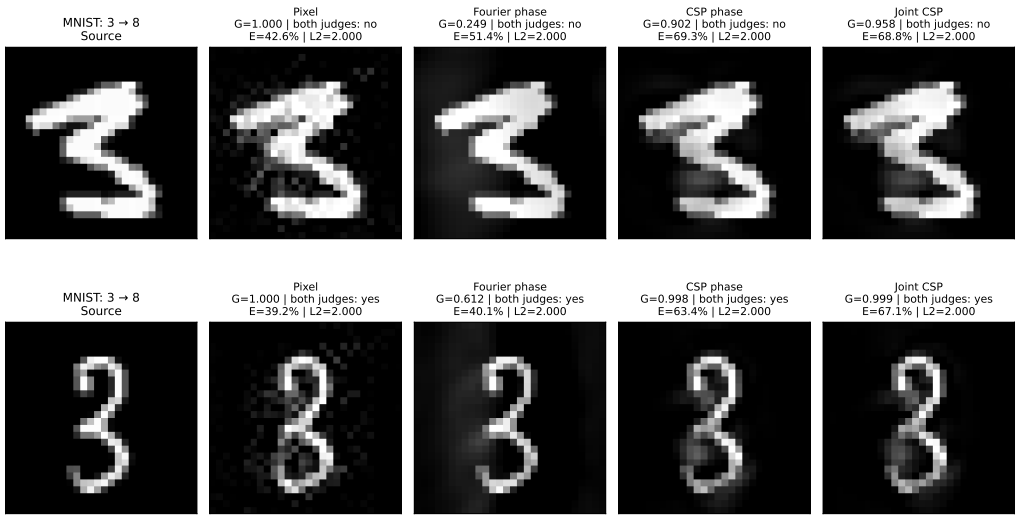


Figure 7: MNIST $3 \rightarrow 8$: two common-attainment sources with pixel, Fourier-phase, CSP-phase, and joint-CSP endpoints. Each row is matched in rendered image distance. Confidence, joint target agreement, and smooth reconstruction describe distinct properties; none certifies human recognition.

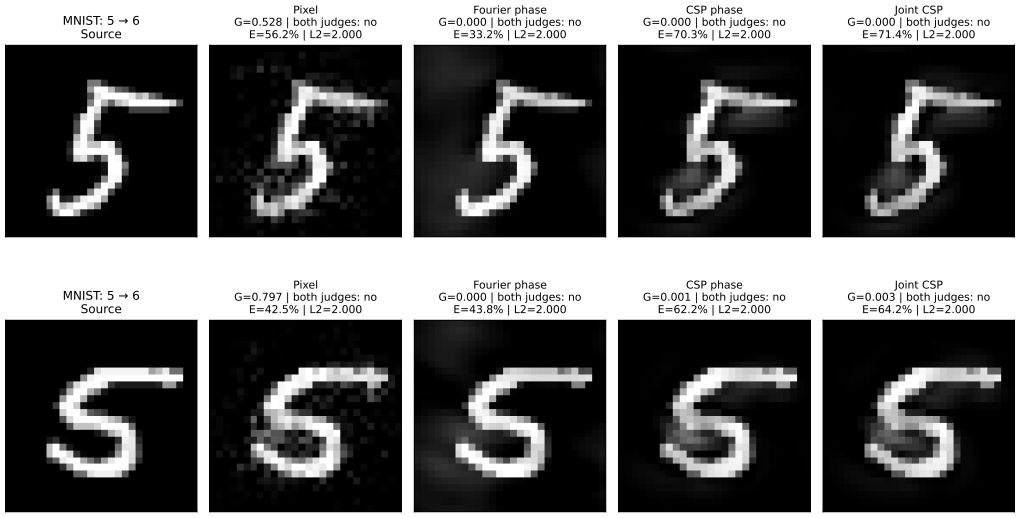


Figure 8: MNIST $5 \rightarrow 6$: two common-attainment sources with pixel, Fourier-phase, CSP-phase, and joint-CSP endpoints. Each row is matched in rendered image distance. Confidence, joint target agreement, and smooth reconstruction describe distinct properties; none certifies human recognition.



Figure 9: CelebA smile addition: two common-attainment sources with pixel, Fourier-phase, CSP-phase, and joint-CSP endpoints. Each row is matched in rendered image distance. Confidence, joint target agreement, and smooth reconstruction describe distinct properties; none certifies human recognition.

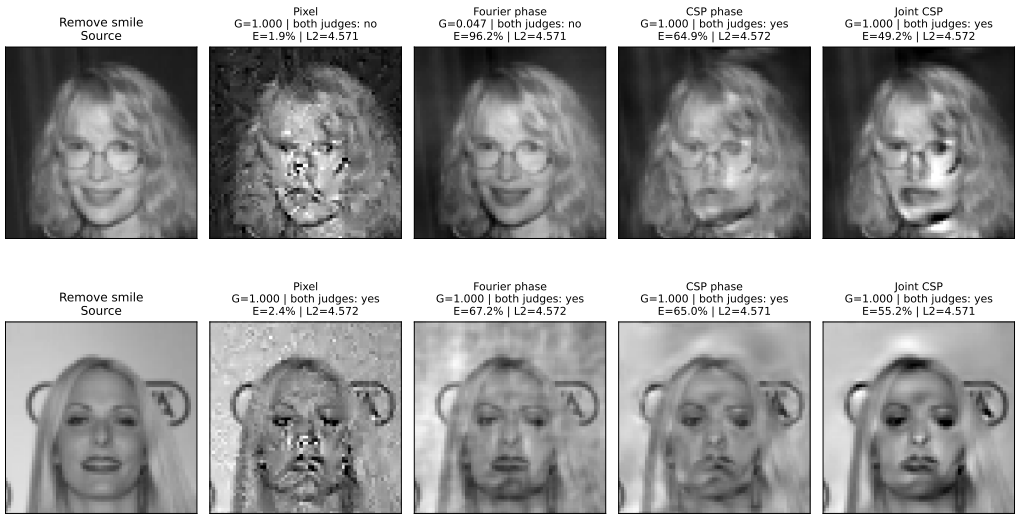


Figure 10: CelebA smile removal: two common-attainment sources with pixel, Fourier-phase, CSP-phase, and joint-CSP endpoints. Each row is matched in rendered image distance. Confidence, joint target agreement, and smooth reconstruction describe distinct properties; none certifies human recognition.

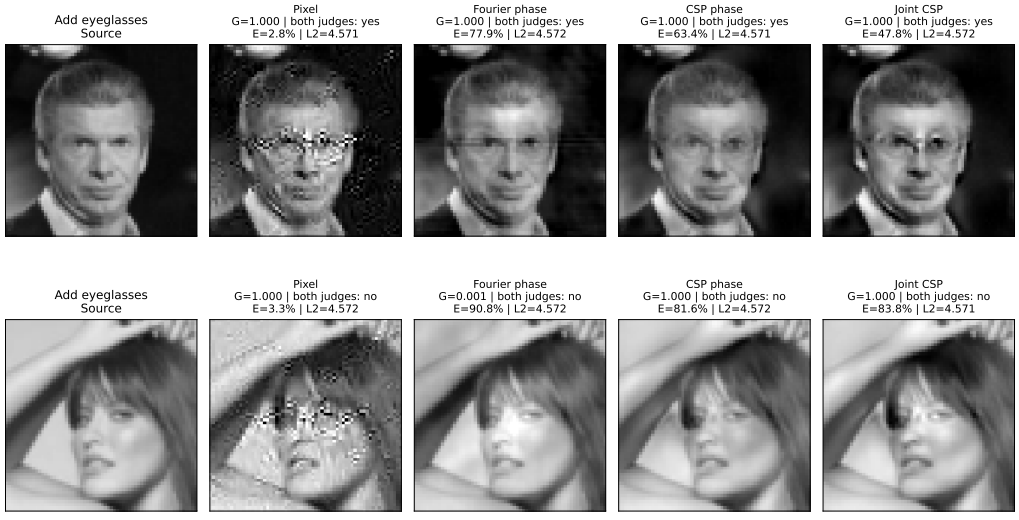


Figure 11: CelebA eyeglasses addition: two common-attainment sources with pixel, Fourier-phase, CSP-phase, and joint-CSP endpoints. Each row is matched in rendered image distance. Confidence, joint target agreement, and smooth reconstruction describe distinct properties; none certifies human recognition.

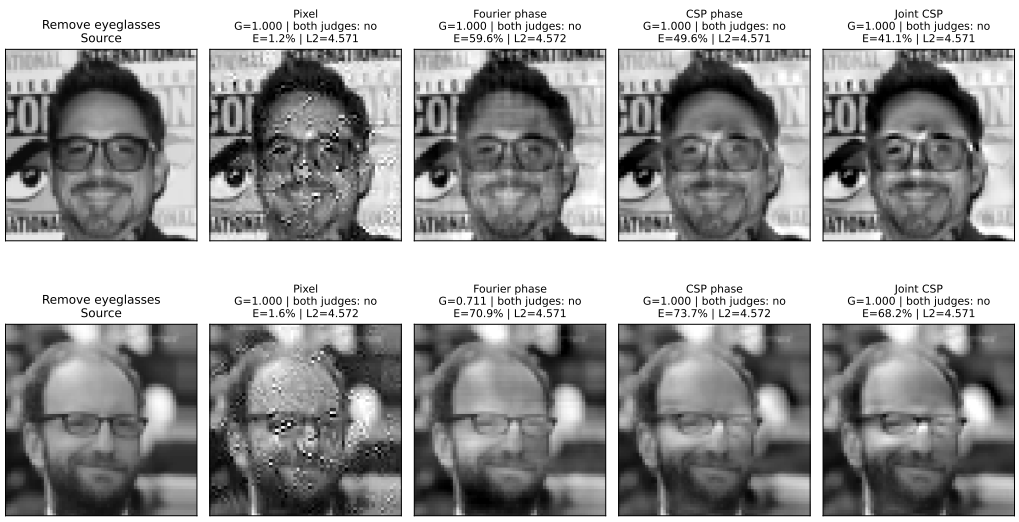


Figure 12: CelebA eyeglasses removal: two common-attainment sources with pixel, Fourier-phase, CSP-phase, and joint-CSP endpoints. Each row is matched in rendered image distance. Confidence, joint target agreement, and smooth reconstruction describe distinct properties; none certifies human recognition.

F NATURAL-IMAGE STEP-LOCAL TRAJECTORIES

We use three local cat photographs at 224×224 with ImageNet-pretrained classifiers (Deng et al., 2009; He et al., 2016); the photographs are not drawn from the ImageNet validation set. Frozen single-band probes use a standard ResNet-18 driver and a robust ResNet-50 judge. They compare separate RGB-channel, Y-only, and shared-RGB phase interventions at matched attained image distances, using a target-minus-source logit-margin gradient evaluated once at the source. Their driver responses differ across these choices, and none attains $L_2 = 8$ within the tested phase cap and search limits. This is non-attainment under the tested protocol, not a representational impossibility.

The all-band step-local experiment uses the reference complex-pyramid backend, a robust ResNet-50 driver G, and a standard ResNet-18 classifier O excluded from optimization. We generate twelve-frame trajectories for two selected source–target pairs, cat toward Egyptian cat and cat toward tiger cat, at per-step phase caps of 0.25, 0.5, and 1 radian. At each step we convert the current RGB image to YIQ, reanalyze Y, and compute the gradient of target cross-entropy with respect to all oriented-band phase increments. We normalize this gradient by its largest absolute component across bands and step in the descent direction. Backtracking tests up to twelve phase scales, starting at the cap and halving after each rejection. A candidate is accepted only if target cross-entropy decreases on the rendered RGB image; otherwise the current frame is repeated.

Band magnitudes and residuals remain fixed within each step. The current frame's I/Q channels are retained during RGB reconstruction, followed by clipping to $[0, 1]$. Clipping and reanalysis can change the recovered magnitudes, residuals, and chrominance, so these quantities are not constrained to remain equal to their original-source values throughout the trajectory. Both classifiers score the rendered RGB image.

At phase cap 1, the final target probabilities for cat toward Egyptian cat are 0.606 under G and 0.301 under O, at $L_2 = 27.84$. For cat toward tiger cat they are 0.534 and 0.238, at $L_2 = 28.07$. These distances exceed the frozen-band budgets and differ across settings. Model roles and objectives also differ between the frozen probes and live runs, so these experiments do not isolate the effect of refreshing the gradient. They illustrate repeated classifier-directed steering without independently establishing recognizable category conversion. The historical local-pyramid illustrations in Appendices G and H use a different operator and are reported separately.

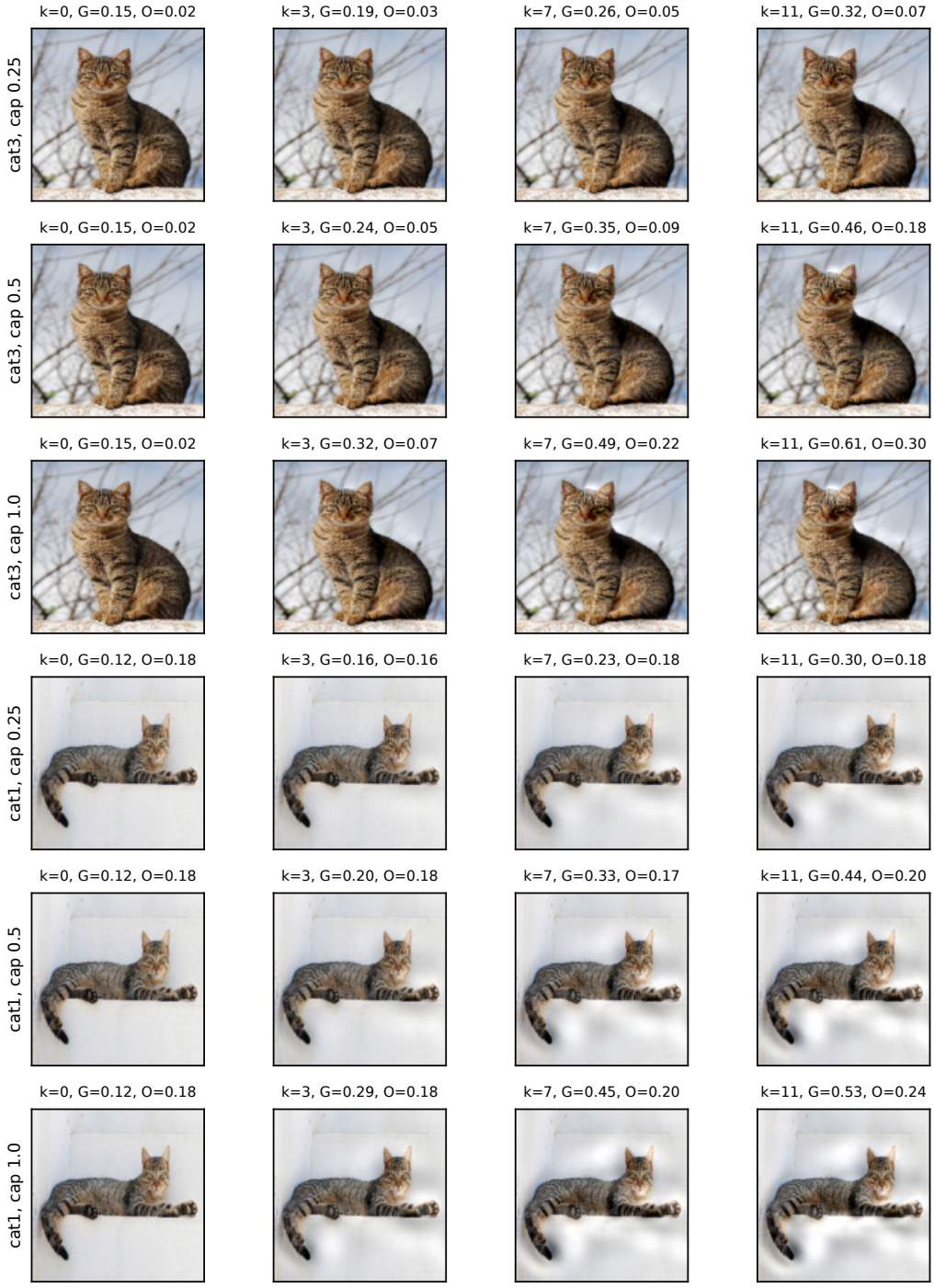


Figure 13: All six reference-backend YIQ trajectories: cat toward Egyptian cat and cat toward tiger cat, each at per-step phase caps 0.25, 0.5, and 1 radian. Frames 0, 3, 7, and 11 of each twelve-frame trajectory are displayed. The G and O annotations give the requested target's probability under robust ResNet-50 and standard ResNet-18, respectively. Image distances are not matched across trajectories; these are selected illustrations rather than a method-ranking benchmark.

G ADDITIONAL CAT–DOG CLOSE-UP STEERING ILLUSTRATIONS

These seven selected strips supplement the reference-backend trajectories in Appendix F. They use a different, historical `LocalPyramid` backend (order 4, automatic height), all-band live luminance-phase editing, and twelve frames at 224×224. The driver is a robust ResNet-50 and the additional classifier is a standard ResNet-18. Gradients are refreshed during the trajectory; magnitudes and residuals are held fixed within each step, not guaranteed fixed relative to the source across steps. The historical amplification settings 120 and 1000 are coordinate-specific scalars, not matched image budgets or magnitude amplifiers. The magnitude amplifier is zero in these runs.

The requested target is English springer for the cat close-up and Egyptian cat for the three dog close-ups. Labels above each frame give each classifier's top prediction, its probability for the requested target, and the rendered image distance from the source. Top predictions and target probabilities are distinct: a non-target top prediction can coexist with nonzero target probability. These selected examples display classifier-driven appearance and texture changes, including substantial degradation at large amplification. They do not establish species conversion, preservation of identity, or superiority over another representation. They were not included in the matched-budget structural analysis or the CelebA human study.

The figures display all twelve frames of each trajectory with the original annotations.

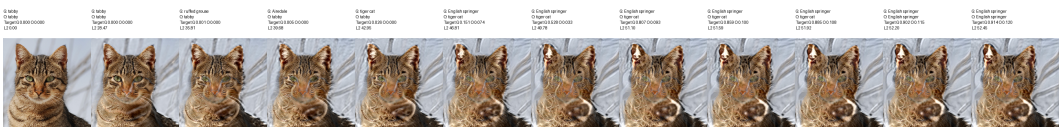


Figure 14: Selected cat-face trajectory toward the English springer target, historical amplification 120. Driver and judge target probabilities and image distance are shown for each frame. This is an unmatched step-local illustration from the historical local-pyramid backend.

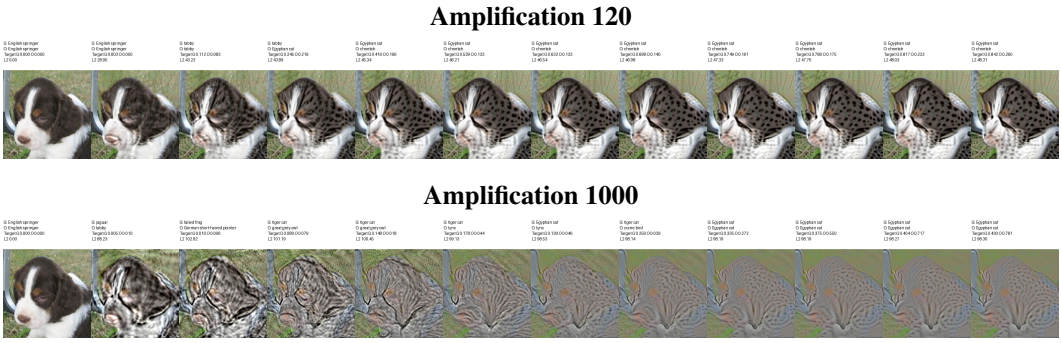


Figure 15: Dog close-up 1 steered toward Egyptian cat, with all twelve frames retained at each amplification. Final values are amplification 120: $G = 0.842$, $O = 0.260$, $L_2 = 48.31$; amplification 1000: $G = 0.430$, $O = 0.781$, $L_2 = 98.30$. The two trajectories are not matched in image distance; higher amplification does not necessarily yield higher driver confidence or a more recognizable target. Both use the historical local-pyramid backend.

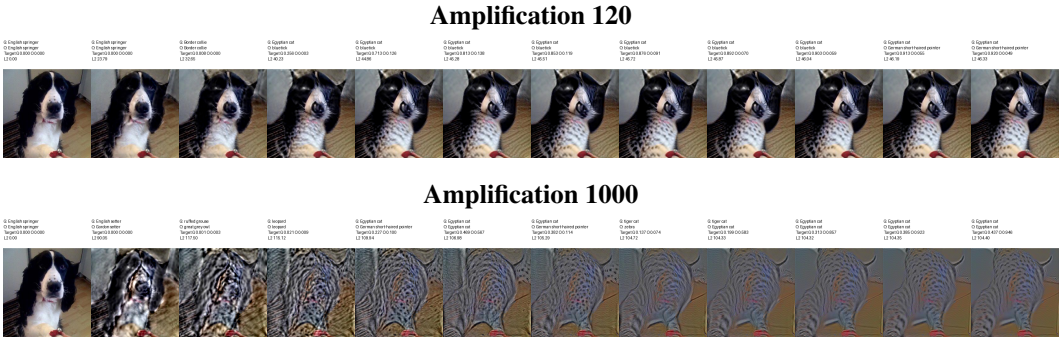
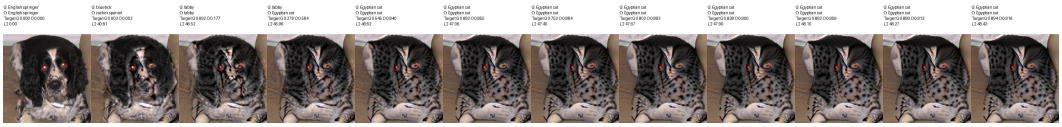


Figure 16: Dog close-up 2 steered toward Egyptian cat, with all twelve frames retained at each amplification. Final values are amplification 120: $G = 0.920$, $O = 0.049$, $L_2 = 46.33$; amplification 1000: $G = 0.437$, $O = 0.946$, $L_2 = 104.40$. The two trajectories are not matched in image distance; higher amplification does not necessarily yield higher driver confidence or a more recognizable target. Both use the historical local-pyramid backend.

Amplification 120



Amplification 1000

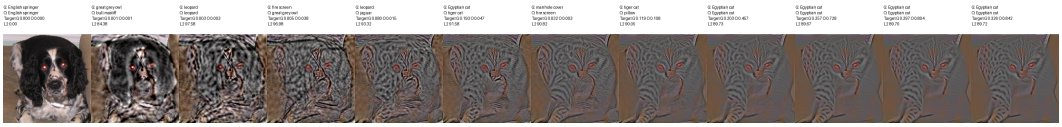


Figure 17: Dog close-up 3 steered toward Egyptian cat, with all twelve frames retained at each amplification. Final values are amplification 120: $G = 0.894$, $O = 0.916$, $L_2 = 48.43$; amplification 1000: $G = 0.326$, $O = 0.842$, $L_2 = 89.72$. The two trajectories are not matched in image distance; higher amplification does not necessarily yield higher driver confidence or a more recognizable target. Both use the historical local-pyramid backend.

H WITHIN-CAT LUMINANCE TRAJECTORIES

These two additional strips reproduce historical all-band live luminance-phase trajectories using the local-pyramid implementation. They are separate from the reference-backend, phase-capped runs in Appendix F. The source–target pairs and amplification settings were selected for illustration. These figures retain the original driver class labels and confidence annotations from the historical exports. They do not show independent-judge scores, and we do not use these annotations for quantitative comparison. The strips do not provide matched image budgets or a comparison with pixel and Fourier editing.

I FOUR-METHOD MNIST TRAJECTORIES: FROZEN AND RECOMPUTED GRADIENTS

The matched-budget MNIST panels in Appendix E compare pixel, Fourier phase, CSP phase, and joint CSP at $L_2 = 2$. Figure 20 additionally displays all four methods with frozen and recomputed gradients on the same source, at requested distances 1, 2, and 4. Here "joint CSP" means oriented-band phase and magnitude editing with fixed residuals; "CSP phase" fixes the supplied band magnitudes as well.

This historical comparison uses a target-minus-source margin, a frozen scalar sweep or calibrated iterative SGD, and persistent source coordinates. It is distinct from the headline adaptive cross-entropy replication and from step-local reanalysis. The source is the first recorded protocol source, without outcome-based selection. Non-attained budgets are shown explicitly; blank placeholders do not represent zero-valued images. The figure illustrates schedule dependence under this protocol, not a comparison across identical optimizers in every experiment.



Figure 18: cat toward Egyptian cat, historical amplification 1000. All twelve displayed frames are retained. These selected live trajectories illustrate the historical operator and driver response; they do not validate semantic conversion or superiority over other methods.



Figure 19: cat toward tiger cat, historical amplification 120. All twelve displayed frames are retained. These selected live trajectories illustrate the historical operator and driver response; they do not validate semantic conversion or superiority over other methods.

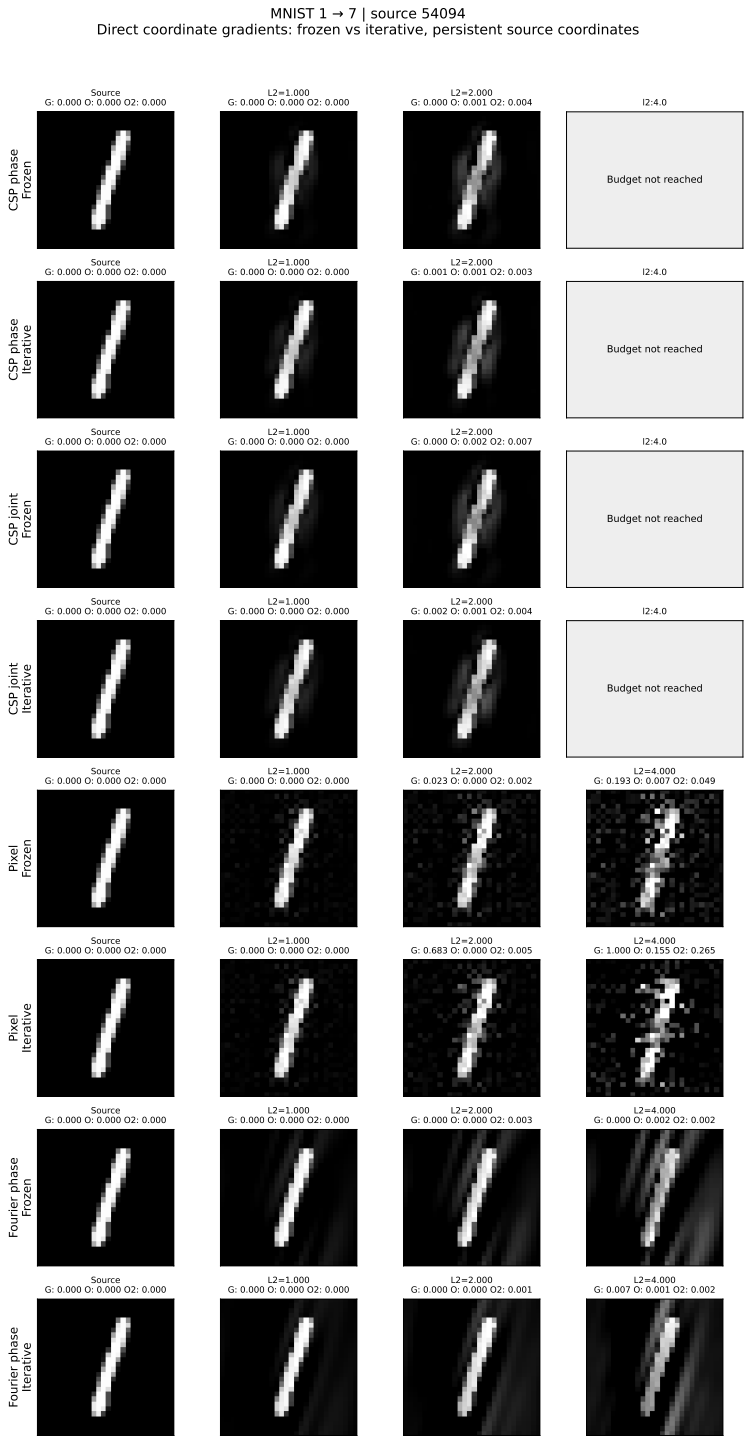


Figure 20: Four MNIST intervention methods, each with frozen and recomputed gradients, on one source. Columns show the source and requested L_2 budgets 1, 2, and 4. G, O, and O_2 annotations are target probabilities. Explicit non-attainment prevents interpreting a smaller edit as budget-matched. This historical margin-based comparison is separate from the main adaptive-CE results.

J SELECTED JOINT-CSP ITERATIVE MNIST TRAJECTORIES

The following examples were selected for illustration from a gallery filtered by budget attainment and driver target probability. Its selection required attainment of budgets $0.5, 1, 2, 3, 4$ and driver target probability at least 0.95 at $L_2 = 4$ and at the final endpoint. It uses direct-coordinate, persistent joint-CSP optimization with refreshed gradients and adaptive target cross-entropy; source residuals remain fixed. The source gradient is not frozen, and each frame is not reanalyzed as a new anchor.

These single-method strips are supplementary illustrations, not unbiased samples or four-method comparisons. The source is followed by the five attained budget endpoints and the final optimizer endpoint, which may exceed $L_2 = 4$. Annotations report the update index, rendered distance, and target probabilities under G, O, and O_2. High driver probability is an explicit selection criterion, not independent evidence of recognizable conversion. Several transitions reuse the same source image, and judges may disagree despite high driver confidence.

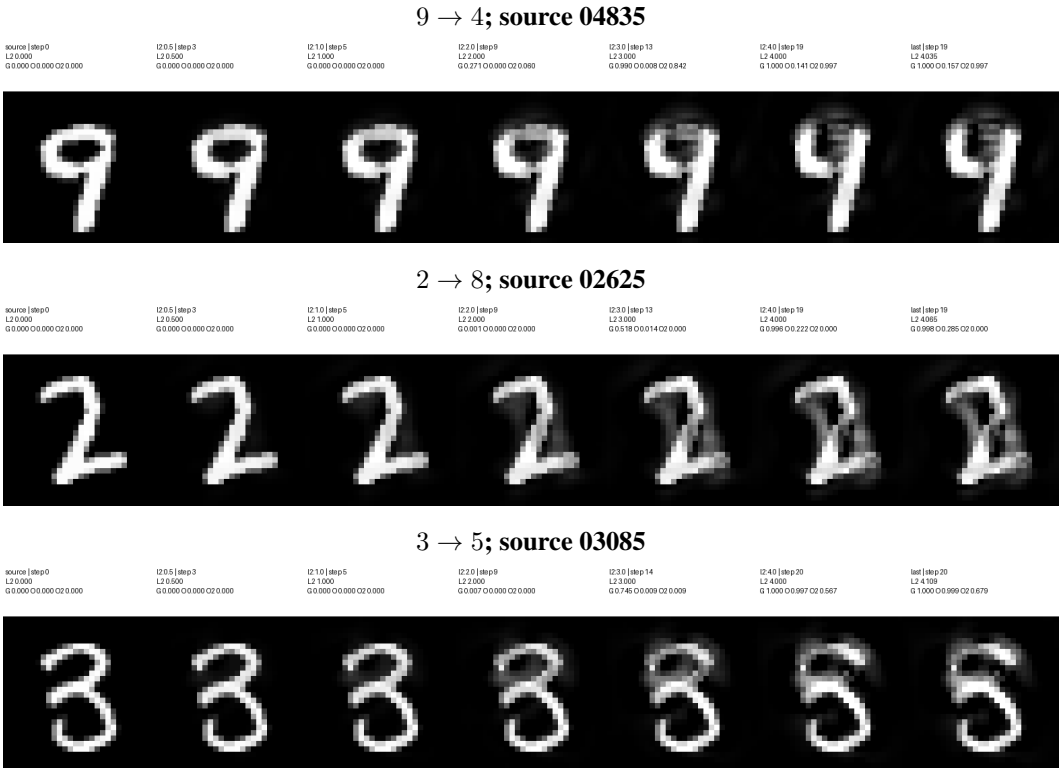
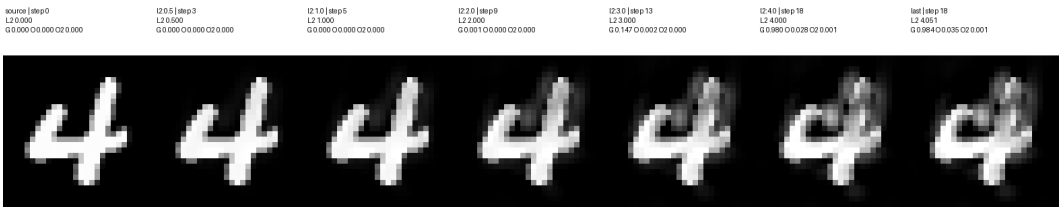


Figure 21: Selected joint-CSP iterative MNIST trajectories. All rows were selected from a high-driver-confidence, budget-attaining gallery; they are not representative success-rate evidence. Distances and independent-judge probabilities remain visible, including low judge responses.

$4 \rightarrow 9$**; source 06144**



$7 \rightarrow 8$**; source 07796**



$8 \rightarrow 7$**; source 09609**

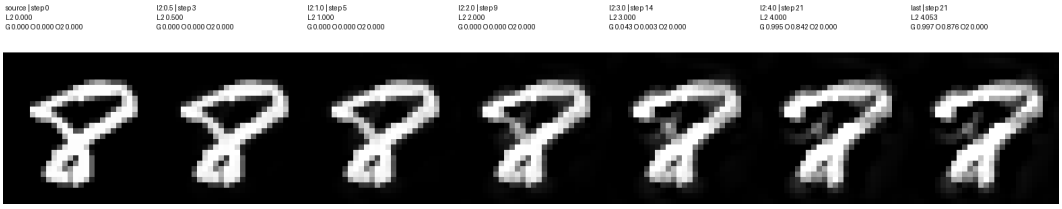


Figure 22: Selected joint-CSP iterative MNIST trajectories. All rows were selected from a high-driver-confidence, budget-attaining gallery; they are not representative success-rate evidence. Distances and independent-judge probabilities remain visible, including low judge responses.

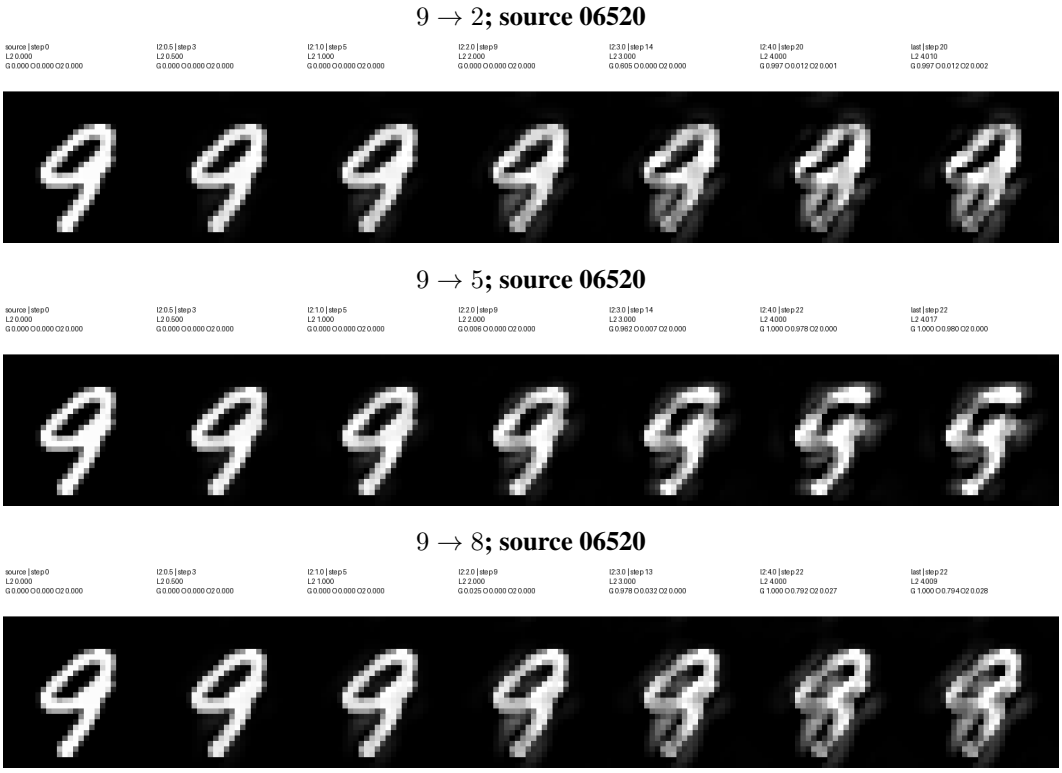


Figure 23: Selected joint-CSP iterative MNIST trajectories. All rows were selected from a high-driver-confidence, budget-attaining gallery; they are not representative success-rate evidence. Distances and independent-judge probabilities remain visible, including low judge responses.

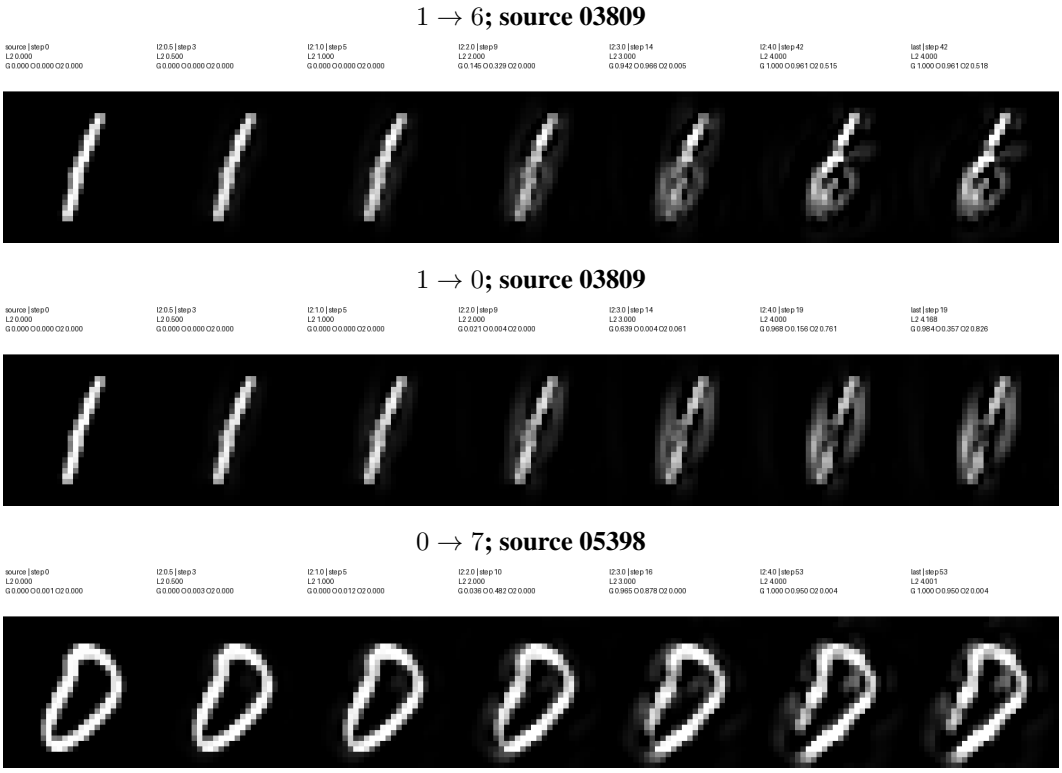


Figure 24: Selected joint-CSP iterative MNIST trajectories. All rows were selected from a high-driver-confidence, budget-attaining gallery; they are not representative success-rate evidence. Distances and independent-judge probabilities remain visible, including low judge responses.