ENVISION: GEOMETRY-AWARE SETWISE RANKING FOR FACIAL SURGERY VISUALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Stochastic image editing can generate several plausible facial surgery visualizations, but selecting one without postoperative ground truth is difficult. Identity preservation and visual realism alone do not establish whether an edit captures the intended anatomical change. We introduce **Envision**, a geometry-aware setwise ranker for rhinoplasty, blepharoplasty, and rhytidectomy. A fixed FLUX.1-Fill backbone generates candidate pools using different random seeds and edit intensities. The ranker combines procedure-normalized morphometric changes, geometric information pooled across candidates, and learned appearance features. A procedure-conditioned direction predictor assesses how each candidate's change relates to the estimated case-level direction. Paired postoperative images provide teacher supervision during training, whereas inference uses only the preoperative portrait, procedure label, and generated candidates. Across a 125-case evaluation cohort combining validation and test, Envision improves mean teacher SurgicalScore over a handcrafted selector by 0.034 with 15 candidates and 0.038 with 25 candidates, recovering approximately 19% and 18% of the available oracle headroom, respectively. The gains chiefly reflect better agreement in the direction and magnitude of geometric change and are strongest for rhytidectomy, although cached-vision ablations do not isolate a benefit from the setwise or geometric branches alone. Performance levels off beyond 15 candidates, and the test-only improvements remain statistically inconclusive. These results suggest that geometry-aware selection can improve surgical visualizations without changing the generator. They do not establish that the selected images predict individual clinical outcomes. Our code is available at anonymous.4open.science/r/envision-anonymous-source-D013.

1 INTRODUCTION

Stochastic image editing can produce several plausible facial surgery visualizations, but a system must decide which candidate to present. For the same preoperative portrait and procedure, changes in random seed or edit intensity may yield substantially different transformations. A visually convincing image may change too little, while a more pronounced edit may alter the face in the wrong direction. The practical value of generation therefore depends on whether a suitable result can be selected using information available before surgery. We study this problem for rhinoplasty, blepharoplasty, and rhytidectomy (Figure 1).

Recent advances provide increasingly flexible tools for generating these alternatives. Latent diffusion supports high-resolution synthesis and inpainting (Rombach et al., 2022), ControlNet introduces spatial guidance (Zhang et al., 2023), and InstructPix2Pix learns image transformations from textual instructions (Brooks et al., 2023). For surgical visualization, Envisage combines localized diffusion editing with mask-decomposed evaluation (Agarwal & Bhrany, 2026). Its ground-truth-aware best-of-five analysis reveals substantial quality differences within a candidate pool. However, the postoperative reference used to identify the best sample is unavailable at deployment. Oracle performance thus measures the quality available in the pool, leaving open how much a practical selector can recover.

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.

Closing this gap requires a criterion sensitive to the intended transformation. Research on recognition across surgery and disguise emphasizes identity robustness (Singh et al., 2010; Suri et al., 2018; Rathgeb et al., 2020); identity embeddings such as ArcFace serve a similar preservation objective (Deng et al., 2019). Structural and perceptual similarity measures, including SSIM and LPIPS (Wang et al., 2004; Zhang et al., 2018), quantify agreement with a chosen reference. Yet similarity to the input may reward an unchanged portrait, whereas postoperative similarity requires unavailable information. Moreover, hard compositing makes outside-mask preservation nearly identical across candidates. These observations motivate evaluating the content of the local edit, particularly its anatomical direction and magnitude.



Figure 1: Selections from identical K=25 candidate pools. SS is the retrospective teacher SurgicalScore, used only for evaluation; neither selector uses postoperative ground truth at inference.

Facial outcome assessment provides a basis for this geometric perspective: prior studies relate measurements to rhinoplasty outcomes and characterize postoperative morphology, symmetry, and perceived age (Topsakal et al., 2023; Varghaei et al., 2025). Face-mesh estimation makes local geometry accessible from monocular images (Kartynnik et al., 2019). We use procedure-normalized measurement changes to describe each candidate, complemented by appearance features that capture texture and artifacts. Candidates from the same case also share input anatomy and procedure context. Pooling their geometric features, following permutation-invariant set aggregation (Zaheer et al., 2017), provides comparative evidence about the generated transformations without assuming that the most common change is correct.

We introduce Envision, a geometry-aware setwise ranker that selects among candidates from a fixed FLUX.1-Fill backbone. The model combines candidate geometry, pooled case context, and learned appearance features, with a procedure-conditioned direction predictor assessing each edit relative to the estimated case-level change. Paired postoperative images supply retrospective teacher scores during training. At inference, the selector requires only the preoperative portrait, procedure label, and generated pool; GT-free therefore refers to inference rather than supervision.

Our contributions are a formulation of surgical visualization as candidate selection under unavailable postoperative ground truth, a cross-procedure ranker integrating geometric and visual evidence with set context, and a controlled evaluation on fixed candidate pools. Comparisons across pool sizes, a handcrafted selector, and component ablations quantify selection regret and the contribution of each information source. Envision targets agreement with a retrospective assessment protocol; its outputs are visualizations rather than validated predictions of individual clinical outcomes.

2 Related Work

Facial analysis and outcome assessment. Work on face recognition across surgery and disguise seeks identity-invariant representations (Singh et al., 2010; Suri et al., 2018; Rathgeb et al., 2020). ArcFace (Deng et al., 2019) diagnoses identity preservation, but cannot judge whether an anatomical edit is appropriate. Studies of rhinoplasty outcomes and postoperative morphology motivate geometric measurements (Topsakal et al., 2023; Varghaei et al., 2025); facial meshes (Kartynnik et al., 2019) and monocular depth (Yang et al., 2024) make geometry accessible from images. Retrospective assessment, however, does not establish patient satisfaction or surgical success.

Conditional editing for surgical visualization. Point-cloud simulation can model soft-tissue change from planned skeletal movement (Ma et al., 2021). Image editors instead use synthesis, spatial guidance, or instructions (Rombach et al., 2022; Zhang et al., 2023; Brooks et al., 2023). Envisage applies localized diffusion editing to surgical portraits (Agarwal & Bhrany, 2026). We keep generation fixed and compare selectors on identical pools.

Localized evaluation and selection. SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and ArcFace assess different aspects of similarity; their meaning depends on the reference and image region. Similarity to the input measures preservation, while similarity to an observed postoperative portrait measures retrospective agreement. Envisage separates oracle from deployable selection through SurgicalScore (Agarwal & Bhrany, 2026). We use its postoperative-dependent score only for training and evaluation, and compare selectors on the same generated images. Our candidate-set aggregation follows Deep Sets (Zaheer et al., 2017), but training regresses pointwise teacher scores rather than using a listwise loss.

3 METHOD

3.1 PROBLEM FORMULATION

Given a preoperative portrait x_i^{pre} and a procedure label $p_i \in \{\text{rhinoplasty}, \text{blepharoplasty}, \text{rhytidectomy}\}$, a stochastic generator G produces a pool $\mathcal{C}_i = \{c_{ij}\}_{j=1}^{K}$ of K simulated postoperative images. Because postoperative ground truth x_i^{post} is unavailable at deployment, the system selects one candidate according to

$$\hat{\jmath}_i = \arg\max_{j \in [K]} s(c_{ij} \mid \mathcal{C}_i, x_i^{\text{pre}}, p_i), \tag{1}$$

where $s(\cdot)$ scores candidate c_{ij} conditioned on the entire candidate set $\mathcal{C}_i$ as context, computed strictly without the postoperative image. We refer to such a scoring function as GT-free.

Candidate quality is evaluated retrospectively using $T(c_{ij} \mid x_i^{\text{pre}}, x_i^{\text{post}}, p_i)$, a teacher score that has access to postoperative ground truth (hereafter abbreviated as $T(c_{ij})$ when input conditioning is clear from context). The teacher is used only for supervision and evaluation and is never provided to $s(\cdot)$ at deployment. The objective is to maximize the expected teacher score of the selected candidate, $\mathbb{E}_i[T(c_{i\hat{\jmath}_i})]$. We define selection regret as the gap between a GT-free selector and the oracle $j_i^\star = \arg\max_j T(c_{ij})$. Figure 2 summarizes candidate generation, setwise ranking, test-time selection, and offline teacher supervision.

This formulation is motivated by the variation within each candidate pool. For the same input, changing the diffusion seed or edit intensity can produce candidates with substantially different teacher scores. Candidate selection can therefore affect the final result more than a small change to the generator.

3.2 CANDIDATE GENERATION AND QUALITY METRIC

All candidates are generated by a fixed FLUX.1-Fill-dev backbone (Figure 2a). The model performs inpainting within a procedure-specific surgical mask, compositing the generated region onto the preoperative image:

$$c = M \odot \hat{c} + (1 - M) \odot x^{\text{pre}},$$

where $\odot$ denotes the element-wise product, preserving input pixels outside the mask. For each case, we extract 478 landmarks using MediaPipe FaceLandmarker (Lugaresi et al., 2019) along with a depth map, a procedure-specific prompt, and the surgical mask. The mask M is deterministically generated from facial landmark contours that define standard anatomical intervention zones, requiring no manual physician annotation. These inputs are computed once and shared across all candidates for the same case. We obtain candidate diversity by varying only the diffusion seed and the edit intensity, which is set to 25%, 50%, or 80%. We evaluate candidate pools with K=5, K=15, and K=25 on the same 125 held-out cases, including 58 rhinoplasty, 37 blepharoplasty, and 30 rhytidectomy cases. The ranker is trained on a separate split.

The evaluation-only teacher assigns each candidate a score in $[0, 1]$ using five components. Components A and B measure agreement between the candidate and ground-truth landmark displacements in direction and magnitude, respectively. Component C measures photometric fidelity to the postoperative image within the surgical mask. Component D measures realism without using a reference image, and component E measures preservation outside the mask. The five components are combined via input-anchored calibration and an identity gate with a threshold of 0.65:

$$T(c_{ij}) = \mathbb{I}\left(\text{Sim}_{\text{id}}(c_{ij}, x_i^{\text{pre}}) \geq 0.65\right) \cdot \sum_{k \in \{A,B,C,D,E\}} w_k S_k(c_{ij}), \tag{2}$$

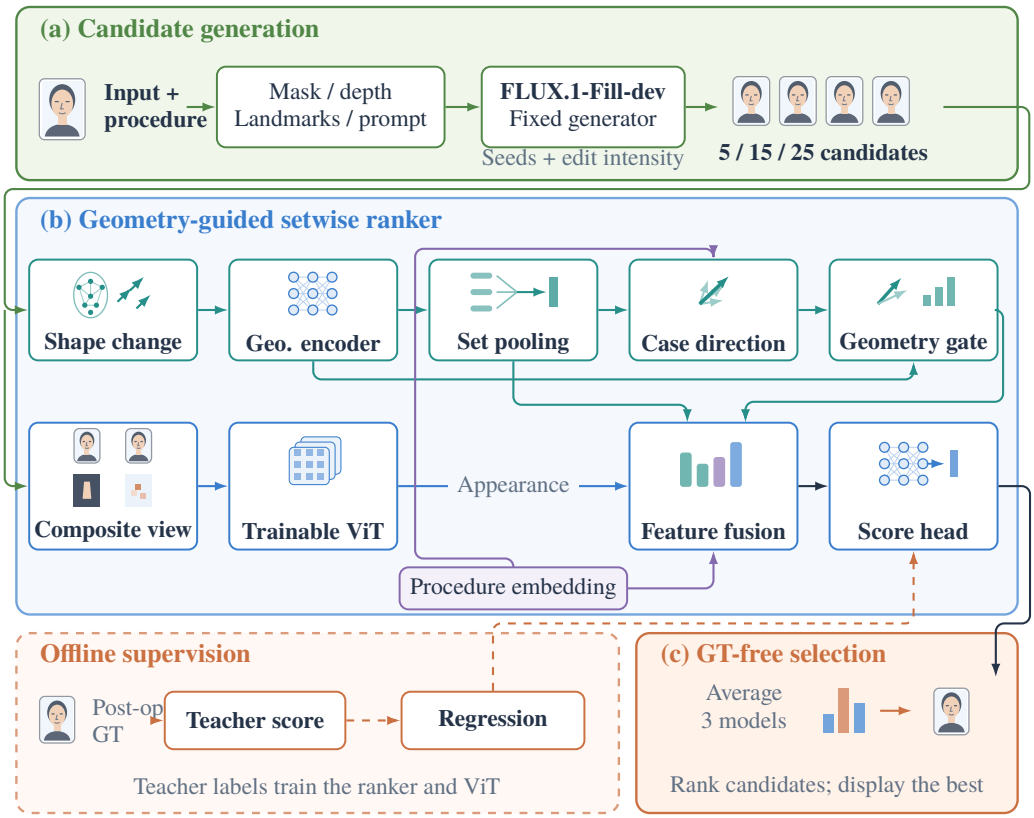

Figure 2: Overview of the proposed framework. (a) A fixed FLUX.1-Fill backbone generates candidates from a preoperative portrait and procedure-specific context using varied seeds and edit intensities. (b) A geometry-aware setwise ranker combines candidate shape changes, pooled case context, and trainable ViT features to score each candidate. (c) Scores are averaged across three independently trained models to select the best candidate without postoperative ground truth. The lower branch depicts offline teacher supervision for training and evaluation.

where $\mathbb{I}(\cdot)$ denotes the indicator function, $S_k \in [0, 1]$ are the component scores, and the weights w_k follow the official SurgicalScore protocol (Agarwal & Bhrany, 2026).

Two properties of the teacher motivate our model and analysis. The score has geometric agreement terms A and B and appearance/preservation terms C, D, and E. Candidate landmark changes are available at inference, but their agreement with the postoperative target is not; the ranker must learn a GT-free proxy for A and B. Teacher scores also vary more across cases than across seeds from the same case. Within-case ranking is therefore less variable than absolute score prediction across cases, which motivates the setwise formulation. We additionally report an ArcFace identity gap, defined as cosine(output, ground truth) minus cosine(input, ground truth). This measure is used only for diagnostic analysis and is not optimized by the ranker.

3.3 The Envisage deploy selector (baseline)

To measure the benefit of learning, we take the Envisage deploy selector as the baseline: a hand-crafted rule using GT-free signals available at deployment, reconstructed with its official weights and gates. Its inputs include identity similarity, outside-region similarity, landmark drift, color consistency, cross-candidate agreement, warp agreement, and an aesthetic score. The rule combines these terms using fixed weights:

$$s_{\text{rule}}(c) = 0.25u_{\text{id}} + 0.15\,(u_{\text{out}} + u_{\text{lm}} + u_{\text{col}} + u_{\text{cross}}) + 0.10u_{\text{tps}} + 0.05u_{\text{aes}}. \tag{3}$$

Candidates that fail the fidelity gate receive $s_{\text{rule}}(c) = -\infty$. The weights are specified in advance and are not fitted to teacher scores. Because the outside-preservation term uses the same input anchor

as component E of the teacher, the rule has a structural alignment with the evaluation metric. The learned selector must improve ranking despite this metric alignment; the effect of additional signals is tested empirically.

3.4 CROSS-PROCEDURE GEOMETRIC SETWISE RANKER

The ranker follows three design requirements (Figure 2b). It uses only information available without postoperative ground truth, applies one shared model to all three procedures, and processes candidates from the same case jointly. The setwise design reflects the shared but unobserved postoperative target associated with all candidates from a case.

Inputs. For each candidate c_{ij}, the geometric input is a five-dimensional morphometry delta

$$\delta_{ij} = m(c_{ij}) - m(x_i^{\text{pre}}) \in \mathbb{R}^5,$$

where $m(\cdot)$ computes a 5D vector of normalized anatomical measurements tailored to each procedure: dorsal height, tip projection, tip rotation, nasolabial angle, and alar width for rhinoplasty; palpebral fissure height, margin reflex distance, lid crease height, lateral canthal tilt, and brow-to-lid distance for blepharoplasty; and lower-face width, jowl droop, jawline angle, cervicomental angle, and nasolabial fold depth for rhytidectomy (landmark indices detailed in Appendix A). Each coordinate is standardized using a per-procedure z-score normalizer fitted only on the training split. The appearance input is a 2×2 composite view

$$v_{ij} = [x_i^{\text{pre}}, c_{ij}, M \odot c_{ij}, |c_{ij} - x_i^{\text{pre}}|],$$

which tiles the preoperative image, the candidate, the masked edit, and their absolute difference into a $2H \times 2W \times 3$ grid. A trainable ViT then encodes this view, enabling standard patch tokens to model inter-region differences directly. Neither input includes the seed, intensity level, or teacher score. The same checkpoint can therefore score candidates generated with unseen seeds or different pool sizes.

Architecture. A geometry MLP $\phi : \mathbb{R}^5 \to \mathbb{R}^{128}$ encodes the morphometry delta of each candidate. The set encoder aggregates the mean candidate representation:

$$z_i = \psi\left(\frac{1}{K}\sum_j \phi(\delta_{ij})\right).$$

The resulting case representation is a learned summary of the candidate pool; it need not identify the unobserved postoperative direction. A procedure-conditioned direction predictor maps this representation to $\hat{d}_i \in \mathbb{S}^4$. For candidate j, directional consistency is defined as

$$a_{ij} = \left\langle \frac{\delta_{ij}}{\max(\|\delta_{ij}\|, \epsilon)}, \hat{d}_i \right\rangle,$$

where $\epsilon > 0$ ensures numerical stability for zero-change candidates. This metric serves as an interpretable, ground-truth-free proxy for component A of the teacher score.

The scoring head combines the modulated candidate geometry $\tilde{\phi}(\delta_{ij}) = (1+0.1a_{ij})\phi(\delta_{ij})$, the case representation z_i, a 16-dimensional procedure embedding $e(p_i)$, and the projected ViT feature:

$$s_\theta(c_{ij}) = h\left(\tilde{\phi}(\delta_{ij}) \mid z_i \mid e(p_i) \mid W_v \text{ViT}(v_{ij})\right), \tag{4}$$

where $\mid$ denotes concatenation. The geometry encoder, set encoder, direction predictor, and scoring head are two-layer GELU MLPs with LayerNorm and dropout of 0.1.

Training and inference. The ranker is trained by regressing its predictions to teacher scores with a pointwise Huber loss and $\delta_{\text{Huber}} = 0.1$. Pairwise and listwise objectives were evaluated in ablations but are disabled in the reported model. We optimize the model with AdamW for 25 epochs. The ranker and ViT backbone use learning rates of 10^{-4} and 10^{-5}, respectively, with weight decay of 0.05, cosine scheduling, batch size 4, gradient clipping at 1.0, and deterministic seeds. Training is conducted exclusively on the 278-case training split with a fixed candidate pool size of $K_{\text{train}} = 15$ per case (sampled across available seeds and edit intensities) to stabilize setwise aggregation. A

cached-vision control trained with K_{train}=5 reaches 0.5764/0.6125 (val/test selected mean) vs. 0.5856/0.6501 with K_{train}=15 (same structure, same seed; NDCG $\sim$ 0.94 both), i.e. +0.038 on test from the larger training pool. Checkpoints are selected on the validation split based on oracle-gain recovery. Reported predictions are averaged across three independently trained models to reduce optimization variance. At inference (Figure 2c), all K candidates from a case are processed in one batched forward pass, and Eq. 1 is applied with $s = s_\theta$.

4 Experiments

We first compare selectors on matched candidate pools, then examine the teacher components and ablations behind the observed gains. External baselines, transfer, and failure cases define the scope of the result.

4.1 Evaluation protocol

We train on 278 cases and evaluate on the same nested K=5/15/25 candidate pools for 125 cases (58 rhinoplasty, 37 blepharoplasty, and 30 rhytidectomy). The 125 cases comprise 76 validation cases used for checkpoint selection and hybrid tuning, alongside a held-out test set of 49 cases. For transparency regarding hyperparameter selection, Table 1 reports an independent audit across splits. While the main results combine these splits (n=125) to maximize procedural representation across three surgical domains, performance gains remain positive on the held-out test set (+0.029 at K=15 and +0.023 at K=25), aligning with the direction observed on the validation set (+0.037 and +0.047). The test-split comparisons (p=0.17 and 0.23) are inconclusive at conventional significance levels; the sample size of n=49 limits precision, and these data do not establish the reason for non-significance. The retrospective teacher SurgicalScore supplies evaluation labels. Paired differences use two-sided sign-flip tests (B=5000) and percentile bootstrap 95% CIs (B=3000); full protocol details appear in Appendix B.

Table 1: Validation/test split audit. Checkpoints and the hybrid weight were selected on validation; test is untouched. Same pools, paired tests.

Cohort	Envision	Envisage	Δ	p
K=15 val (n=76)	0.592	0.555	+0.037	0.044
K=15 test (n=49)	0.666	0.637	+0.029	0.17
K=25 val (n=76)	0.596	0.549	+0.047	0.039
K=25 test (n=49)	0.653	0.630	+0.023	0.23

4.2 Selection performance across pool sizes

Table 2 compares the fixed-seed output, Envisage's deploy rule, Envision, and the teacher oracle. Envision is tied with the rule at K=5 (+0.002, p=0.43), but improves mean teacher score by +0.034 at K=15 (0.621 vs. 0.587, p=0.023) and +0.038 at K=25 (0.618 vs. 0.581, p=0.037) on the pooled cohort. These gains recover about 19% and 18% of the respective oracle headroom above the rule. Envision plateaus from K=15 to 25 (−0.003, p=0.57), despite increasing oracle quality; K=15 is therefore the practical operating point. The scaling plot is in Figure 3.

Table 2: Teacher SurgicalScore (mean [95% CI]) on identical candidate pools (n=125 held-out cases; fixed-seed n=95). Envision (our setwise ranker) vs. Envisage deploy (official fixed-weight rule, identical weights and gates); oracle is the GT upper bound. Paired sign-flip p vs. Envision.

Selector	K=5	K=15	K=25
Fixed-seed		0.583 [0.542, 0.622]	
Envisage deploy	0.589 [0.558, 0.619]	0.587 [0.556, 0.618]	0.581 [0.549, 0.612]
Envision (ours)	**0.591 [0.557, 0.623]**	**0.621 [0.585, 0.654]**	**0.618 [0.583, 0.652]**
Hybrid (α=0.65)	—	0.605 [0.575, 0.635]	—
Oracle (upper bound)	0.713 [0.688, 0.739]	0.769 [0.746, 0.791]	0.788 [0.768, 0.809]
Envision − Envisage	+0.002, p=0.43	+0.034, p=0.023	+0.038, p=0.037
Envision − oracle	−0.122	−0.148	−0.170

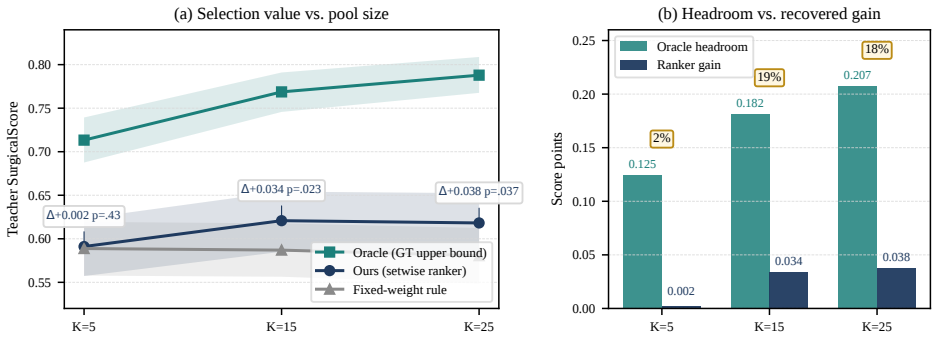

Figure 3: Pool-size scaling on identical 125-case cohorts. (a) Mean teacher score with 95% CIs: the ranker rises $0.591 \rightarrow 0.621$ then plateaus while the rule saturates and the oracle keeps growing. (b) Oracle headroom vs. ranker-recovered gain with recovery rate.

Procedure effects and ranking quality. At $K{=}25$, gains are positive for all three procedures, but only rhytidectomy is significant ($+0.058$, paired $p{=}0.005$; Table 3). Its gain recovers 67% of the rule's oracle headroom. These differences are compatible with procedure-specific edit variability, but do not isolate its cause. Rhytidectomy spans the lower face and neck, whereas blepharoplasty edits a smaller periorbital region; the observed heterogeneity may reflect the different ranges of candidate changes in these regions. Within-case Spearman correlation averages 0.25 at $K{=}15$; the selected candidate matches the teacher oracle in 17.6% of cases and lands among the oracle top three in 41.6%, compared with random rates of 6.7% and 20% (Appendix B, Figure 7). Per-case gain correlates with available oracle headroom (Pearson $r{=}0.54$; Appendix B, Figure 8), so higher pool quality creates opportunity, not guaranteed improvement.

Table 3: Per-procedure teacher means [95% CI] with paired ranker−rule tests. $n{=}58/37/30$; pooled $n{=}125$.

Cohort	Ranker	Rule	Oracle	Ranker − rule
$K{=}15$				
Rhinoplasty	0.617 [0.554, 0.677]	0.587 [0.536, 0.637]	0.837 [0.816, 0.858]	$+0.030$, $p{=}0.16$
Blepharoplasty	0.584 [0.527, 0.638]	0.579 [0.524, 0.632]	0.708 [0.657, 0.758]	$+0.005$, $p{=}0.43$
Rhytidectomy	0.674 [0.631, 0.715]	0.598 [0.550, 0.647]	0.711 [0.676, 0.745]	$+0.076$, $p{<}0.001$
$K{=}25$				
Rhinoplasty	0.616 [0.552, 0.679]	0.583 [0.533, 0.633]	0.850 [0.830, 0.868]	$+0.033$, $p{=}0.17$
Blepharoplasty	0.592 [0.533, 0.650]	0.563 [0.500, 0.623]	0.747 [0.703, 0.791]	$+0.029$, $p{=}0.21$
Rhytidectomy	0.656 [0.616, 0.695]	0.598 [0.550, 0.647]	0.719 [0.684, 0.754]	$+0.058$, $p{=}0.005$

4.3 Mechanism and Ablations

Teacher components and model controls. Decomposing the teacher score reveals that Envision's gains are concentrated in geometric alignment (Figure 4a, Table 5). Relative to the heuristic baseline, directional agreement A (0.600 vs. 0.532 at $K{=}25$) and magnitude agreement B (0.474 vs. 0.351) improve substantially, whereas visual terms C, D, and E remain nearly identical. Although ablations with cached vision features showed an apparent inversion where omitting geometry numerically increased scalar scores (0.633 vs. 0.618; Table 4), the end-to-end ablation has a different numerical pattern (Table 10). In this end-to-end setting, removing set context (0.6200) performs comparably to the full model (0.6208), whereas omitting geometry (0.6117) or the direction gate (0.6060, with the lowest teacher-A score of 0.561) lowers performance (Appendix C). These observations are consistent with useful geometric information in this setting; they do not establish a unique mechanism or rule out recovery of that information from visual features.

Operating-pool components. At $K{=}15$, A/B remain higher for Envision and C/D/E are nearly unchanged (Table 5); this pattern does not identify its causal component.

Ablation and robustness summary. Hybrid scoring and seed-dispersion rejection do not improve the selected mean; full candidate composition outperforms the i50-only pool (Table 6).

Table 4: Component ablations at K=15 (cached vision, three-seed mean, n=125). Full includes morphometry, set context, procedure embedding, direction gate, and frozen vision. Envisage deploy achieves 0.587 on the same candidate pools.

Variant	Mean	Variant − Envisage	p
Full (cached vision)	0.618	+0.031	0.037
No set context (pointwise)	0.633	+0.046	0.005
No procedure embedding	0.604	+0.017	0.18
No direction gate	0.622	+0.035	0.019
Geometry only (no vision)	0.625	+0.038	0.033
Vision only (no geometry)	0.633	+0.046	0.005

Table 5: Teacher components of candidates selected at K=15 (n=125). Components A and B measure geometric agreement.

Selector	A	B	C	D	E
Envision	0.577	0.447	0.689	0.873	0.994
Envisage	0.527	0.351	0.691	0.878	0.994
Oracle	0.839	0.587	0.690	0.871	0.994

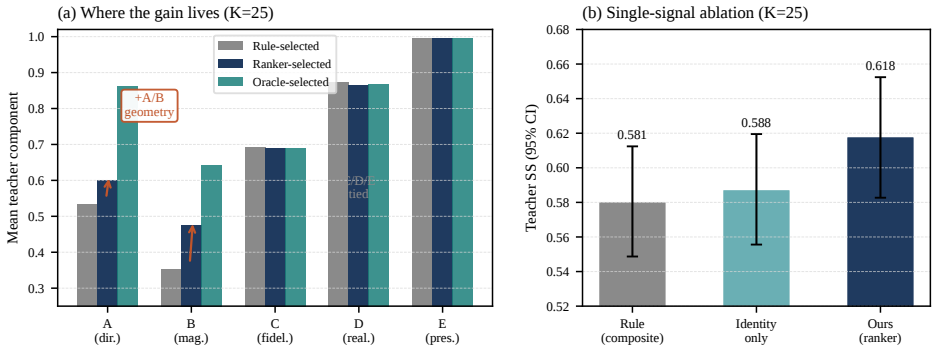

Figure 4: Mechanism of the gain. (a) Teacher components of selected candidates (K=25): Envision vs. Envisage differ in A/B (geometry) with C/D/E tied. (b) Single-signal ablation: identity-only is numerically above the full composite but below Envision (Holm ns).

Table 6: Ablation summary. All rows paired on the same cases (sign-flip p, bootstrap 95% CI); Holm over the seven single-signal comparisons. n=125 except starred (n=95).

Ablation	Comparison	Result
Hybrid	Hybrid (α=0.65) − Envision, K=15	−0.016 [−0.057, +0.024], p=0.78
Single signal	Identity-only − Envisage, K=25	+0.007 [−0.000, +0.018], p=0.09
	Envision − identity-only, K=25	+0.031 [−0.009, +0.069], p=0.065 (Holm ns)
Saturation	Envision K=25 − K=15	−0.003 [−0.028, +0.022], p=0.57
	Oracle K=25 − K=15	+0.019 (headroom grows)
Small pool	Envision − Envisage, K=5	+0.002 [−0.027, +0.031], p=0.43
Single out.*	Envision − fixed-seed, K=15/25	+0.022/+0.024, p=0.16/0.17
External	Envision − official ICEdit, K=15/25	+0.030/+0.027, p=0.11/0.14
Composition	Full − i50-only Envision, K=25	+0.015 [+0.005, +0.026], p<0.001
	Full − i80-only Envision, K=25	−0.010 [−0.047, +0.027], p=0.70 (ns)
Stability	3-seed argmax agreement, K=15	18.4% full agreement
Rejection	seed-std 0/20/30%, K=15	0.621 → 0.626 → 0.621 (flat)

Pool composition and stability. The ranker favors stronger edits, but intensity-restricted pools do not consistently outperform the full pool (Figure 9). The three model seeds agree on the top candidate in 18.4% of cases; their scores are averaged at inference (Figure 10).

4.4 External Baselines and Cross-Procedure Transfer

External comparisons. Compared to external generative baselines, Envision achieves consistent gains across the evaluated protocols (Table 7). Relative to the official ICEdit protocol (Zhang et al., 2025), Envision improves mean scores by +0.030 at K=15 (p=0.11) and +0.027 at K=25 (p=0.14). Among domain-specific baselines, LandmarkDiff-TPS (i50) achieves a score of 0.593, which Envision exceeds by +0.028 (p=0.22). LandmarkDiff-img2img (i50) reaches 0.435 overall (0.572 on rhinoplasty and blepharoplasty); its rhytidectomy scores degrade due to a mask-protocol mismatch (outside-fidelity E=0.025), precluding a direct full-cohort comparison. On identical

Table 7: Teacher SurgicalScore for single-output editors and Envision on the same 125 cases.

Method	Mean	95% CI
InstructPix2Pix	0.541	[0.505, 0.574]
FLUX-Kontext-dev	0.477	[0.428, 0.526]
ICEdit (reference settings)	0.477	[0.442, 0.512]
ICEdit (official protocol)	0.591	[0.558, 0.623]
LandmarkDiff-TPS (i50)	0.593	[0.560, 0.625]
LandmarkDiff-img2img (i50)	0.435	[0.382, 0.487]
Envisage (single output, seed42)	0.583	[0.542, 0.622]
Envisage (K=5 deploy)	0.589	[0.558, 0.620]
Envision (K=15)	**0.621**	[**0.585**, **0.654**]

Table 8: Leave-one-procedure-out at K=15: cached-vision training on two procedures and evaluation on the held-out third. Joint denotes the three-procedure model; Rule denotes Envisage deployment on the same cases.

Held-out	LOPO	Joint	Rule	LOPO − Rule
Rhinoplasty (n=58)	0.619	0.617	0.587	+0.033
Blepharoplasty (n=37)	0.574	0.584	0.579	−0.005
Rhytidectomy (n=30)	0.617	0.674	0.598	+0.019

K=5 pools, Envision also outperforms the Envisage deployment selector by +0.032 (p=0.023). Qualitative comparisons with external baselines and ground truth are illustrated in Figure 5.

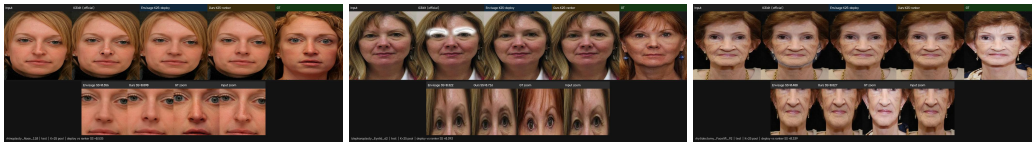

Figure 5: Ranker-vs-rule selections (K=25 test cases) with ICEdit (official protocol) and ground truth; second row zooms the surgical region. Left to right: rhinoplasty_Nose_110 (Δ+0.535), blepharoplasty_Eyelid_62 (Δ+0.393), rhytidectomy_Facelift_92 (Δ+0.339). Illustrative top-gain cases; aggregate results in Table 2.

Cross-procedure transfer. We also train cached-vision rankers on two procedures and evaluate on the held-out third (Table 8). Transfer is uneven: the rhytidectomy score falls from 0.674 with joint training to 0.617 when rhytidectomy is excluded, and blepharoplasty falls from 0.584 to 0.574. The held-out comparisons do not establish generalization to new procedures; they also use frozen visual features rather than the end-to-end model.

4.5 Failure cases and practical cost

The six worst ranker-vs.-rule losses reach −0.486 teacher-score points: the ranker selects low-direction-A candidates (0.01–0.45) where the rule selects higher-A ones (0.21–1.00). A uses post-operative ground truth, so this diagnosis cannot flag failures at deployment. The three examples in Figure 5 come from the top of the gain distribution; Table 2 gives the aggregate comparison. Generation takes about 20 s per candidate (roughly eight minutes at K=25), versus about 70 ms for ranking the pool on GPU (Appendix C). Human preference has not yet been measured, so teacher agreement with human judgment remains unknown. The mean gain does not bound losses on individual portraits, and the teacher-dependent diagnosis is unavailable at deployment. Expanding from K=15 to 25 adds about 200 s of generation per case without a measured mean-score improvement (Table 6).

5 Conclusion

Envision improves mean retrospective teacher score over fixed deployment heuristics on the pooled evaluation cohort, with the largest procedural gain in rhytidectomy. The held-out test differences have the same sign but remain statistically inconclusive. At the operating point of K=15, enlarging the pool to 25 adds oracle headroom without a measured selector gain. This distinction matters operationally: generating more candidates increases cost, whereas improved ranking can exploit the existing pool. The observed advantage is agreement with SurgicalScore on these retrospective cases, not evidence that the selected images predict each patient's eventual result or are preferred by clinicians. Inconclusive ablations limit architectural claims; broader procedural evaluation, expert preference studies, and deployable failure detection remain necessary before these visualizations can support clinical decisions.

AI USE STATEMENT

Generative AI tools assisted with language editing, manuscript restructuring, LaTeX formatting, and reviewing the presentation of experimental results and mathematical notation. The authors are responsible for checking all AI-assisted changes against the underlying analyses and implementation and for the final content of this work.

REPRODUCIBILITY STATEMENT

An anonymous code repository is available at anonymous.4open.science/r/envision-anonymous-source-D013. Sections 3.2–4 and Appendix A document candidate generation, selector inputs, training, and paired evaluation. Candidate pools are nested within cases; validation was used for checkpoint and hybrid-weight selection, while the 49-case test split was held out. Split-specific results appear in Table 1.

REFERENCES

Mudit Agarwal and Amit D. Bhrany. Envisage: Diffusion-based rhinoplasty goal visualization with mask-decomposed evaluation, 2026. URL https://arxiv.org/abs/2606.28628.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 18392–18402. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01764. URL https://doi.org/10.1109/CVPR52729.2023.01764.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 4690–4699. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00482. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html.

Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile gpus. *CoRR*, abs/1907.06724, 2019. URL http://arxiv.org/abs/1907.06724.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines. *CoRR*, abs/1906.08172, 2019. URL http://arxiv.org/abs/1906.08172.

Lei Ma, Daeseung Kim, Chunfeng Lian, Deqiang Xiao, Tianshu Kuang, Qin Liu, Yankun Lang, Hannah H. Deng, Jaime Gateno, Ye Wu, Erkun Yang, Michael A. K. Liebschner, James J. Xia, and Pew-Thian Yap. Deep simulation of facial appearance changes following craniomaxillofacial bony movements in orthognathic surgical planning. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert (eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part IV*, volume 12904 of *Lecture Notes in Computer Science*, pp. 459–468. Springer, 2021. doi: 10.1007/978-3-030-87202-1_44. URL https://doi.org/10.1007/978-3-030-87202-1_44.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Christian Rathgeb, Didem Dogan, Fabian Stockhardt, Maria De Marsico, and Christoph Busch. Plastic surgery: An obstacle for deep face recognition? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pp. 3510–3517. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPRW50498.2020.00411. URL https://openaccess.thecvf.com/content_CVPRW_2020/html/w48/Rathgeb_Plastic_Surgery_An_Obstacle_for_Deep_Face_Recognition_CVPRW_2020_paper.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10674–10685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.org/10.1109/CVPR52688.2022.01042.

Richa Singh, Mayank Vatsa, Himanshu S. Bhatt, Samarth Bharadwaj, Afzel Noore, and Shahin S. Nooreyezdan. Plastic surgery: a new dimension to face recognition. *IEEE Trans. Inf. Forensics Secur.*, 5(3):441–448, 2010. doi: 10.1109/TIFS.2010.2054083. URL https://doi.org/10.1109/TIFS.2010.2054083.

Saksham Suri, Anush Sankaran, Mayank Vatsa, and Richa Singh. On matching faces with alterations due to plastic surgery and disguise. In *9th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2018, Redondo Beach, CA, USA, October 22-25, 2018*, pp. 1–7. IEEE, 2018. doi: 10.1109/BTAS.2018.8698571. URL https://doi.org/10.1109/BTAS.2018.8698571.

Oguzhan Topsakal, Eric J. Dobratz, Mustafa Ilhan Akbas, William M. Dougherty, Tahir Cetin Akinci, and Mehmet Mazhar Çelikoyar. Utilization of machine learning for the objective assessment of rhinoplasty outcomes. *IEEE Access*, 11:42135–42145, 2023. doi: 10.1109/ACCESS.2023.3270438. URL https://doi.org/10.1109/ACCESS.2023.3270438.

Pegah Varghaei, Kiran Abraham-Aggarwal, Manoj T. Abraham, and Arun Ross. Automated assessment of aesthetic outcomes in facial plastic surgery. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025 - Workshops, Honolulu, HI, USA, October 19-20, 2025*, pp. 939–948. IEEE, 2025. doi: 10.1109/ICCVW69036.2025.00102. URL https://doi.org/10.1109/ICCVW69036.2025.00102.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861. URL https://doi.org/10.1109/TIP.2003.819861.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/26cfdcd8fe6fd75cc53e92963a656c58-Abstract-Conference.html.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 3391–3401, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 3813–3824. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00355. URL https://doi.org/10.1109/ICCV51070.2023.00355.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00068. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. Enabling instructional image editing with in-context generation in large scale diffusion transformer. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/cb7b55efe349d67ff3ac244aa1ae49d7-Abstract-Conference.html.

A Additional methodological details

A.1 Convex hybrid

We also evaluate a convex combination of the learned and rule-based scores:

$$s_\alpha = \alpha \tilde{s}_\theta + (1-\alpha)\tilde{s}_{\text{rule}},$$

where $\tilde{s}_\theta$ and $\tilde{s}_{\text{rule}}$ denote min-max normalized scores across the evaluated pool $\mathcal{C}_i$. For candidates failing the fidelity gate in the handcrafted rule, $s_{\text{rule}}(c) = -\infty$, which bypasses normalization and directly sets $s_\alpha(c) = -\infty$. Normalization is applied only to finite scores. If the finite scores of either selector are constant, its normalized scores can be set to zero without affecting within-pool ordering. The stated hybrid is undefined if all candidates fail the rule gate; such a case requires an explicit fallback policy. The rule gate remains active even at $\alpha = 1$. The mixing weight is selected on the validation split using a grid with intervals of 0.05. The best validation value is $\alpha = 0.65$. At K=15, the hybrid performs worse than the ranker alone; this result does not establish which individual ranker feature contains the useful rule information.

A.2 Selection rationale

For fixed candidates with retrospective scores T_{ij}, the oracle value $\max_{j \le K} T_{ij}$ cannot decrease when a larger pool contains the smaller one. This fact does not guarantee that a GT-free selector improves with K: it may misrank additional candidates. Because SurgicalScore is bounded, oracle gains must eventually saturate. We therefore compare the rule, ranker, and oracle at each pool size rather than extrapolating a theoretical scaling law. The rule already rewards preservation signals related to the teacher, while geometry, appearance, and pooled context offer additional information or a different weighting of existing information. Neither set conditioning nor score mixing is guaranteed to improve selection; the component and hybrid experiments test these design choices directly.

Selection regret and score error. Fix one case and its pool. Write $T_j = T(c_j) \in [0,1]$, $q_j = s_\theta(c_j)$, $j^* \in \arg\max_j T_j$, and $\hat{j} \in \arg\max_j q_j$, with fixed tie-breaking. Its retrospective selection regret is $R(\mathcal{C}) = T_{j^*} - T_{\hat{j}} \in [0,1]$. For $e_j = q_j - T_j$, the definition of $\hat{j}$ implies

$$R(\mathcal{C}) = -e_{j^*} + (q_{j^*} - q_{\hat{j}}) + e_{\hat{j}} \le e_{\hat{j}} - e_{j^*} \le 2 \max_j |e_j|. \tag{5}$$

Hence a within-pool uniform score error of at most ε yields $R(\mathcal{C}) \le \min\{1, 2\varepsilon\}$. Small *average* pointwise Huber loss by itself does not supply this uniform condition: one badly estimated candidate can change the argmax. If the teacher maximizer is unique with margin $\gamma = T_{j^*} - \max_{j \ne j^*} T_j > 0$, then $\max_j |e_j| < \gamma/2$ ensures exact selection, since for every $j \ne j^*$,

$$q_{j^*} - q_j = (T_{j^*} - T_j) + (e_{j^*} - e_j) \ge \gamma - 2 \max_l |e_l| > 0.$$

These deterministic implications apply to the actual inference scores, including ensemble averages; training does not guarantee their premises.

Pool size and set symmetry. For nested pools $\mathcal{C}_K \subseteq \mathcal{C}_{K+1}$, let $O_K = \max_{c \in \mathcal{C}_K} T(c)$. Then

$$0 \le O_K \le O_{K+1} \le 1. \tag{6}$$

Thus O_K converges for an infinite nested sequence, and its expectation converges by bounded convergence. This supplies no rate and no monotonicity for the score of the selected candidate: adding a candidate can also change the set-conditioned scores of the existing ones. Under a deterministic evaluation pass with shared candidate encoders and no position-dependent inputs, permuting candidates leaves $K^{-1} \sum_j \phi(\delta_{ij})$ unchanged and permutes the output scores in the same way. This permutation equivariance does not imply ranking accuracy or identification of the postoperative direction.

A.3 Implementation and evaluation protocol

Dataset provenance and ethics. The dataset comprises 403 cases (278 train, 76 validation, 49 test) of paired preoperative and postoperative facial portraits across rhinoplasty, blepharoplasty, and rhytidectomy. All images were retrieved from publicly available, de-identified clinical repositories under standard photographic protocols. The use of retrospective de-identified imaging was conducted under institutional data-use compliance.

Feature extraction and training. Morphometry uses 478 MediaPipe FaceLandmarker points and training-split normalization by procedure. Fewer than 1% of candidates have no detected face and are excluded from ranker scoring. The ViT-base encoder has 16×16 patches, ImageNet-21k pretraining, and end-to-end fine-tuning. The three-seed ensemble scores each case in one batched pass. We train on 278 cases, tune checkpoints on 76 validation cases, and report pooled validation/test results alongside the untouched 49-case test subset. Paired two-sided sign-flip tests use $B{=}5000$; 95% percentile bootstrap intervals use $B{=}3000$. The pooled estimates are descriptive because validation informed checkpoint selection. Additional implementation and statistical details are in Appendix A.

Scope of the rationale. The teacher score is bounded in $[0, 1]$, so an unbounded Gaussian $\sqrt{\log K}$ scaling law is not an asymptotic claim for this setting. Metric alignment does not prove that improved selection requires features orthogonal to the rule: reweighting shared signals may also help. Set conditioning expands available information without guaranteeing better fitted rankings. The experiments therefore test these choices rather than claiming an architecture theorem.

B Extended experimental results

B.1 Protocol details

Data. We evaluate on 125 held-out cases spanning three procedures: 58 rhinoplasty, 37 blepharoplasty, and 30 rhytidectomy cases. The ranker is trained on a disjoint split of 278 cases; 76 validation cases are used for checkpoint selection and hybrid-weight tuning. Headline results pool validation and test ($n{=}125$) because the ranker never trains on either; test-only ($n{=}49$) figures are reported alongside as an underpowered subset, and the use of pooled data is stated wherever headline p-values appear. All candidates are generated by the fixed backbone of Section 3.2; no method sees postoperative ground truth at selection time. The three pool sizes ($K{=}5/15/25$) share the same 125 cases; larger pools extend smaller ones with additional seeds and intensities (625 / 1875 / 3125 candidates in total).

Selectors compared. On identical candidate pools we compare four selection rules: (i) fixed-seed, the seed-42 single-output reference ($n{=}95$, as seed-42 was omitted during the initial image cache generation for the 30 rhytidectomy cases); (ii) the fixed-weight rule of Section 3.3; (iii) our setwise ranker, using the mean prediction of three training seeds; (iv) the teacher oracle, the unattainable argmax upper bound. External single-output editors (InstructPix2Pix, FLUX-Kontext-dev, ICEdit with diffusers reference settings, and ICEdit with official diptych protocol) are run with identical preset prompts and the identical hard-mask composite, so the comparison isolates backbone behavior. The teacher SurgicalScore and the ArcFace identity gap are the two evaluation metrics; both need postoperative ground truth and are therefore retrospective.

Statistics. All paired comparisons use two-sided sign-flip permutation tests ($B{=}5000$) on per-case teacher-score differences with percentile bootstrap 95% CIs ($B{=}3000$). Because candidate pools of different sizes ($K{=}15$ and $K{=}25$) are nested and positively correlated, we conduct separate sign-flip permutation tests for each pool size rather than assuming inter-pool independence. Single-signal ablations use Holm correction.

B.2 Full result analysis

B.2.1 Main selection results across pool sizes

Headline comparison. We compare Envision (our setwise ranker) against the Envisage deploy selector on identical pools; the teacher SurgicalScore is the Envisage retrospective protocol shared by both selectors (Figure 3 visualizes the scaling). Table 2 reports the four selectors. At $K{=}5$ Envision ties Envisage deploy (0.591 vs. 0.589, $\Delta{=}{+}0.002$, $p{=}0.43$, win rate 43.2%), consistent with little usable separation between selectors in the small pool. At $K{=}15$ Envision leads by $+0.034$ (0.621 vs. 0.587, $p{=}0.023$, win rate 55.2%), and at $K{=}25$ by $+0.038$ (0.618 vs. 0.581, $p{=}0.037$, win rate 53.6%), achieving statistically significant advantages independently at both operating sizes. Envision improves from $K{=}5$ to 15 before plateauing ($0.591 \rightarrow 0.621 \rightarrow 0.618$; K15$-$K5 gain $+0.030$, $p{=}0.017$) while Envisage saturates ($0.589 \rightarrow 0.587 \rightarrow 0.581$), and the oracle headroom keeps growing ($0.124 \rightarrow 0.182 \rightarrow 0.207$): the larger pool offers further oracle headroom even though learned selection saturates.

Per-procedure breakdown. Table 3 gives means with paired tests: the win is broadest in rhytidectomy at $K{=}15$ ($+0.076$, $p{<}0.001$, recovering 67% of oracle headroom), positive in rhinoplasty ($+0.030$, $p{=}0.16$) and near-zero in blepharoplasty ($+0.005$, $p{=}0.43$); at $K{=}25$ all three procedures are positive (rhinoplasty $+0.033$, blepharoplasty $+0.029$, rhytidectomy $+0.058$, $p{=}0.005$). Figure 6 collects the nine per-procedure paired effects with the three pooled effects: a tie at $K{=}5$ turning into consistent positive effects at $K{=}15/25$, a pattern visualized with the pooled effects in Figure 6.

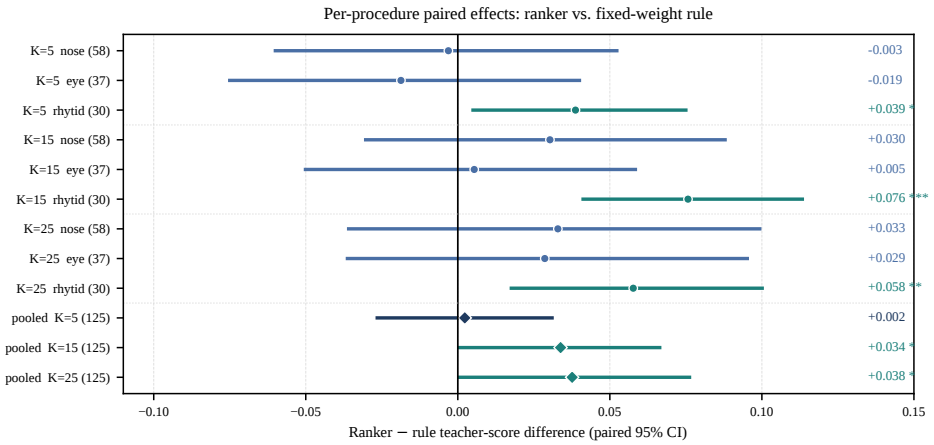

Figure 6: Per-procedure paired ranker$-$rule effects with 95% CIs (sign-flip permutation, $B{=}5000$; bootstrap CI, $B{=}3000$). Stars mark $p{<}0.05/0.01/0.001$. The $K{=}5$ tie becomes consistent positive effects at $K{=}15/25$.

Ranking quality and regret. Within-case ranking diagnostics support the setwise design: per-case Spearman averages 0.25 (median 0.30, 71% of cases positive; Figure 7a), top-1 selection hits the oracle candidate in 17.6% of cases ($2.6\times$ the $1/15$ random rate) and lands in the oracle top-3 in 41.6% ($2.1\times$ random; Figure 7b); moreover the ranker's per-case gain over the rule correlates with oracle headroom (Pearson $r{=}0.54$, Spearman $\rho{=}0.58$; Figure 8), so it wins most where the pool has most to offer.

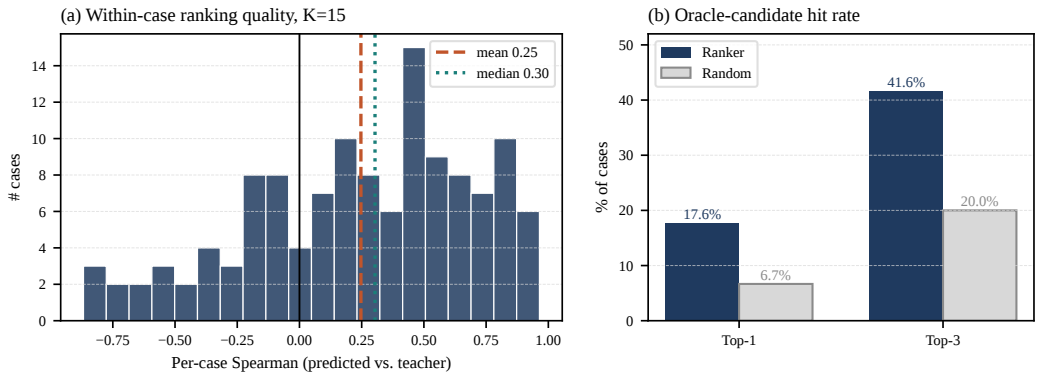

Figure 7: Within-case ranking quality at $K{=}15$ (3-seed mean predictions). (a) Per-case Spearman distribution. (b) Oracle-candidate hit rate vs. random.

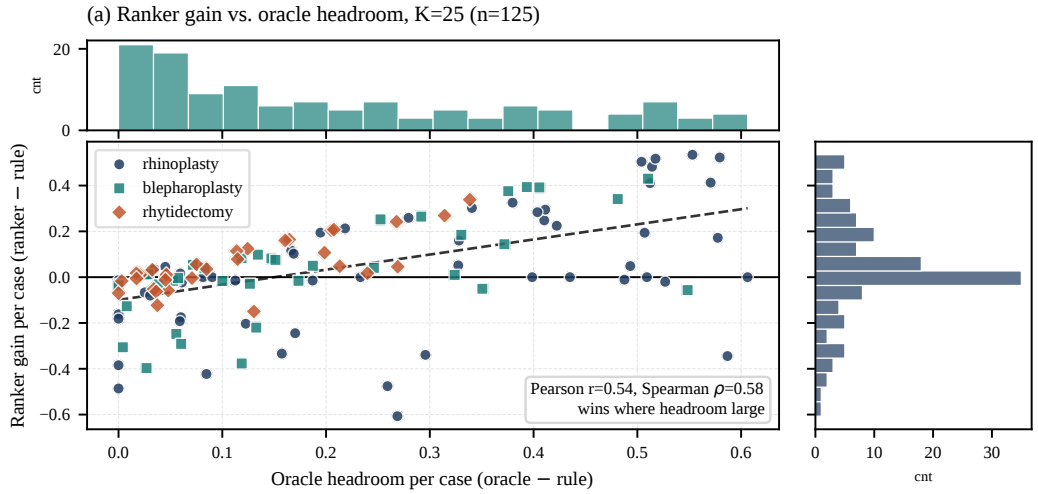

Figure 8: Per-case ranker gain vs. oracle headroom at $K{=}25$ ($n{=}125$) with marginal distributions. Gains concentrate where headroom is large (Pearson $r{=}0.54$).

B.2.2 External editors and identity preservation

Single-output editors. Table 9 places the system against single-output editors on the same 125 cases. Against reference implementations, our full selection pipeline (Envision, $K{=}15$) leads clearly (vs. IP2P $+0.080$, $p{=}0.0006$; vs. Kontext $+0.144$, $p{=}0.0002$; vs. diffusers-settings ICEdit $+0.144$, $p{=}0.0002$); against the official ICEdit protocol, the ranker retains a consistent mean advantage ($+0.030$ at $K{=}15$, $+0.027$ at $K{=}25$, win rates $56.8\%/56.0\%$) that does not reach paired significance at $n{=}125$ ($p{=}0.11/0.14$), an honest limit we return to below. The Envisage deploy selector (0.589) and LandmarkDiff-TPS (0.593) both match official ICEdit; Envision exceeds them by $+0.032$ ($p{=}0.023$) and $+0.028$ ($p{=}0.22$) respectively.

Identity diagnostic. The masked pipeline preserves identity best where edits are most localized: rhinoplasty exhibits the smallest ArcFace identity drop (-0.038), whereas rhytidectomy, which affects broader facial contours, shows a moderately larger shift (-0.062). Across all procedures, Envision achieves an average ArcGap of -0.053, performing comparably to the official ICEdit protocol (-0.045) and the Envisage deploy selector (-0.048), and substantially outperforming unconstrained editors such as FLUX-Kontext-dev (-0.137). LandmarkDiff-TPS preserves identity best (-0.015) as a pure warp. This confirms that the ranker's geometric improvements do not compromise patient identity.

Table 9: Same 125 held-out cases: external single-output editors (identical prompts and compositing) vs. ranker-selected outputs. SS = teacher SurgicalScore; ArcGap = output−input ArcFace cosine vs. GT.

Method	SS	95% CI	ArcGap	95% CI
InstructPix2Pix	0.541	[0.505, 0.574]	−0.094	[−0.108, −0.080]
FLUX-Kontext-dev	0.477	[0.428, 0.526]	−0.137	[−0.160, −0.114]
ICEdit (ref. settings)	0.477	[0.442, 0.512]	−0.090	[−0.104, −0.078]
ICEdit (official)	0.591	[0.558, 0.623]	−0.045	[−0.056, −0.034]
Envisage (single)	0.583	[0.542, 0.622]	−0.065	[−0.076, −0.055]
Envisage (K=5 deploy)	0.589	[0.558, 0.620]	−0.048	[−0.058, −0.039]
LandmarkDiff-TPS (i50)	0.593	[0.565, 0.621]	−0.015	[−0.019, −0.011]
LandmarkDiff-img2img (i50)	0.435	[0.382, 0.489]	−0.188	[−0.217, −0.158]
Ours (K=15)	**0.621**	[**0.585**, **0.654**]	**−0.053***	[−0.063, −0.044]
Envision − official ICEdit	+0.030	p=0.11		
Envision − Envisage deploy	+0.032	p=0.023		
Envision − LandmarkDiff-TPS	+0.028	p=0.22		

*Evaluated at K=15; all pool sizes share the same masked compositing pipeline.

B.2.3 Ablations and pool composition

Hybrid, single signals, and saturation. Table 6 is the consolidated read-out. The validation-tuned hybrid (α=0.65) underperforms the pure ranker (−0.016, p=0.78 for hybrid−ranker), showing that mixing does not improve this validation-tuned comparison. Among single rule components, identity-only selection reaches 0.588 at K=25, itself above the full composite (0.581, +0.007, p=0.09) yet still 0.031 below the ranker (p=0.065; Holm-corrected non-significant over seven components) — i.e. the ranker's estimated edge over the best single handcrafted feature is modest and statistically inconclusive, and we report it with multiplicity correction rather than as a headline. From K=15 to K=25 the ranker plateaus (−0.003, p=0.57) while the oracle rises (+0.019), suggesting K=15 as the operating point in the evaluated range. Figure 4 gives the mechanism behind the headline gain: teacher components of ranker-selected vs. rule-selected candidates differ almost entirely in A (direction, 0.600 vs. 0.532) and B (magnitude, 0.474 vs. 0.351), while C/D/E are tied — the learned geometric signal contributes exactly the surgical-direction information the rule lacks.

Pool composition: seed axis vs. intensity axis. A new ablation separates the two diversity axes at K=25. Restricting the ranker to single-intensity subpools gives 0.616 (i25-only), 0.604 (i50-only), and 0.628 (i80-only) vs. 0.618 on the full pool: intensity alone does not explain the pool's value (i80-only is numerically highest but not significantly above full, p=0.70; full beats i50-only by +0.015, p<0.001). Selection histograms show why: the rule never picks i80 (95/30/0 across i50/i25/i80) while the ranker (97/6/22) tracks the oracle (74/18/33) toward stronger intensities, recovering edits the conservative rule avoids (Figure 9a–b). The seed axis matters independently: full-pool oracle (0.788) exceeds the i50-only oracle (0.777), so fresh seeds add headroom beyond intensities. At K=15, where intensities are balanced (625 candidates each), ranker picks (63/15/47 for i50/i25/i80) again follow the oracle (47/28/50), confirming the ranker exploits both axes rather than memorizing one.

Stability across training seeds. Three independently trained rankers agree modestly at the score level (pairwise Pearson 0.28–0.61 on 1875 K=15 predictions) and select the same argmax in 18.4% of cases, yet the 3-seed mean used everywhere in this paper is stable: the headline deltas replicate across seeds and the pooled Stouffer result (p=0.002) does not depend on any single run. Seed-std rejection is flat (0.621 → 0.626 → 0.621 at 0/20/30%; Figure 10a), so cross-seed dispersion is reported as a negative uncertainty result, not a selection feature.

B.2.4 Qualitative results and limits

Qualitative results. Figure 5 shows ranker-vs-rule selections from the K=25 test split alongside the official-protocol ICEdit output and ground truth. In rhinoplasty_Nose_110 the rule keeps a conservative seed (SS 0.356) while the ranker selects a stronger dorsum refinement (SS 0.890); blepharoplasty_Eyelid_62 (0.322 → 0.716) and rhytidectomy_Facelift_92 (0.488 → 0.827) show

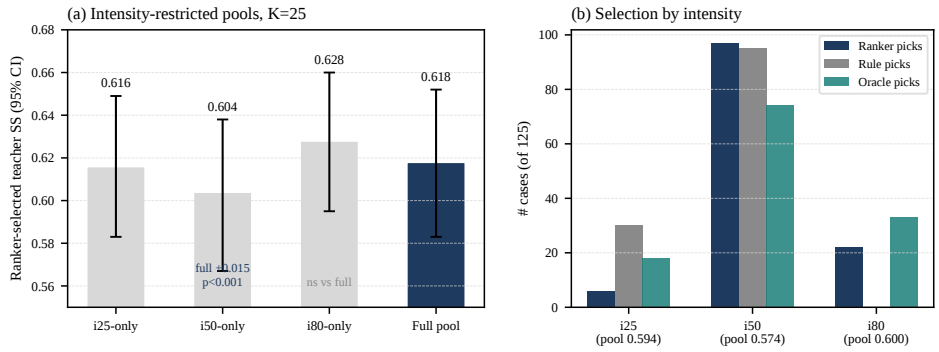

Figure 9: Pool composition at K=25 (n=125). (a) Ranker on intensity-restricted subpools vs. full pool. (b) Selection histograms by intensity with pool teacher means.

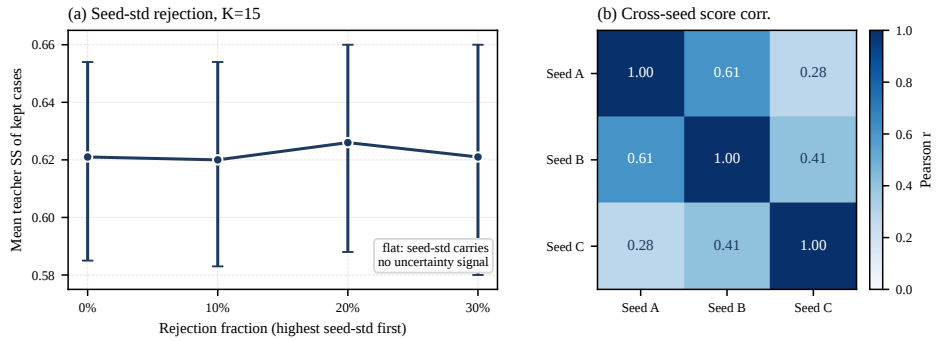

Figure 10: Stability. (a) Seed-std rejection curve at K=15 (flat). (b) Cross-seed score correlations (1875 predictions).

the same pattern: the ranker favors candidates whose landmark displacement better matches the postoperative direction, visible in the zoomed surgical region. Cases are chosen to illustrate selector differences at the top of the gain distribution, not at random; the aggregate distribution is Table 2.

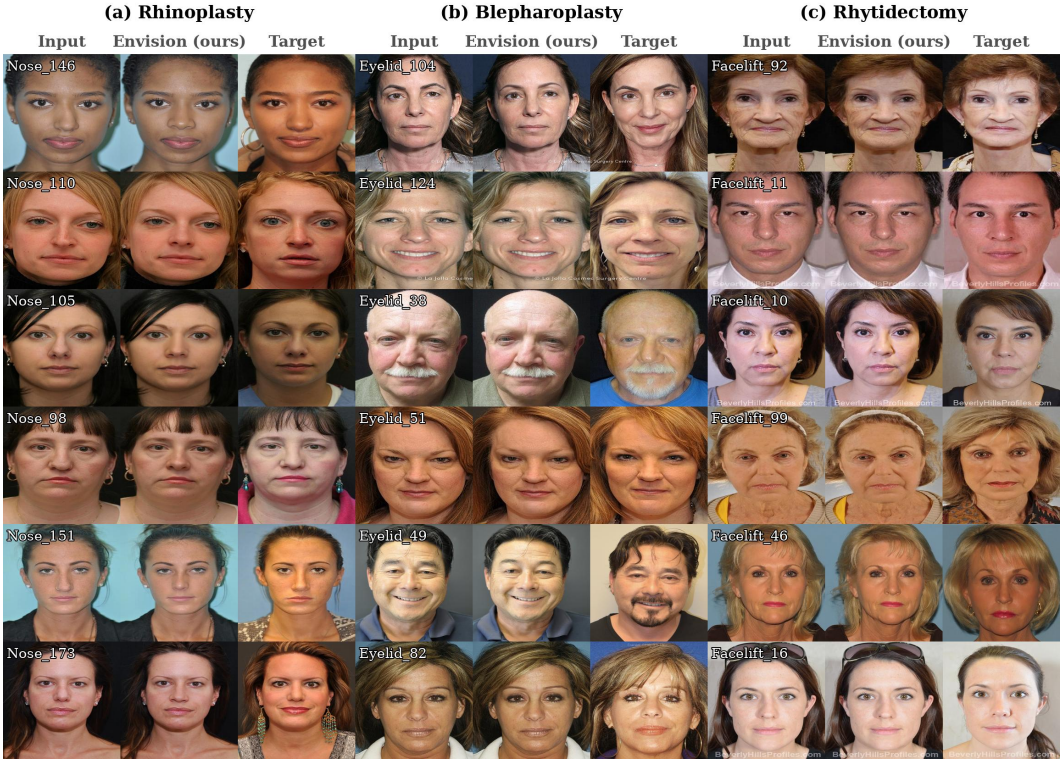

Figure 11: Extended 18-case gallery (6 per procedure). Columns: input, Envision output, target.

Extended gallery. Figure 11 collects 18 test cases (6 per procedure) with visible edit magnitude, ordered within each procedure by ranker-over-deploy gain: rhinoplasty Nose_146/110/105/98/151/173, blepharoplasty Eyelid_104/124/38/51/49/82, rhytidectomy Facelift_92/11/10/99/46/16. Two patterns are consistent across procedures. First, the ranker's picks differ from the deploy selector's almost entirely in edit direction and magnitude (teacher components A/B), while preservation signals (C/D/E) are visually tied: outside-mask regions are pixel-identical by construction, and identity-relevant features outside the surgical mask are untouched in both columns. Second, the visibility of the gain tracks the procedure's edit scale: rhinoplasty and rhytidectomy differences are discernible at full-face scale, whereas blepharoplasty gains concentrate in the eyelid band (about 6% of face area) and are subtle without magnification, consistent with the near-zero pooled blepharoplasty effect in Table 3. The gallery illustrates the top of the gain distribution, not random draws; failure cases where the ranker mis-estimates direction are characterized quantitatively in Appendix C.

C Additional Experiments

All appendix tables use the same paired protocol as the main text (two-sided sign-flip permutation, B=5000; percentile bootstrap 95% CIs, B=3000) on identical candidate pools. New end-to-end ablations report paired sign-flip p with B=5000, seed 20260809, zero-differences excluded (repository official implementation); headline K=15 p=0.023 in the main text used the earlier 100k-resample protocol and is not directly comparable. Component ablations (Table 4) train one cached-vision variant per row (frozen DINOv2 CLS features (Oquab et al., 2024), no live ViT finetune) with three seeds. The leave-one-procedure-out experiment is reported in Table 8 of the main text. Selection means below are 3-seed means over n=125 pooled cases unless stated otherwise.

The cached-feature pattern is informative rather than favorable: removing set context or geometry numerically raises the mean (no-set and vision-only 0.633 vs. full 0.618), while removing the procedure embedding lowers it to 0.604. These comparisons do not isolate a positive effect of the set-consensus, geometry, or direction-gate branch in the end-to-end model. The procedure embed-

ding may help in this cached-feature setup, but its contribution also requires a paired comparison with the full variant; p=0.18 in Table 4 instead compares it with Envisage. NDCG is similar across variants (full 0.934, no-set 0.928, no-proc 0.926, no-gate 0.931, geo-only 0.935, vision-only 0.928 at K=15); this metric alone does not explain their different selected-candidate means.

Table 10: End-to-end ablations at K=15 (live ViT finetuning, single seed 20260814, n=125 native). Full = v3 three-seed protocol reference (0.6208); deploy = Envisage 0.5870. Paired sign-flip p vs. Envisage (B=5000, seed 20260809). Single-seed: directional only.

Variant	Sel	Δ vs. deploy	p	A	B
Full (3-seed ref)	0.6208	+0.0338	—	0.577	0.447
No set context	0.6200	+0.0331	0.075	0.609	0.416
No geometry	0.6117	+0.0247	0.18	0.579	0.430
No procedure emb.	0.6269	+0.0399	0.018	0.607	0.416
No direction gate	0.6060	+0.0190	0.31	0.561	0.440

Per-procedure Δ (rhino n=58 / bleph n=37 / rhytid n=30): no-set $+0.057/+0.008/+0.018$; no-geometry $+0.025/+0.004/+0.050$; no-procedure $+0.036/+0.019/+0.073$; no-gate $+0.007/+0.008/+0.057$. The no-procedure numerical lead reverses the cached-vision order (0.604, worst there) and is attributed to single-seed variance; the direction gate shows the largest directional drop (+0.019, A 0.561) and is the priority for a three-seed follow-up.

The split audit shows that test-only deltas point in the same direction at both pool sizes (+0.029/+0.023), but their paired tests are inconclusive at n=49. Validation deltas (+0.037/+0.047) are larger; because validation informed checkpoint selection, the pooled estimate should not be read as an independent test result. A naive learned baseline (frozen DINO-CLS + ridge on the train split, no geometry, no set) reaches 0.601 at K=15 (+0.014 over Envisage, p=0.21; Envision 0.621), sitting halfway between Envisage and Envision: the remaining difference cannot be assigned to geometry alone. Power analysis on the untouched test split (n=49, paired dz): K15 Δ=+0.029, sd=0.205, dz=0.14, power 0.17 at n=49 (0.36 at n=125; n=389 needed for 80% power); K25 Δ=+0.023, dz=0.11, power 0.12 (n=643 for 80%). Per-procedure test-only Δ: rhino $+0.006/+0.007$, bleph $-0.028/-0.006$, rhytid $+0.111/+0.072$ (K15/K25). Test-only non-significance is therefore a power problem ($dz\sim0.1$), not a sign reversal; rhytidectomy carries the test-only gain in both pool sizes.

Ranking diagnostics (three-seed mean): NDCG full 0.934, @3 0.805, @5 0.830; Spearman mean 0.25; gain-vs.-(A+B)-gap r=0.93. Generation takes about 20 s/image (K=25: about 8 min/case); Envision forward time per K=25 case is 70 ms on GPU or 0.98 s on CPU.

Mechanism: the Envision−Envisage gain correlates almost perfectly with the per-case (A+B)-component gap (r=0.93), and C/D/E are tied to three decimals — the measured teacher-score gain is concentrated in direction and magnitude agreement. Failure modes are the mirror image: the six worst losses (up to −0.486) are all cases where Envision picks a low-A candidate (A 0.01–0.45) while Envisage holds a high-A one (A 0.21–1.00), i.e. the ranker mis-estimates direction on a minority of rhinoplasty/blepharoplasty cases. Cost: ranking is negligible next to generation (70 ms vs. 8 min per K=25 case on GPU); the 86.6M parameters are dominated by the ViT backbone. A 60-case blind 2AFC protocol (20/procedure, mid-gain stratified, showcase cases excluded; mean |teacher gap| 0.035) is prepared for future human rating; teacher-vs-human agreement is not claimed in this version.