RETARGETINGFACE: BRIDGING HETEROGENEOUS FACIAL-CONTROL SPACES WITH A SHARED EXPRESSION REPRESENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Facial animation systems encode the same expressions in incompatible native control spaces. Pairwise vector regression can bridge these spaces, but requires a separate mapping for each source–target direction. We therefore study cross-system facial retargeting as *native-control interoperability*: preserving the expression being conveyed while retaining each system's original control parameterization. Based on this perspective, we introduce **RetargetingFace**, a shared expression interface that connects different facial-control systems using condition-level correspondence rather than exact cross-rig coefficient pairs. RetargetingFace extracts neutral-relative visual evidence from expressive and neutral renderings, reducing dependence on source facial shape while preserving expression-induced changes. Expression supervision, conditional rig confusion, and cross-rig alignment organize these observations into a shared representation. Each rig-specific predictor trains on its own latent–control pairs and is reused across source rigs. We construct a controlled GNM-derived dataset with 532 expression states, strength and FLAME-shape variation, and condition-matched native FLAME/ARKit realizations. On held-out FLAME↔ARKit transfers, RetargetingFace improves mean human similarity by **18.7%** over the best evaluated existing baseline; the mean gain on external captured expressions is **9.9%**. With the representation frozen, independently trained GNM and MHR predictors extend the interface to new target and source rigs. On the reported FLAME/GNM/MHR transfers, human ratings and first-hop cosine favor our method over direction-specific Direct MLP.

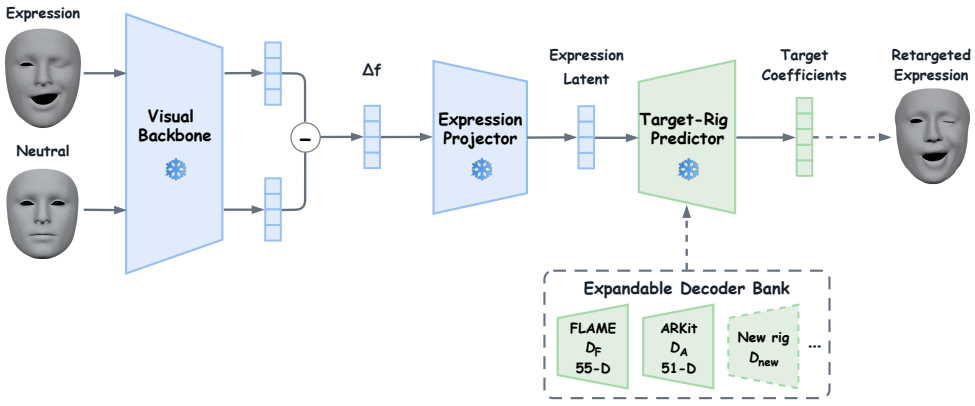

Figure 1: **RetargetingFace at inference.** A source expression and its matched neutral are encoded into a shared expression latent, which the selected target-rig predictor maps to native controls. The decoder bank accommodates additional target systems through separately trained predictors. All modules are frozen at inference.

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.

1 INTRODUCTION

Facial performances must move between animation systems while retaining native editability. FLAME, ARKit, GNM, and MHR encode expressions with incompatible controls, making this transfer difficult.

Native coefficients specify how a rig realizes an expression, not the expression itself. The same mouth opening, for example, may use jaw pose and expression coefficients in FLAME but several blendshapes in ARKit. Direct vector regression can learn this relationship for one ordered pair; extending it to another source or target requires a new mapping. The central challenge is to train native-control predictors once and reuse them across sources.

Semantic representations (Josi et al., 2025), mesh-transfer controls (Qin et al., 2023), and multi-head learning from heterogeneous labels (Fan et al., 2024) address related aspects, but not reuse of predictors trained solely within each existing native-control system (Sec. 2).

We therefore separate expression encoding from native-control learning. **RetargetingFace** learns a shared latent from expressive renders relative to their matched neutral, reducing source-shape information while retaining facial actions and strengths. Shared action–strength conditions align the representation without exact cross-rig coefficient pairs. Each rig then learns its own latent-to-control predictor from local data; the target predictor accepts latents from other supported sources without direction-specific retraining.

Our Contributions. ***(1) New Perspective.*** We study cross-system facial retargeting as native-control interoperability: independently designed facial systems communicate through a shared expression interface while retaining their original control parameterizations. ***(2) New Method.*** We learn a neutral-referenced shared expression representation using condition-level correspondence. With this representation fixed, each native predictor trains by same-rig reconstruction and is reused across sources; GNM and MHR follow the same design beyond FLAME and ARKit. ***(3) New Dataset.*** We construct a GNM-derived cross-rig dataset varying action composition, strength, source instance, and FLAME shape, with condition-matched native FLAME/ARKit realizations. ***(4) Fidelity and Reuse.*** On held-out FLAME↔ARKit transfers, RetargetingFace improves mean human similarity by **18.7%** over the best evaluated existing baseline per direction; external expressions and newly attached rigs further test predictor reuse.

2 RELATED WORK

2.1 FACIAL-CONTROL REPRESENTATIONS AND RETARGETING

Native rigs expose incompatible controls: statistical face models (Blanz & Vetter, 1999), FLAME's expression and pose (Li et al., 2017), ARKit blendshapes (Apple Inc., 2026a), and the parameterizations of GNM (Ploumpis et al., 2026) and MHR (Ferguson et al., 2025). Their deformation bases and neutral configurations prevent direct coefficient exchange.

Retargeting has used identity–expression factorization (Vlasic et al., 2005), neural transfer (Wu et al., 2018; Mo et al., 2022), and learned controls (Qin et al., 2023; Cha et al., 2025). Kim et al. (2021) share an image latent across rigs, then synthesize a target image before predicting its controls. Direct-control approaches instead use designated output vocabularies, such as ARKit blendshapes (Lee et al., 2025), region-specific Motion Units (Zhu et al., 2025), or FACS-pose weights (Anadon Leon & Bütepage, 2026).

Calibration and rig construction provide other routes: tracker refinement (Omens et al., 2026), local video-patch controls (Choi et al., 2025), and learned rigs across topologies (Ma et al., 2025; Mohammadi et al., 2026). Our setting retains existing native controls and asks whether a predictor learned within one rig can serve multiple sources.

2.2 LEARNED FACIAL EXPRESSION REPRESENTATIONS

Neural Face Rigging (Qin et al., 2023) and Neural Face Skinning (Cha et al., 2025) transfer controls or deformations across meshes. SEREP (Josi et al., 2025) learns expression semantics from unpaired meshes with common topology and decodes target-identity-conditioned meshes. UniTalker (Fan

et al., 2024) uses a shared backbone and multiple heads for heterogeneous audio-driven animation labels.

RetargetingFace uses shared action–strength conditions to train a neutral-referenced interface, then trains each native-control predictor only on its own latent–coefficient pairs. This decoupling permits cross-source predictor reuse without direction-specific retraining.

3 Controlled Cross-Rig Expression Dataset

Why a Controlled Cross-Rig Dataset? Learning a shared expression interface requires observing comparable actions across native control systems while varying expression strength and face shape. We construct a GNM-derived (Ploumpis et al., 2026) dataset that exposes these factors jointly, complementing captured-expression datasets with condition-matched FLAME and ARKit realizations.

Expression Coverage and Native Realization. The atlas combines 19 base actions and three mixtures for each of 171 action pairs into 532 states, each with 50 source variants. SMIRK (Retsinas et al., 2024) and MediaPipe (Google AI Edge, 2026) independently extract 55-dimensional FLAME and 51-dimensional ARKit controls from each source-state image under one fixed source appearance. Native peaks are scaled over 11 strengths from neutral to full activation, then rendered on ten FLAME shapes or one ARKit rig with matched neutral references.

Each sample has condition $q = (k, s)$: action state k and relative strength s. This supplies cross-rig semantic relations without asserting coefficient-wise correspondence. Shape changes at fixed controls isolate identity variation; strength paths supply expression ordering. Source variants describe expression realizations, separately from rendering shape.

Controlled Evaluation. Train, Validation, and Test separate source variants, FLAME shapes, and mixed-action pairs. All base actions occur in Train; held-out states test new action combinations, while interleaved strengths test amplitude transfer. Split counts and extraction settings appear in Appendix A.1.

Table 1 shows broader detector-observed local-action and cross-region coverage than the captured-expression samples under the same measurement protocol. These counts characterize action range, not ground-truth FACS labels (Appendix A.2).

Table 1: **Observed facial-action range.** C_{act}: responsive controls among 43 MediaPipe channels; C_{pair}: coactive cross-region pairs among 609 possibilities, after within-source-group centering. Multiface uses five frontal frames per nonneutral script. Sampling and threshold sensitivity: Appendix A.2.

Dataset	$C_{\text{act}} \uparrow$	$C_{\text{pair}} \uparrow$
KDEF-dyn (Calvo et al., 2018)	29/43	151/609
FaceGrabber (Merget et al., 2016)	26/43	113/609
Multiface (Wuu et al., 2022)	25/43	75/609
GNM-derived controlled set (Ours)	37/43	286/609

4 Method

4.1 Overview

RetargetingFace separates expression semantics from native controls through neutral-referenced visual features, a shared *Expression Latent*, and modular rig-specific predictors. Each predictor learns native controls by same-rig reconstruction and connects to the fixed shared interface.

Problem Formulation. Given source controls $\mathbf{c}_s \in \mathcal{C}_s$, we seek expression-preserving target controls $\hat{\mathbf{c}}_t \in \mathcal{C}_t$. Representation learning uses shared conditions $q = (k, s)$, with action state k and strength s; native coefficient supervision remains within each rig. Matched source instances also

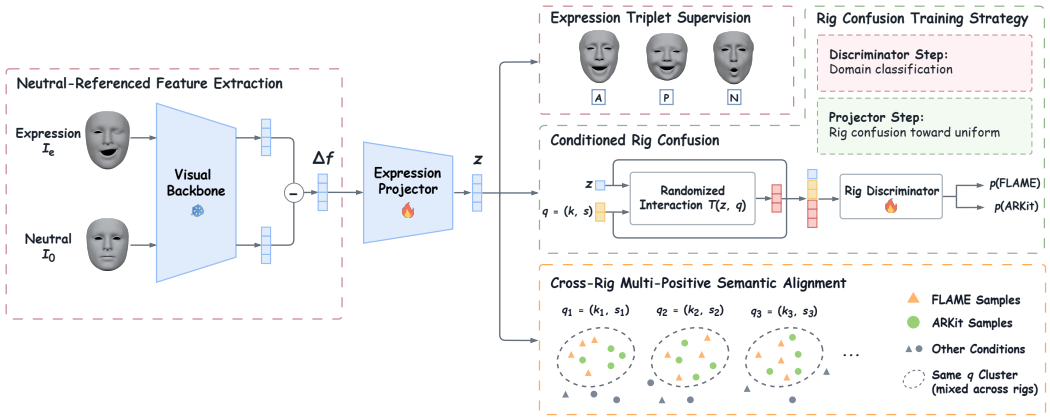

Figure 2: **Shared expression representation learning.** Triplet supervision, conditional rig confusion, and multi-positive alignment train the Expression Projector on neutral-referenced features. Fire and snowflakes mark trainable and frozen modules.

permit supervised vector-to-vector regression, without requiring coefficient-wise correspondence. Matched expressive and neutral renders define the source latent $\mathbf{z}_s$; retargeting applies $\hat{\mathbf{c}}_t = D_t(\mathbf{z}_s)$.

4.2 Neutral-Referenced Feature Extraction

Motivation. Incompatible native controls still produce comparable visual changes. Referencing those changes to the same face at rest reduces static appearance information before learning the shared representation.

Visual Backbone and Expression Projector. For rig d, expressive and neutral images I_d^e and I_d^0 share identity, camera, lighting, and material. We compute

$$\mathbf{z}_d = E\big(\mathcal{N}_f[\phi(I_d^e) - \phi(I_d^0)]\big), \qquad \|\mathbf{z}_d\|_2 = 1, \tag{1}$$

where ϕ is the CLS feature of the frozen DINOv3 ViT-B/16 *Visual Backbone*, $\mathcal{N}_f$ is Train-fitted feature normalization, and E is the learned *Expression Projector*. The resulting 128-dimensional latent emphasizes expression relative to the matched neutral. Sec. 5.3 tests the effect of this reference when source face shape varies.

Condition-based Expression Sampling. For condition $q = (k, s)$ and source variant v, rig d realizes $\mathbf{c}_{d,k,v}(s) = s\mathbf{p}_{d,k,v}$ from its independently extracted peak $\mathbf{p}_{d,k,v}$. Shared conditions specify action and relative activation along the two native paths, not equality of their coefficients (Sec. 3).

4.3 Shared Expression Representation Learning

Motivation. Neutral-referenced features can retain rig-dependent geometry. We jointly preserve expression relations, suppress conditional rig information, and align cross-rig conditions, as shown in Figure 2.

Expression Triplet Supervision. Strict FLAME triplets hold native controls fixed across shapes; semantic triplets match action and strength across source variants while contrasting different conditions. Neutral-relative distance ordering preserves increasing activation. These relations jointly constrain shape invariance and expression discrimination:

$$\mathcal{L}_{\text{expr}} = \mathcal{L}_{F,\text{strict}} + \mathcal{L}_{F,\text{semantic}} + \mathcal{L}_{A,\text{semantic}} + \mathcal{L}_{\text{mono}}. \tag{2}$$

Appendix A.5 specifies sampling and margins.

Conditional Rig Confusion. Conditioning on action and strength makes rig suppression specific to comparable expressions. A Rig Discriminator C predicts $d \in \{F, A\}$ from $\mathbf{z}$, q, and their randomized interaction $T(\mathbf{z}, q)$. Its updates minimize $\mathrm{CE}(C(\mathrm{sg}(\mathbf{z}), q), d)$ with detached embeddings. Projector updates fix C and drive its posterior toward the uniform prior:

$$\mathcal{L}_{\mathrm{conf}} = \mathbb{E}_{\mathbf{z},q}\left[-\tfrac{1}{2}\log p_F - \tfrac{1}{2}\log p_A - \log 2\right] = \mathbb{E}_{\mathbf{z},q}[D_{\mathrm{KL}}(U(F,A) \,\|\, p(d \mid \mathbf{z}, q))]\,. \tag{3}$$

Cross-Rig Multi-Positive Semantic Alignment. Confusion suppresses rig cues; alignment brings corresponding expressions together across rigs, allowing multiple source variants and shapes as positives. For anchor i, positives are $P(i) = \{j : q_j = q_i,\ d_j \neq d_i\}$, and candidates are $A(i) = \{j : \neg(q_j = q_i \wedge d_j = d_i)\}$. The objective is

$$\mathcal{L}_{\mathrm{xrig}} = -\frac{1}{N}\sum_i \frac{1}{|P(i)|}\sum_{p\in P(i)} \log \frac{\exp(\mathbf{z}_i^\top \mathbf{z}_p/\tau_{\mathrm{xrig}})}{\sum_{a\in A(i)}\exp(\mathbf{z}_i^\top \mathbf{z}_a/\tau_{\mathrm{xrig}})}. \tag{4}$$

Same-condition variants from the anchor rig are excluded from the candidate set. The combined Projector objective is

$$\mathcal{L}_E = \mathcal{L}_{\mathrm{expr}} + \lambda_{\mathrm{conf}}(t)\mathcal{L}_{\mathrm{conf}} + \lambda_{\mathrm{xrig}}\mathcal{L}_{\mathrm{xrig}}. \tag{5}$$

4.4 Rig-Specific Coefficient Learning

Motivation. A shared expression space lets each rig learn its own native realization. We freeze the Visual Backbone and Expression Projector and train each predictor using its own rig's latent–coefficient samples. Adding a target rig trains only its predictor; the representation and existing predictors remain fixed.

Same-Rig Reconstruction. For each supported rig r, predictor D_r maps its training sample's latent $\mathbf{z}_r$ to native controls $\mathbf{c}_r$ by minimizing

$$\mathcal{L}_{D_r} = \mathbb{E}_{(\mathbf{z}_r,\mathbf{c}_r)\sim\mathcal{D}_r^{\mathrm{Train}}}\left[\ell_r(\mathbf{z}_r, \mathbf{c}_r)\right]. \tag{6}$$

The native-control loss ℓ_r supervises expression, jaw, and eyelid controls for FLAME, and blendshapes with auxiliary activation supervision for ARKit. Their checkpoints are selected by Validation coefficient reconstruction (Appendix A.6). GNM and MHR follow the same reconstruction formulation with their respective native controls (Appendix A.7).

Cross-Rig Inference. At inference, the source latent $\mathbf{z}_s$ replaces a target-rig training latent at D_t, producing $\hat{\mathbf{c}}_t = D_t(\mathbf{z}_s)$. The target predictor is reused across source rigs supported by the shared representation, with all modules frozen. Sec. 5.4 evaluates this design with new target predictors and unseen source rigs. Optional predictor-training comparisons appear in Appendix A.8.

5 Experiments

We test whether the shared interface preserves expressions across native controls, which representation choices enable this transfer, and how it extends to unseen rigs and externally captured expressions.

5.1 Experimental Setup

Data and Evaluation. We fit on Train and select on Validation using the splits in Sec. 3. Test fixes source cases, target geometry, and scenes. Since native coefficients have different meanings across rigs, we compare rendered expressions in a common space: $S_{\mathrm{emb}} = \mathbf{z}_s^\top \mathbf{z}_t$, using unit embeddings from the frozen DINOv3 Visual Backbone and RetargetingFace Expression Projector. Cosine is equivalent to squared Euclidean distance through $\|\mathbf{z}_s - \mathbf{z}_t\|_2^2 = 2 - 2S_{\mathrm{emb}}$; human ratings assess perceptual fidelity independently.

Baselines and Inputs. The released conversion modules of B2F (Jang et al., 2025) and EMAGE (Liu et al., 2024) map native controls. MediaPipe, SMIRK, EMICA, EMOCA v2, and MediaPipe-to-FLAME infer controls from source renders. ReenactNet+BPNet (Kim et al., 2021) and Direct MLP learn direction-specific mappings. All outputs share target geometry and scenes, with matched neutrals available for calibration. Appendix A.3 specifies inputs and adaptations.

5.2 RETARGETING FIDELITY ON TRAINING RIGS

How faithfully does the interface retarget FLAME and ARKit expressions? We first validate the readout independently of retargeter rankings. Identical FLAME coefficients define 24 actions at three strengths across face shapes, generated after training and excluded from model selection. Correspondence is fixed by native controls, rather than the Projector. Each representation uses its native input pipeline; ours also receives the matched neutral. Matched cosine 0.9618, match–nonmatch separation 0.7675, and R@1 96.48% support both cross-shape consistency and expression discrimination (Table 2). Shape robustness alone could arise from a constant code; the discrimination tests rule out this explanation.

Table 2: **Cross-shape expression readout.** Identical coefficients define 72 action–strength conditions per shape. $\cos^+$ measures matched cosine; $\Delta_{\cos}$ measures match–nonmatch separation. AUC and R@1 assess exact-condition matching; mAP groups strengths of one action. Values average ordered shape pairs (Appendix A.4).

Representation	$\cos^+ \uparrow$	$\Delta_{\cos} \uparrow$	AUC↑	R@1 (%)↑	mAP (%)↑
EMOCA v2 (Daněček et al., 2022)	0.8184	0.2679	0.9084	38.38	34.93
EMICA (Daněček, 2023)	0.8969	0.3168	0.9440	52.55	36.77
SMIRK (Retsinas et al., 2024)	0.8484	0.2243	0.9013	28.19	36.59
RetargetingFace (Ours)	**0.9618**	**0.7675**	**0.9995**	**96.48**	**42.06**

Agreement with Human Judgments. On 50 shared cases per direction, agreement between cosine and 20-rater mean preferences is 86.1% and 75.2%, averaged within cases. It remains 83.3% and 68.9% without Ours, supporting comparisons among baselines (Appendix A.3). All methods share frozen evaluation weights and rendering; our base FLAME–ARKit predictors are trained and selected by coefficient reconstruction, not Test cosine. These tests support a task-aligned common readout. Because evaluation shares our representation, independent human evidence remains essential to the fidelity claim.

Quantitative Fidelity. RetargetingFace leads full-Test cosine and human means in both directions (Figure 3). Direct MLP learns from matched cross-rig vectors with the same predictor architecture and target-training budgets (Appendix A.3). It reaches cosine $0.8303/0.9336$, compared with $0.8956/0.9624$ for our same-rig predictors. Human means also favor our model: $2.33/2.48$ versus $2.17/2.32$. Across three training seeds, our full-Test cosine is $0.8867 \pm 0.0079/0.9595 \pm 0.0025$ (mean $\pm$ sample standard deviation; Appendix A.4). The key distinction is reuse: Direct MLP trains a separate mapping for each ordered rig pair, whereas our target predictor is trained once within its own rig and reused across source rigs. The fidelity gains complement this modularity.

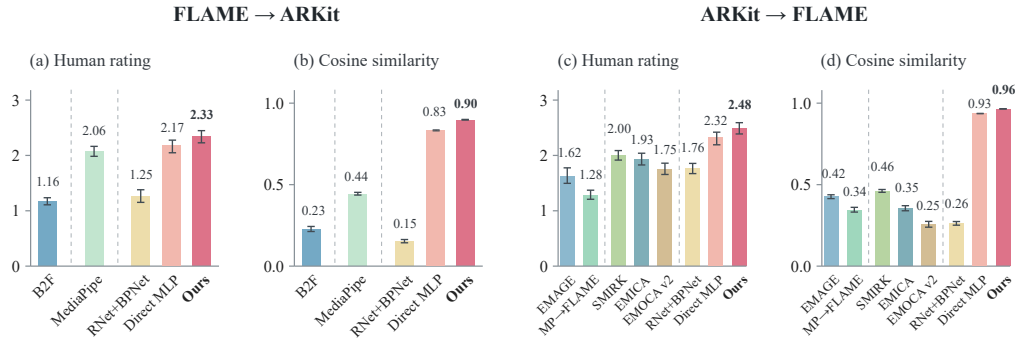

Figure 3: **Bidirectional retargeting fidelity.** Selected seed 20260719. Human means (50 cases, 20 raters; left) and cosine (full Test; right). Error bars: 95% cluster-bootstrap intervals. Higher is better. Dashed lines separate method families. RNet+BPNet abbreviates ReenactNet+BPNet; MP→FLAME abbreviates MediaPipe-to-FLAME. Protocol: Appendix A.3.

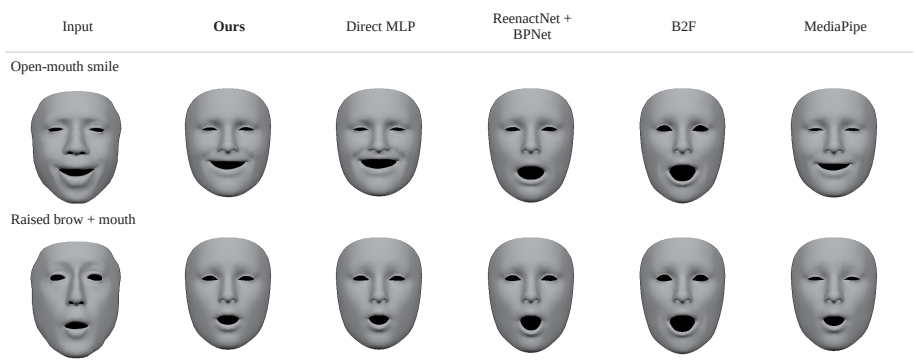

Figure 4: **FLAME→ARKit.** Open-mouth smile (top) and raised brows with mouth opening (bottom). Input is the FLAME source; all methods predict the same Said target.

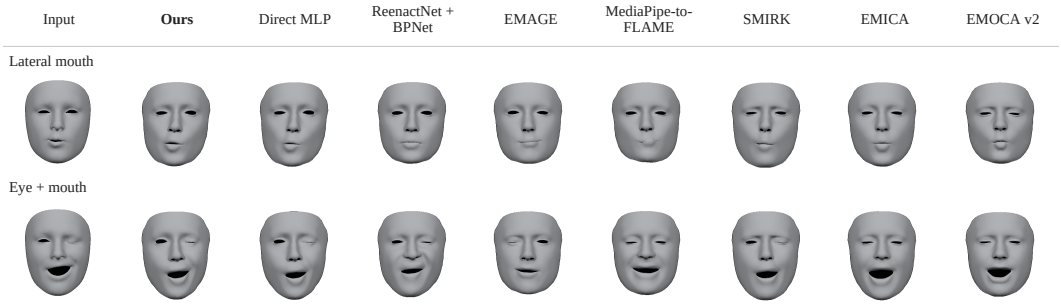

Figure 5: **ARKit→FLAME.** Lateral mouth motion (top) and unilateral eye closure with mouth opening (bottom). Said inputs; canonical-zero FLAME outputs with a common face mask.

Qualitative Comparisons. Figures 4 and 5 compare compound and asymmetric expressions under shared target scenes. FLAME source shapes are unseen by both our model and the ReenactNet+BPNet adaptation. Further examples appear in Appendix A.9.

5.3 Representation Analysis and Ablations

Which parts of the shared representation enable cross-rig transfer? We test neutral differencing by removing it while retaining all learning objectives. The PCA reference retains differencing but replaces the Projector with Train-fitted, fixed 128-dimensional PCA. Each variant trains its own same-rig predictors.

RetargetingFace outperforms the fixed PCA reference in both cosine and cycle recovery, supporting learned expression organization beyond dimensionality reduction (Table 3). Across three seeds, removing differencing lowers cosine by 0.1551 for multiple-shape FLAME versus 0.0028 for fixed-shape ARKit (paired means); both gaps are positive in every seed, consistent with reduced shape interference. Component diagnostics connect expression supervision, rig confusion, and alignment to strength ordering, conditional rig discrepancy, and cross-rig retrieval (Appendix A.4). On the base FLAME–ARKit pair, adding cross-rig predictor losses brings no consistent first-hop gain (Appendix A.8), supporting same-rig reconstruction for the base predictors.

5.4 Generalization to Unseen Native Control Systems

Can the interface transfer to and from unseen rigs? We test GNM and MHR as targets and sources outside representation training. Their predictors use the same-rig reconstruction objective while the Visual Backbone, Expression Projector, and existing predictors remain fixed. Table 4 compares the resulting transfers with direction-specific Direct MLP; Figure 6 shows corresponding renders.

Table 3: **Representation ablation.** Mean ± sample standard deviation over three seeds on 9,560 matched Test cases per direction. S_{emb}: common-readout cosine; E_{cyc}: native cycle NRMSE. PCA is a fixed-feature reference (Appendix A.4).

	F→A		A→F	
Variant	$S_{\text{emb}} \uparrow$	$E_{\text{cyc}} \downarrow$	$S_{\text{emb}} \uparrow$	$E_{\text{cyc}} \downarrow$
W/o neutral differencing	0.7309 ± 0.0040	0.5350 ± 0.0158	0.9566 ± 0.0006	0.3547 ± 0.0023
Neutral-difference PCA	0.5687 ± 0.0093	0.8591 ± 0.0066	0.6541 ± 0.0057	0.8441 ± 0.0219
RetargetingFace (Ours)	$\mathbf{0.8860 \pm 0.0081}$	$\mathbf{0.4401 \pm 0.0025}$	$\mathbf{0.9595 \pm 0.0025}$	$\mathbf{0.3523 \pm 0.0074}$

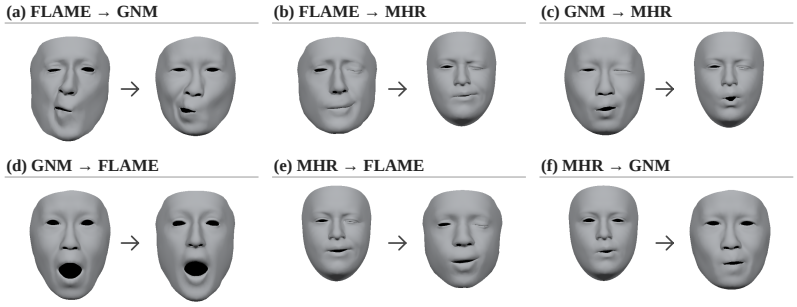

Figure 6: **Cross-rig transfer with newly attached GNM and MHR.** The panels show representative Test expressions for transfers among FLAME, GNM, and MHR. In every panel, the source render is on the left and the predicted target render is on the right. The Visual Backbone and Expression Projector remain frozen, while GNM and MHR use Validation-selected same-rig predictors. When either new rig is the source, its render is generated from a matched FLAME Test latent and re-encoded before transfer.

Same-Rig Learning for New Targets. For FLAME→GNM/MHR, each predictor learns only its native latent–coefficient mapping yet decodes FLAME latents. Both first-hop cosine and cycle NRMSE improve over direction-specific Direct MLP without cross-rig predictor supervision.

Cross-Source Reuse and Perceptual Evidence. GNM/MHR renders drive the unchanged FLAME predictor, while their target predictors accept each other's latents without pair-specific training. RetargetingFace leads direction-specific Direct MLP in first-hop cosine and mean human ratings across the reported transfers, with human-score margins of 0.07–0.34 points (Table 4). These independent ratings corroborate perceptual fidelity beyond the learned readout. Cycle NRMSE also improves except for MHR→GNM: Direct MLP recovers controls more closely, but raters prefer ours (2.11 versus 2.02). Appendix A.7 details the protocol.

Table 4: **Unseen-rig transfer.** Human: 50 cases/direction, 20 method-blinded raters/case (1–3). S_{emb}: first-hop cosine; E_{cyc}: source-space cycle NRMSE, each on the same 9,560 Test cases/direction for both methods. Direct MLP is trained per direction; row blocks test new targets, new sources, and new-rig exchange.

	Direct MLP			RetargetingFace (ours)		
Direction	Human↑	$S_{\text{emb}} \uparrow$	$E_{\text{cyc}} \downarrow$	Human↑	$S_{\text{emb}} \uparrow$	$E_{\text{cyc}} \downarrow$
FLAME→GNM	2.06	0.6435	0.6510	**2.13**	**0.7210**	**0.5993**
FLAME→MHR	1.90	0.4257	0.9507	**2.24**	**0.5445**	**0.6937**
GNM→FLAME	1.99	0.6503	0.7356	**2.29**	**0.7983**	**0.7048**
MHR→FLAME	2.04	0.4130	0.8107	**2.14**	**0.7633**	**0.7352**
GNM→MHR	2.05	0.3401	0.9375	**2.17**	**0.6005**	**0.8247**
MHR→GNM	2.02	0.3243	**0.8295**	**2.11**	**0.6707**	0.8469

5.5 Transfer to External Captured Expressions

Does the frozen model transfer to externally captured expressions? Across KDEF-dyn, FaceGrabber, and Multiface, the frozen model leads cosine and blinded human means in both directions (Figure 7). These native-fitted captured expressions test transfer beyond the constructed atlas, supported by independent perceptual judgments (Appendix A.2).

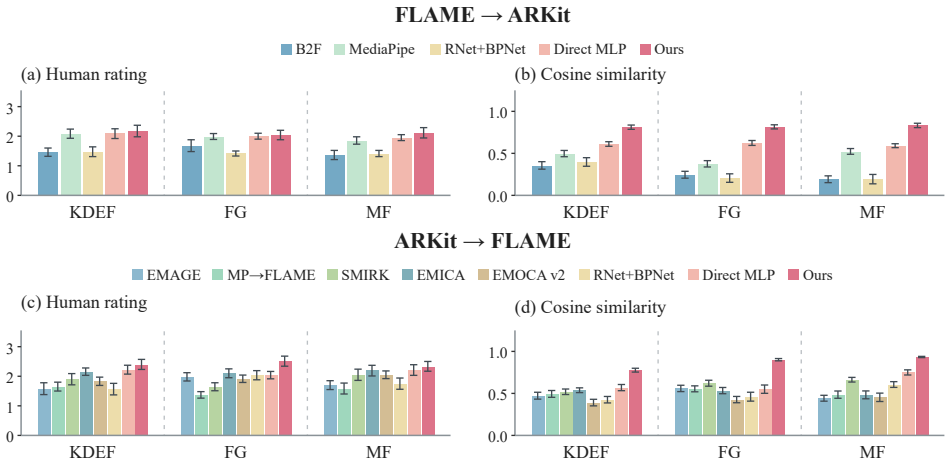

Figure 7: **External expression transfer.** Human: 50 cases, 20 raters/case (1–3); cosine: 100 cases. Error bars: 95% case-bootstrap intervals. KDEF: KDEF-dyn; FG: FaceGrabber; MF: Multiface. Protocol: Appendix A.2.

6 Conclusion

RetargetingFace makes heterogeneous native controls interoperable by separating shared expression encoding from rig-local prediction. A new target needs only a predictor trained on its own latent–control pairs; that predictor can serve different supported sources without pairwise retraining. Controlled readout tests, held-out human comparisons, and external-expression evaluations support expression preservation, while GNM/MHR transfers test reuse beyond the base rig pair. Together, these results establish native-control retargeting through reusable modules rather than a collection of ordered mappings.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experimental design, refine conceptual frameworks and hypotheses, assist with implementing and debugging data-processing, training, and evaluation code, and help interpret experimental results. Additionally, we used these tools to summarize literature, explore research directions, brainstorm, and revise the wording and organization of the manuscript. Dataset images and control coefficients were obtained through the native-control extraction and rendering procedures described in the paper. Formal mathematical claim formulation and proof development were not applicable to this work.

We reviewed all AI-assisted work through iterative revision and checks against the intended method and recorded experimental results. These checks included comparing reported numerical values with evaluation outputs and human-rating records, inspecting rendered figures and their source–target correspondence, and checking the consistency of equations, pseudocode, citations, and figure captions. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The study evaluates synthetic face renderings and expressions derived from existing datasets. Human raters participated with informed consent, and method names were hidden during scoring. We report aggregate ratings rather than participant identifiers. Retargeted outputs should be identified as synthesized facial animation and should not be presented as evidence of a person's actual facial performance.

REPRODUCIBILITY STATEMENT

The appendix specifies native control schemas, source-variant and action-state splits, neutral references, rendering settings, training objectives, update budgets, and Validation checkpoint selection. It also details the fixed evaluation populations, direct-regression baseline, human-rating scale, and uncertainty estimates. Our experiment records bind checkpoints, sample manifests, and cached predictions by file hashes, allowing reported values to be traced to the evaluated artifacts.

REFERENCES

Hector Anadon Leon and Judith Bütepage. Pose-based facial animation retargeting. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 9(1):1–18, 2026. doi: 10.1145/3804499.

Apple Inc. ARFaceAnchor.BlendShapeLocation. Apple Developer Documentation, 2026a. URL `https://developer.apple.com/documentation/arkit/arfaceanchor/blendshapelocation`. Online documentation, accessed 2026-08-31.

Apple Inc. ARKit face tracking: Blend shapes. Apple Developer Documentation, 2026b. URL `https://developer.apple.com/documentation/arkit/arfaceanchor/blendshapes`. Accessed 2026-09-18.

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 187–194, 1999. doi: 10.1145/311535.311556.

Manuel G. Calvo, Andrés Fernández-Martín, Guillermo Recio, and Daniel Lundqvist. Human observers and automated assessment of dynamic emotional facial expressions: Kdef-dyn database validation. *Frontiers in Psychology*, 9:2052, 2018. doi: 10.3389/fpsyg.2018.02052.

Sihun Cha, Serin Yoon, Kwanggyoon Seo, and Junyong Noh. Neural face skinning for mesh-agnostic facial expression cloning. *Computer Graphics Forum*, 44(2):e70009, 2025. doi: 10.1111/cgf.70009.

Yeonsoo Choi, Inyup Lee, Sihun Cha, Seonghyeon Kim, Sunjin Jung, and Junyong Noh. Deep-learning-based facial retargeting using local patches. *Computer Graphics Forum*, 44(1):e15263,

2025. doi: 10.1111/cgf.15263. URL https://onlinelibrary.wiley.com/doi/10.1111/cgf.15263.

Radek Daněček. INFERNO: Set the world on fire with FLAME. GitHub repository, 2023. URL https://github.com/radekd91/inferno. FaceReconstruction module including EMICA.

Radek Daněček, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20311–20322, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Danecek_EMOCA_Emotion_Driven_Monocular_Face_Capture_and_Animation_CVPR_2022_paper.html.

Xiangyu Fan, Jiaqi Li, Zhiqian Lin, Weiye Xiao, and Lei Yang. Unitalker: Scaling up audio-driven 3d facial animation through a unified model. In *European Conference on Computer Vision*, pp. 204–221. Springer, 2024.

Aaron Ferguson, Ahmed AA Osman, Berta Bescos, Carsten Stoll, Chris Twigg, Christoph Lassner, David Otte, Eric Vignola, Fabian Prada, Federica Bogo, et al. MHR: Momentum human rig. *arXiv preprint arXiv:2511.15586*, 2025.

Google AI Edge. Mediapipe face landmarker. Google AI for Developers, 2026. URL https://ai.google.dev/edge/mediapipe/solutions/vision/face_landmarker. Face Landmarker documentation, accessed 2026-09-18.

Bokyung Jang, Eunho Jung, and Yoonsang Lee. B2F: End-to-end body-to-face motion generation with style reference. In *Pacific Graphics Conference Papers, Posters, and Demos*. The Eurographics Association, 2025. doi: 10.2312/pg.20251256. URL https://diglib.eg.org/handle/10.2312/pg20251256.

Arthur Josi, Luiz Gustavo Hafemann, Abdallah Dib, Emeline Got, Rafael MO Cruz, and Marc-André Carbonneau. Serep: Semantic facial expression representation for robust in-the-wild capture and retargeting. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14538–14548. IEEE, 2025.

Seonghyeon Kim, Sunjin Jung, Kwanggyoon Seo, Roger Blanco i Ribera, and Junyong Noh. Deep learning-based unsupervised human facial retargeting. *Computer Graphics Forum*, 40(7):45–55, 2021. doi: 10.1111/cgf.14400.

Sol Lee, Hak-Bum Lee, Ye-Won Jang, Jung-Woo Kim, and Young-Ho Seo. A new blendshape-based retargeting for 3D facial expression. *International Journal of Computers and Applications*, 47(12):1007–1020, 2025. doi: 10.1080/1206212X.2025.2571553.

Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, 36(6):194:1–194:17, 2017. doi: 10.1145/3130800.3130813. URL https://doi.org/10.1145/3130800.3130813.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J. Black. EMAGE: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1144–1154, 2024. doi: 10.1109/CVPR52733.2024.00115. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_EMAGE_Towards_Unified_Holistic_Co-Speech_Gesture_Generation_via_Expressive_Masked_CVPR_2024_paper.html.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/ab88b15733f543179858600245108dd8-Abstract.html.

Wenchao Ma, Dario Kneubuehler, Maurice Chu, Ian Sachs, Haomiao Jiang, and Sharon Xiaolei Huang. Riganyface: Scaling neural facial mesh auto-rigging with unlabeled data. In *Advances in Neural Information Processing Systems*, volume 38, pp. 51772–51796, 2025. doi: 10.52202/085713-1543. URL https://arxiv.org/abs/2511.18601.

Daniel Merget, Tobias Eckl, Martin Schwörer, Philipp Tiefenbacher, and Gerhard Rigoll. Capturing facial videos with kinect 2.0: A multithreaded open source tool and database. In *IEEE Winter Conference on Applications of Computer Vision*, 2016.

Langyuan Mo, Haokun Li, Chaoyang Zou, Yubing Zhang, Ming Yang, Yihong Yang, and Mingkui Tan. Towards accurate facial motion retargeting with identity-consistent and expression-exclusive constraints. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 1981–1989, 2022.

Arad Mohammadi, Sebastian Weiss, Jakob Buhmann, Loic Ciccone, Robert W. Sumner, Derek Bradley, and Martin Guay. CANRig: Cross-attention neural face rigging with variable local control. *Computer Graphics Forum*, pp. e70426, 2026. doi: 10.1111/cgf.70426.

Dalton Omens, Allise Thurman, Jihun Yu, and Ronald Fedkiw. Improving facial rig semantics for tracking and retargeting. *Computer Graphics Forum*, pp. e70417, 2026. doi: 10.1111/cgf.70417.

Stylianos Ploumpis, Jan Bednarik, Gaspard Zoss, Ruslan Guseinov, Luca Prasso, Prashanth Chandran, Oliver Boyne, Vasileios Choutas, Timo Bolkart, Daoye Wang, et al. GNM head: A generative anthropometric model of the human head. *arXiv preprint arXiv:2607.23687*, 2026.

Dafei Qin, Jun Saito, Noam Aigerman, Thibault Groueix, and Taku Komura. Neural face rigging for animating and retargeting facial meshes in the wild. In *ACM SIGGRAPH 2023 conference proceedings*, pp. 1–11, 2023.

Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3D deep learning with PyTorch3D. *arXiv preprint arXiv:2007.08501*, 2020. doi: 10.48550/arXiv.2007.08501. URL https://arxiv.org/abs/2007.08501.

George Retsinas, Panagiotis P. Filntisis, Radek Daněček, Victoria F. Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3D facial expressions through analysis-by-neural-synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2490–2501, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Retsinas_3D_Facial_Expressions_through_Analysis-by-Neural-Synthesis_CVPR_2024_paper.html.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Schroff_FaceNet_A_Unified_2015_CVPR_paper.html.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. Face transfer with multilinear models. *ACM Transactions on Graphics*, 24(3):426–433, 2005. doi: 10.1145/1073204.1073209.

Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *european conference on computer vision*, pp. 622–638. Springer, 2018.

Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Xuhua Huang, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shoou-I Yu, and Yaser Sheikh. Multiface: A dataset for neural face rendering. *arXiv preprint arXiv:2207.11243*, 2022. doi: 10.48550/arXiv.2207.11243.

Peizhi Yan. Real-time MediaPipe-based FLAME animation driver. `https://github.com/PeizhiYan/mediapipe-blendshapes-to-flame`, 2025. Public mapping code and weights; accessed September 20, 2026.

ChangAn Zhu, Sima Soltanpour, and Chris Joslin. A region-based facial motion analysis and retargeting model for 3D characters. In *Pacific Graphics Conference Papers, Posters, and Demos*, 2025. doi: 10.2312/pg.20251257.

SUPPLEMENTARY MATERIAL

A IMPLEMENTATION DETAILS

A.1 CROSS-RIG DATASET CONSTRUCTION

Semantic states and source instances. The controlled dataset is constructed from a source expression atlas containing 19 base expression classes and 171 unordered class pairs. Each pair is instantiated at mixture coefficients $\alpha \in \{0.25, 0.50, 0.75\}$, giving

$$19 + 171 \times 3 = 532 \tag{7}$$

semantic states. Each state contains 50 source variants: alternative expression realizations indexed independently of rendered facial appearance. We use one fixed source appearance, `id_0000`, for native coefficient extraction, resulting in 26,600 source instances. Source variants are therefore not 50 facial identities. The ten FLAME shapes are a separate factor introduced when native controls are rendered. The state key, including its endpoint labels and mixture coefficient, defines the action index k. The mixture coefficient α is part of the semantic state and is independent of the neutral-to-peak strength s used below.

Both native extractors process each already-rendered source state image. SMIRK (Retsinas et al., 2024) provides the FLAME (Li et al., 2017) control state and MediaPipe Face Landmarker (Google AI Edge, 2026) provides the ARKit-style blendshape control state directly from that mixed-state image. FLAME uses the 55-dimensional SMIRK control contract, consisting of 50 expression, 3 jaw, and 2 eyelid controls, whereas the ARKit branch uses the 51 non-neutral MediaPipe blendshape channels in the fixed rig order, following the named ARKit facial-action convention (Apple Inc., 2026b). For every extracted peak, we generate the neutral-to-peak path at

$$s \in \{0, 0.1, \ldots, 0.9, 1.0\}. \tag{8}$$

All native controls of a source instance are scaled by the same s. Equal s specifies the same relative position along each rig's independently extracted coefficient path and provides the shared strength coordinate used throughout training.

Data partitioning. Source variants are globally partitioned into 30 training, 10 validation, and 10 test variants. The same source variant retains its split across all semantic states. Semantic states are independently partitioned into 424 training, 54 validation, and 54 test states. All 19 base classes belong to the training-state set. Mixed-expression pairs are split at the pair level, keeping their three mixture coefficients together. Thus, unseen states test new combinations of observed base actions, rather than new atomic action classes. FLAME is rendered with ten fixed natural shapes: `shape_00--05` for training, `shape_06--07` for validation, and `shape_08--09` for test. ARKit uses one fixed rig geometry.

The physical dataset stores all 11 strength levels. Training uses

$$s \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}, \tag{9}$$

while the interleaved nonzero levels $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ form the unseen-strength partition. Validation and test each contain four views formed by crossing seen or unseen semantic states with seen or unseen nonzero strengths, using their own held-out source variants and FLAME shapes. The neutral reference $s = 0$ is included in both strength partitions by construction. In each of Validation and Test, the two views built from seen states contain 76,320 records each, while the two views built from unseen states contain 9,720 records each.

Across both domains, the ordinary physical dataset contains 1,580,040 records. For each semantic state and strength, ARKit contributes 50 records, while FLAME contributes

$$30 \times 6 + 10 \times 2 + 10 \times 2 = 220 \tag{10}$$

records from the source-variant and shape roles, yielding $532 \times 11 \times 270 = 1{,}580{,}040$ records in total. The Train view contains 534,240 records. An auxiliary FLAME *canonical-zero-shape* view containing 157,740 records is stored separately for training relations that require the same native controls under a zero FLAME shape vector. This auxiliary view is reserved for canonical-shape relation construction, while normalization statistics are computed from the ordinary nonzero Train records. Here, *canonical zero* denotes the zero FLAME shape vector; expression controls remain those of the corresponding state.

Table 5: **Controlled cross-rig dataset statistics.** Ordinary records exclude the auxiliary canonical-zero-shape FLAME view. Train counts include the neutral strength; the nonzero Train counts are used for feature and coefficient statistics.

Quantity	FLAME	ARKit
Native control dimension	55	51
Natural rig identities	10 shapes	1 fixed rig
Ordinary physical records	1,287,440	292,600
Train records	457,920	76,320
Nonzero Train records	381,600	63,600

Rendering and cached visual features. Expression and neutral renders use the same shape, camera, lighting, material, and orientation. Images are rendered at 512×512 with PyTorch3D (Ravi et al., 2020) using an orthographic camera in native screen orientation, a black background, and a fixed gray material with RGB value $(0.62, 0.64, 0.67)$. The orthographic scale is 0.2956, with

$$\begin{aligned} \mathbf{p}_{\text{target}} &= (-0.00198, 0.01127, -0.04035), \\ \mathbf{p}_{\text{cam}} &= (-0.00198, 0.01127, 1.29870). \end{aligned} \tag{11}$$

Illumination implements the fixed `upper_detail` preset as three PyTorch3D point lights. The stored key/fill/lower energy values $(5.043, 5.043, 0.5043)$ are normalized to diffuse weights summing to 0.94; ambient strength is 0.01 and exposure is -0.15.

The frozen DINOv3 ViT-B/16 (Siméoni et al., 2025) receives 384×384 inputs and produces a 768-dimensional CLS feature. For every record, the cached residual is exactly

$$\Delta \mathbf{f} = \phi(I_{\text{expr}}) - \phi(I_{\text{neutral}}). \tag{12}$$

Feature statistics are estimated from nonzero Train records. The FLAME and ARKit feature moments are combined with equal domain weight, giving the two rigs equal contribution to feature normalization. Coefficient means and standard deviations are computed from the corresponding unique nonzero Train native states. The same frozen statistics are used for cached features, validation, test, and inference.

A.2 Dataset Capability Comparison and External-Transfer Protocol

Original data and experimental units. The GNM-derived controlled set contains 532 designed action states, 50 source variants per state, 11 neutral-to-peak strength levels, two independently extracted native-control spaces, and ten FLAME shapes. One source-variant–state peak is an independent expression unit for coverage; repeated strengths and shape renders are grouped with it. KDEF-dyn (Calvo et al., 2018) contains 240 neutral-to-emotion morph clips from 40 identities and six basic emotion endpoints; one identity–emotion clip is a unit. FaceGrabber (Merget et al., 2016) captures six facial-expression sequences from each of 40 subjects; one subject–expression sequence is a unit. Repeated frames are observations of the same action, not additional expression categories. The 532 designed states and six emotion categories are different label units and are not directly compared as class counts. The Multiface expression-script inventory evaluated here (Wuu et al., 2022) contains 66 scripts, covering 64 nonneutral action categories and two neutral references including the `E001` neutral-eyes-open state. We use two identities to cover this inventory. Five frontal-camera frames from each nonneutral script and three neutral images yield 323 measured images; the five frames remain observations of one script category. This sample does not represent every frame in the original Multiface collection.

Controlled factors and native interfaces. KDEF-dyn provides same-identity neutral endpoints and a morph path, which is not a native rig-strength control. FaceGrabber provides captured RGB-D sequences and separate neutral head scans. Neither public source originally supplies matched FLAME/ARKit coefficients or fixed native coefficients re-rendered across FLAME shapes. Native extraction or zero-expression neutral rendering used for our frozen-model input is a processing step and is not attributed to the original public dataset.

Measured coverage. We apply a frozen MediaPipe Face Landmarker (v0.10.35) to all 26,600 GNM source-state peak images, the expression endpoint of each of the 240 KDEF-dyn clips, and one fixed middle frame from each of the 240 FaceGrabber sequences. FaceGrabber's wide camera frames are cropped by one fixed rule around independently estimated OpenFace landmarks; no frame is selected by the measured action score. All 27,080 units yield valid Face Landmarker output. For each of the 64 nonneutral Multiface scripts, we select the first, 25%, 50%, 75%, and last numbered frame of the same frontal-camera sequence before applying the detector. We add the three neutral reference images described above; all 323 images yield valid detections. Of the detector's 52 channels, we exclude the neutral channel and eight eye-gaze channels, leaving 43 expression controls grouped into brow, eye, midface, jaw, and mouth. For each source group and channel, the 10th percentile across its measured images is a low-response reference. A group is a source variant for GNM and a captured subject identity for the external datasets; GNM retains the fixed `id_0000` appearance throughout. A channel is active in an image when its score exceeds that reference by at least 0.3. C_{act} counts channels active in at least one expression unit, taking the union over its sampled frames. C_{pair} counts pairs of simultaneously active channels in different facial regions within an image under the same rule; there are 609 possible cross-region pairs. S_{det} counts valid detections divided by all enumerated images. Thus, the scores describe detector-observed control range, not manual FACS annotation or perception quality. The same extractor also supplies our ARKit-style native controls, so this measurement is not independent of that extraction pathway.

At activation thresholds 0.2 and 0.4, respectively, GNM yields 37/43 and 37/43 responsive channels and 316/609 and 235/609 pairs; KDEF-dyn yields 29/43 and 25/43 channels and 199/609 and 100/609 pairs; FaceGrabber yields 30/43 and 24/43 channels and 166/609 and 82/609 pairs; and the five-frame Multiface sample yields 27/43 and 20/43 channels and 121/609 and 48/609 pairs. The controlled set remains broader under this detector across these thresholds. At the main threshold, the Multiface sample rises from 20/43 channels and 56/609 pairs with one frame per script to 25/43 and 75/609 with five. Holding the original low-response reference fixed, the three-frame prefix already reaches 25/43 and 75/609. Requiring support from five distinct source groups—GNM variants, or captured identities for KDEF-dyn and FaceGrabber—instead of one observed unit yields 37/43 channels and 218/609 pairs for GNM, 27/43 and 64/609 for KDEF-dyn, and 19/43 and 27/609 for FaceGrabber at the main threshold. The GNM check measures support across expression variants, not identity diversity. The five-identity check is inapplicable to the two-identity Multiface sample. These are ranges over the specified category samples rather than equal-size estimates; the differing numbers of source images and acquisition protocols constrain comparisons of dataset density or per-image diversity.

Frozen external retargeting. The Visual Backbone remains frozen, as do the Expression Projector and target predictors learned on the controlled FLAME–ARKit training set. KDEF-dyn provides 240 subject–emotion peaks over 40 identities and six emotions; FaceGrabber provides one fixed middle frame from each of 240 subject–expression sequences. For each, we select 100 cases before inspecting predictions by a fixed case-ID hash, with emotion quotas $17/17/17/17/16/16$. From the Multiface expression-script inventory, we retain one fixed image from every one of its 64 nonneutral action scripts and 36 additional images from a second identity selected by a fixed script hash, yielding our own 100-image selection. The same selected cases are used in both directions for every applicable method. Fixed native-parameter extraction and zero-expression rendering provide each source white-mesh expression and its matched neutral; the public datasets themselves contain no FLAME–ARKit pairs. FLAME→ARKit predictions are rendered on Said, and ARKit→FLAME predictions on canonical-zero FLAME, with fixed scene conditions. Each source and target render is processed by the same frozen DINOv3 Visual Backbone and RetargetingFace Expression Projector with its own matched neutral, and their unit embeddings are compared by cosine as in Appendix A.3. EMAGE's ARKit handedness is corrected before evaluation. All 100 selected cases per dataset and direction have finite native predictions for every reported method. These externally captured expressions test the frozen retargeter after native fitting. Table 6 combines the learned cosine with a separate human evaluation. For each dataset and direction, all 20 consenting raters evaluate every method on the same 50 source cases, giving 1,000 ratings per method. Cases and candidate results are presented in randomized order, with method names hidden, using the main study's three-point similarity scale. Human means average ratings within cases and then across cases. The human and cosine populations contain 50 and 100 source cases, respectively; action-category coverage refers to the latter.

External uncertainty estimates. Human scores are averaged over the 20 raters within each source case before estimating the mean across 50 cases. Human and cosine error bars are 95% case-bootstrap intervals for their respective 50- and 100-case populations. All ratings for a sampled case stay together; raters, checkpoints, and source selection remain fixed.

Table 6: **External transfer.** Each cell reports human mean / cosine (↑). Human: 50 cases, 20 raters per case, and 1,000 ratings per method/direction, scale 1–3. Cosine: 100 cases. KDEF: KDEF-dyn; FG: FaceGrabber; MF: Multiface, covering all 64 nonneutral action categories in the cosine population.

	FLAME→ARKit		
Method	KDEF	FG	MF
B2F	1.46 / 0.3567	1.68 / 0.2454	1.36 / 0.1916
MediaPipe	2.08 / 0.4972	1.99 / 0.3757	1.85 / 0.5233
RNet+BPNet	1.48 / 0.3968	1.42 / 0.2064	1.41 / 0.1924
Direct MLP	2.10 / 0.6118	2.01 / 0.6246	1.95 / 0.5915
Ours	**2.18 / 0.8124**	**2.04 / 0.8166**	**2.12 / 0.8338**

	ARKit→FLAME		
Method	KDEF	FG	MF
EMAGE	1.57 / 0.4731	1.98 / 0.5595	1.70 / 0.4426
MP→FLAME	1.65 / 0.4948	1.37 / 0.5547	1.58 / 0.4862
RNet+BPNet	1.56 / 0.4244	2.03 / 0.4614	1.74 / 0.6068
SMIRK	1.90 / 0.5195	1.64 / 0.6220	2.05 / 0.6642
EMICA	2.15 / 0.5384	2.10 / 0.5346	2.20 / 0.4831
EMOCA v2	1.84 / 0.3914	1.91 / 0.4270	2.05 / 0.4549
Direct MLP	2.21 / 0.5697	2.05 / 0.5512	2.22 / 0.7513
Ours	**2.40 / 0.7777**	**2.52 / 0.9020**	**2.33 / 0.9337**

A.3 Main-Table Semantic Evaluation

Baseline inputs and comparison status. The B2F and EMAGE baselines use their released FLAME-to-ARKit and ARKit-to-FLAME conversion modules, respectively, rather than their body- or audio-driven motion-generation pipelines. These modules receive native source controls; image-mediated estimators receive source renders with matched neutrals available for calibration. Each output is realized on common target geometry in its supported direction. ReenactNet+BPNet is adapted on the fixed FLAME `shape_00`/Said pair: the FLAME→ARKit inference uses its retrained ARKit predictor, while the reverse direction uses its FLAME predictor. MediaPipe-to-FLAME (Yan, 2025) combines image-based blendshape extraction with its released mapping. The image-based estimators are MediaPipe (Google AI Edge, 2026), SMIRK (Retsinas et al., 2024), EMICA (Daněček, 2023), and EMOCA v2 (Daněček et al., 2022).

Frozen expression-space readout. The frozen DINOv3 Visual Backbone and RetargetingFace Expression Projector are applied again to each rendered source expression and its matched source neutral, and separately to the predicted target expression and the matched target neutral. Let $\mathbf{z}_s$ and $\mathbf{z}_t$ denote the resulting unit-normalized embeddings. We report

$$S_{\cos} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{z}_{s,n}^{\top} \mathbf{z}_{t,n}. \tag{13}$$

The same source and target image preprocessing, DINOv3 feature extraction, Train feature normalization, and RetargetingFace Expression Projector weights are used for every method. Each predicted coefficient vector is rendered on the same fixed target geometry for that direction before re-encoding. The nonneutral Test populations contain 95,600 FLAME-source and 47,800 ARKit-source records. Cosine has a direct geometric meaning here: for unit embeddings, $\|\mathbf{z}_s - \mathbf{z}_t\|_2^2 = 2 - 2\mathbf{z}_s^{\top}\mathbf{z}_t$. Its ranking therefore matches Euclidean distance in the normalized expression space, without comparing heterogeneous native coefficient coordinates. Normalization removes code magnitude as a scoring factor; expression-strength distinctions must be carried by the learned directions.

Three complementary checks motivate this readout. First, coefficient-locked conditions supply a correspondence independent of the learned representation. Table 2 measures both matched consistency and nonmatch discrimination, excluding the trivial explanation that shape robustness comes from a constant code. This comparison concerns the complete encoding pipelines, including our matched-neutral input. Second, every retargeter is evaluated using the same frozen weights, normalization, geometry, and rendering. Our base FLAME–ARKit predictors are trained on coefficient reconstruction and selected by Validation coefficient NRMSE, rather than optimized or selected for cross-rig Test cosine (Appendix A.6). Third, the case-matched human analysis below tests whether cosine preserves perceived method preferences. Together these checks support a task-aligned expression readout for the evaluated conditions. The readout shares our learned representation; independent human judgments provide the perceptual evidence. It is not calibrated as an absolute perceptual score across arbitrary expressions or rigs.

Direct coefficient regression. We train two supervised MLPs that map the 55-dimensional FLAME controls directly to the 51-dimensional ARKit controls and vice versa. The input is the native control vector; no render, Visual Backbone, or Expression Projector is used for prediction. We pair native vectors by action state, strength, and source variant from the same source-state construction. Identical controls repeated across FLAME shapes are collapsed, leaving 76,320 Train pairs: 424 states, six training strengths including neutral, and 30 source variants. Train and Validation follow the original partitions.

The MLPs use the predictor architecture and native output parameterizations of Appendix A.6, replacing the 128-dimensional latent input with normalized source coefficients. Each has hidden width 512, three residual blocks with expansion factor two and dropout 0.05, and its own randomly initialized weights (seed 20260719). Input normalization and target loss scales are frozen from Train. FLAME uses the same block-weighted reconstruction objective and native constraints; ARKit uses normalized coefficient MSE plus the 0.1-weighted activation loss. Batch size is 256, with shuffled cycling through every training pair. We retain the target-specific predictor budgets: 12,750 updates for FLAME→ARKit and 17,000 for ARKit→FLAME, with the same AdamW and cosine learning-rate schedule. Every 255 or 340 updates, respectively, we evaluate the mean per-example native NRMSE in each of the four nonneutral Validation views and select the minimum equally weighted average. The selected steps are 11,220 and 3,400. Test is used only after selection.

Table 7: **Direct coefficient regression.** Mean cosine (↑) after rendering and re-encoding with the common frozen readout, on all 95,600 FLAME-source and 47,800 ARKit-source Test records. The MLP uses matched cross-rig coefficient targets; RetargetingFace trains each native predictor by same-rig reconstruction.

Method	FLAME→ARKit	ARKit→FLAME
Direct coefficient MLP	0.8303	0.9336
RetargetingFace	**0.8956**	**0.9624**

All Test predictions and scores are finite. Table 7 shows that supervised vector mapping is viable without coefficient-wise correspondence. RetargetingFace achieves higher expression-space cosine using separately trained native predictors. Each Direct MLP is tied to one ordered source–target pair. Our predictor instead learns only its target rig and accepts codes from any source supported by the shared representation, so changing the source requires no direction-specific predictor training. Sec. 5.4 tests this reuse on new rigs.

Human evaluation. For each direction, all 20 consenting raters evaluate every method on the same 50 source cases in Figure 3. Method names are hidden and presentation order is randomized. The scale is 1 (not similar), 2 (moderately similar), and 3 (most similar). This yields 1,000 ratings per method and direction. Figure 3 reports the arithmetic mean.

Case-matched cosine–human agreement. We join human ratings to cosine by source-case and method IDs, yielding 50 cases in each direction, with four methods included for FLAME→ARKit and seven for ARKit→FLAME. Each direction contains 50 distinct action-state IDs. The mean of the same 20 raters defines the human score for each method–case item. For this analysis, cosine

is evaluated on the predictions used in the human study; the separately stored ReenactNet+BPNet study predictions are re-scored through the common frozen readout instead of substituting their full-Test prediction cache.

For each source case, we form unordered method pairs and compare the signs of the human-mean and cosine differences. Human ties are uninformative about strict preference and are counted separately; an exact cosine tie receives half credit. We average agreement over informative pairs within each case and then equally over the 50 cases. We also compute Spearman rank correlation within each case, using average ranks for ties, and average these correlations. Table 8 separates all included methods, Ours-versus-baseline pairs, and baseline-only pairs. Intervals use 10,000 paired case-bootstrap resamples (seed 20260925), retaining all methods and 20 ratings within each sampled case. Checkpoints and raters remain fixed. This is an exploratory analysis of the existing study, with no metric fitting or model selection.

Table 8: **Cosine agreement with human judgments on the same cases.** Case-macro estimates with 95% case-bootstrap intervals. Top: pairwise ordering agreement (%). Bottom: mean within-case Spearman ρ. Higher values indicate stronger agreement.

Comparison	FLAME→ARKit	ARKit→FLAME
Pairwise ordering agreement (%)		
All included methods	86.1 [81.2, 90.5]	75.2 [71.5, 78.9]
Ours vs. baselines	89.0 [83.3, 94.0]	89.2 [82.8, 94.7]
Baselines only	83.3 [75.7, 90.7]	68.9 [64.7, 73.2]
Mean within-case Spearman ρ		
All included methods	0.744 [0.653, 0.825]	0.578 [0.495, 0.655]
Baselines only	0.621 [0.477, 0.753]	0.443 [0.351, 0.533]

For all-method comparisons, 39 of 300 pairs in FLAME→ARKit and 105 of 1,050 in ARKit→FLAME are human ties. The corresponding baseline-only counts are 30 of 150 and 87 of 750; Ours-versus-baseline counts are 9 of 150 and 18 of 300. No informative pair has an exact cosine tie, and every case has defined agreement and rank correlation. As a sensitivity analysis, retaining human ties with half credit gives baseline-only agreement of 76.0% and 66.9%. Positive baseline-only association supports method comparison beyond Ours' separation from the alternatives. This within-case result does not establish that cosine is an absolute quality scale across different source expressions. External-study results are reported separately and are not included in this within-case correlation analysis.

Confidence intervals. Figure 3 shows 95% percentile cluster-bootstrap intervals from 10,000 resamples. For cosine similarity, we resample expression-state IDs with replacement, retaining all scored strengths, variants, and shapes within each sampled state and recomputing the record-weighted mean. For human ratings, we first average the 20 ratings within each source case and then resample the 50 case means. The interval endpoints are the 2.5th and 97.5th percentiles of the resampled means. These intervals describe sampling variation across expression states or source cases with the selected checkpoints and raters held fixed; they do not estimate training-seed variation or uncertainty over a population of raters. Numerical labels are rounded to two decimal places; bar heights and intervals use the unrounded scores.

Headline human-score improvements. For the abstract and Introduction, we compare against the highest human mean among the evaluated existing baselines in each setting. Direct MLP is our direct-regression control and is excluded from these prior-method summaries. The relative gain is $100(\bar{h}_{\mathrm{ours}}/\bar{h}_{\mathrm{best}} - 1)\%$. Using unrounded means, we average gains equally over the two main directions (18.7%) or the six external dataset–direction settings (9.9%).

Prediction and statistical audit. For the final same-rig reconstruction model, first-hop scoring covers all 95,600 FLAME-source and 47,800 ARKit-source nonneutral Test records. Every expected sample has a finite prediction and re-encoded score in both directions. The main figures and tables use the Validation-selected seed-20260719 checkpoints; the three-seed results are reported in Appendix A.4.

A.4 Representation and Ablation Evaluation

Comparison scope. The ablation distinguishes two contrasts against RetargetingFace. W/o neutral differencing uses expressive DINOv3 CLS features with its own Train-fitted normalization while retaining expression supervision, conditional rig confusion, and cross-rig alignment. The neutral-difference PCA reference keeps expressive-minus-neutral features but replaces the trained Expression Projector with a fixed 128-dimensional PCA fitted on Train. Each representation has independently trained native predictors selected by Validation. Because PCA has no Expression Projector optimization, this contrast tests the learned representation design as a group rather than any individual learning objective or equal encoder compute.

Operational retargeting diagnostics. Table 3 reports two complementary readouts. For first-hop similarity S_{emb}, the same frozen RetargetingFace Expression Projector reads each source expression and rendered target prediction with their matched neutrals, and their unit embeddings are compared by cosine. For cycle error E_{cyc}, each variant uses its own input representation and native predictors: target coefficients are rendered, re-encoded, and decoded back into source-rig controls. This evaluation-only cycle is not used to train the predictors or perform cross-rig inference. Cycle error is a condition-macro normalized root-mean-square error within the source control space; FLAME coefficients are never subtracted from ARKit coefficients. Normalization uses frozen Train native scales, with FLAME expression/jaw/eyelid blocks weighted $0.6/0.2/0.2$ and only Train-effective ARKit channels included.

Rendering and re-encoding use the 478 action IDs present in Test (424 training states and 54 held-out mixture states), at ten nonneutral strengths and on held-out source variants and shapes, with two distinct source variants per action–strength condition selected by a sample-ID hash. This gives 9,560 matched cases per direction for both readouts. Case IDs, target scenes, and condition-balanced aggregation are fixed across variants. The common RetargetingFace Expression Projector used for S_{emb} need not rank the native cycle error identically, so the two diagnostics are reported separately.

Training-seed stability. We repeat the full method, no-neutral variant, and fixed-PCA reference with seeds 20260719, 20260720, and 20260721. Each seed has independently fitted representation parameters and native predictors selected on Validation. All variants use identical Test case IDs within each direction; cosine uses the same frozen seed-20260719 Full Expression Projector as a common evaluation readout, while cycle recovery uses each run's own representation and predictors. Table 3 reports the mean and sample standard deviation across the three seeds. The paired Full-minus-no-neutral cosine differences are 0.1551 ± 0.0051 for FLAME→ARKit and 0.0028 ± 0.0023 for ARKit→FLAME; Full exceeds the PCA reference in every seed in both directions. On the full 95,600/47,800-record Test populations, Full cosine is 0.8867 ± 0.0079 and 0.9595 ± 0.0025, respectively. All evaluated predictions are finite.

Coefficient-locked encoder readout. For Table 2, we construct a new independent evaluation dataset using 24 actions selected with seed 20260925, strengths 0.2, 0.6, and 1.0, and a fixed set of FLAME face shapes. The dataset is generated separately after model training and is not used for parameter optimization or checkpoint selection. One legal 55-dimensional FLAME state per action–strength condition is rendered on every face shape with the same camera, lighting, material, and orientation; coefficient equality is checked across rendering shapes. This gives 72 conditions per shape, evaluated over ordered shape-pair galleries. All four frozen encoders process the same expressive PNGs using their native input paths; our encoder additionally receives the corresponding neutral PNG through its standard DINOv3 difference and frozen training normalization. No baseline receives this difference or our DINOv3. For query condition q and target-shape gallery G, $\cos^{+}$ is the cosine between unit-normalized codes for the same coefficient-locked condition on different shapes. $\Delta_{\cos}$ subtracts the mean cosine to the 71 other conditions in G from this matched value. AUC is the probability that the match exceeds a nonmatch in cosine, assigning half credit to ties. Exact R@1 and action mAP use the same cosine ranking. We average over ordered shape pairs, and paired uncertainty resamples the 24 actions rather than the correlated shape pairs. The 95% action-bootstrap interval for the $\Delta_{\cos}$ advantage over the strongest baseline, EMICA, is $[0.4311, 0.4691]$; the advantage is positive for every evaluated shape pair. This evaluation supports cross-shape consistency together with expression-condition discrimination for the compared encoding pipelines. It

does not by itself validate every cross-rig facial detail; case-matched human agreement and held-out retargeting results provide the complementary evidence.

Targeted component diagnostics. Table 9 reuses the frozen component runs to examine the representation directly, without a downstream predictor. All runs start from the same selected first-stage checkpoint and use the same second-stage budget and Validation selection rule. Each removes one second-stage ingredient; the first-stage objectives are shared. W/o expression supervision disables all three triplet terms and neutral monotonicity together. These exploratory diagnostics isolate changes from second-stage interventions; they do not establish that each ingredient is necessary when training from scratch.

Table 9: **Second-stage component diagnostics.** Each row uses a metric corresponding to the ingredient's intended role. Our model and its matched intervention share the evaluated cases. Confidence intervals resample action IDs with fixed trained checkpoints.

Removed ingredient	Diagnostic	Ours	Removed
Neutral referencing	$C_{\text{shape}} \uparrow$	0.8551	0.5322
Expression supervision	Strength order (%)↑	99.53	61.21
Conditional rig confusion	$U_{\text{mean}} \downarrow$	0.1372	0.1853
Cross-rig alignment	F→A mAP (%)↑	4.06	1.97
Cross-rig alignment	A→F mAP (%)↑	3.21	1.29

C_{shape} compares the same expression on Test FLAME shapes 08 and 09; neutral referencing raises it by 0.3229 (paired action-bootstrap 95% interval: $[0.3140, 0.3313]$). Strength order uses $d(s) = 1 - \cos(\mathbf{z}_0, \mathbf{z}_s)$ and checks $d(s_l) < d(s_h)$ for all 45 pairs of the ten nonzero strengths along each fixed action–variant–shape path. Averaging within paths, actions, and rigs gives a 38.32 percentage-point advantage ($[37.79, 38.86]$) with expression supervision. For neutral-difference inputs, $\mathbf{z}_0$ is the encoded, Train-normalized zero residual.

For a fixed action–strength condition q, let $\boldsymbol{\delta}_v$ be the mean FLAME code across the two Test shapes minus the ARKit code for source variant v. We measure conditional mean discrepancy using the off-diagonal estimator

$$U_{\text{mean}} = \frac{1}{|Q|} \sum_{q \in Q} \frac{\sum_{v \neq w} \boldsymbol{\delta}_v^\top \boldsymbol{\delta}_w}{V(V-1)}, \qquad V = 10. \tag{14}$$

Rig confusion reduces this discrepancy by 0.0481 ($[0.0450, 0.0509]$); the result concerns conditional means, not removal of every possible rig cue. Cross-rig retrieval ranks opposite-rig codes by cosine with matched action–strength conditions as positives. Alignment improves mAP by 2.10 and 1.92 percentage points, respectively. These representation-specific effects complement the end-to-end transfer comparisons in Table 3.

Ablation controls. The no-neutral variant keeps the RetargetingFace architecture and representation-learning losses, changing only the visual input and its Train-fitted normalization. The neutral-difference PCA reference freezes its representation after Train-only fitting. All three variants train their native predictors independently by Same-Rig Reconstruction and select checkpoints on Validation. PCA fitting does not match the learned Expression Projector's optimization budget, so the PCA row is a fixed-feature reference rather than an equal-compute ablation.

A.5 Expression Relations and Shared Encoding

Source states and extracted peaks. The source atlas provides rendered base expressions and pairwise mixtures across source variants. We retain the original state key, including its endpoint labels and mixture coefficient α, when assigning k. Both extractors process each already-rendered state image and directly produce the peak coefficients for that state. A prototype identifies one domain's extracted peak for one source instance. FLAME rendering shapes are varied independently after extraction, while the retargeted native control vector contains the expression, jaw, and eyelid controls.

For example, the source key `blow_0.50_pucker` identifies a mixture state with $\alpha = 0.5$. Setting $s = 0.5$ then scales each domain's extracted peak to half its coefficient-path amplitude. State composition and neutral-relative strength therefore form two complementary factors of the condition. The shared condition is a supervision prior grounded in common source construction, and source-image correspondence supplies condition-level semantic supervision across rigs.

Visual Backbone, normalization, and Expression Projector. We use

$$\mathcal{N}_f(\mathbf{u}) = (\mathbf{u} - \boldsymbol{\mu}_f) \oslash \boldsymbol{\sigma}_f, \tag{15}$$

with the nonzero Train statistics described above. Neutral subtraction uses a matching render with zero expression, jaw, and eyelid controls and the same identity context. The *Visual Backbone* is a frozen DINOv3 ViT-B/16. The *Expression Projector* is

$$768 \to 512 \to 128, \tag{16}$$

implemented as `Linear--GELU--Linear`, followed by ℓ_2 normalization with numerical epsilon 10^{-12}. Its output is the 128-dimensional unit-normalized *Expression Latent*.

Triplet relations. We use margin-based anchor–positive–negative metric supervision following the standard triplet-learning formulation (Schroff et al., 2015), with task-specific relations defined below. All three triplet streams exclude neutral samples. A base FLAME-strict positive shares the anchor's source instance, state, strength, and complete coefficient vector, but uses a different natural rendering shape. During training, exactly half of the strict positives are deterministically replaced by the same state rendered with the auxiliary canonical-zero shape. Strict negatives contrast a different state at the same strength, an adjacent nonzero strength on the same source path, or a more distant nonzero strength on that path. Their proportions are 55%, 30%, and 15%, respectively, enforced over a sampling quota round. Within each negative category, the negative uses the anchor's or positive's natural shape with equal quotas. Adjacency is defined by rank among the nonzero training strength levels; distant strengths differ by at least two ranks.

A semantic positive has the same state and exact strength as its anchor but a different source variant and prototype. A FLAME semantic positive also has a different rendering shape. Semantic negatives use a third source variant and prototype and contrast state only, strength only, or both, with equal category quotas. FLAME negative-shape roles are balanced as above; ARKit has a fixed rig. These rules use condition labels and source metadata as fixed supervision. The condition q specifies the cross-rig semantic relation used to organize the shared expression space.

For unit embeddings, let $d(\mathbf{u}, \mathbf{v}) = 1 - \mathbf{u}^\top \mathbf{v}$ and $[x]_+ = \max(x, 0)$. The symmetric triplet loss is

$$\ell_{\text{tri}}(a, p, n) = \tfrac{1}{2}\left([m + d(\mathbf{z}_a, \mathbf{z}_p) - d(\mathbf{z}_a, \mathbf{z}_n)]_+ + [m + d(\mathbf{z}_a, \mathbf{z}_p) - d(\mathbf{z}_p, \mathbf{z}_n)]_+\right). \tag{17}$$

The strict-stream margins are 0.20 for a different state at the same strength, 0.05 for adjacent strengths, and 0.10 for distant strengths. The semantic margin is 0.20. Each Projector update draws 72 triplets from each of the FLAME-strict, FLAME-semantic, and ARKit-semantic streams.

Condition structure and instance variation. Strict positives isolate rendering-shape variation by holding the control vector fixed. Semantic positives share their construction condition but may differ in local expression details. The triplet inequalities impose relative-distance ordering while preserving within-condition instance variation. Rig-specific predictors subsequently learn to recover each training sample's own native coefficients from this shared representation.

Strength ordering. Neutral-monotonic tuples contain a matching neutral, a lower nonzero strength, and a higher nonzero strength from the same domain, source path, and rendering shape, with $s_h - s_\ell \geq 0.05$. With these embeddings denoted by $\mathbf{z}_0, \mathbf{z}_\ell, \mathbf{z}_h$, the ordering loss is

$$\ell_{\text{mono}} = [0.05 + d(\mathbf{z}_0, \mathbf{z}_\ell) - d(\mathbf{z}_0, \mathbf{z}_h)]_+. \tag{18}$$

Each Projector update uses 72 monotonic tuples, balanced as 36 FLAME and 36 ARKit samples. Together with the three triplet streams, this gives

$$\mathcal{L}_{\text{expr}} = \mathcal{L}_{F,\text{strict}} + \mathcal{L}_{F,\text{semantic}} + \mathcal{L}_{A,\text{semantic}} + \mathcal{L}_{\text{mono}}, \tag{19}$$

with unit weight for all four active terms.

Conditional Rig Confusion. The condition vector concatenates a 532-way action one-hot vector with the continuous strength:

$$\mathbf{q}_{\text{raw}} = [\text{onehot}(k), s], \qquad \mathbf{q} = \frac{\mathbf{q}_{\text{raw}}}{\sqrt{1+s^2}}, \tag{20}$$

so $\mathbf{q} \in \mathbb{R}^{533}$. A fixed randomized multilinear interaction (Long et al., 2018) uses matrices

$$R_z \in \mathbb{R}^{1024\times 128}, \qquad R_q \in \mathbb{R}^{1024\times 533}, \tag{21}$$

initialized once from the training seed and stored as non-trainable buffers:

$$T(\mathbf{z}, \mathbf{q}) = \frac{(R_z\mathbf{z}) \odot (R_q\mathbf{q})}{\sqrt{1024}}. \tag{22}$$

The *Rig Discriminator* receives the 1,685-dimensional concatenation $[\mathbf{z}, \mathbf{q}, T]$. It applies a linear projection to 256 dimensions, LayerNorm, and GELU, followed by three residual MLP blocks. Each block uses LayerNorm, a $256 \to 512 \to 256$ expansion with GELU and dropout 0.1, and a residual connection. A final LayerNorm and linear layer predict the two rig labels.

The synchronized rig-confusion batch contains 32 nonzero conditions, four source variants shared by both rigs for each condition, and both domains, giving 256 samples. For each condition, one of the four FLAME rows uses the canonical-zero shape while the other three use natural FLAME shapes; the ARKit rows use the corresponding four natural source variants. Thus each batch contains 32 canonical-zero-shape FLAME samples, 96 natural FLAME samples, and 128 ARKit samples. The sampler operates over 2,120 nonzero Train conditions.

For each *Projector Step*, five independent *Discriminator Steps* minimize

$$\mathcal{L}_C = \text{CE}\left(C(\text{sg}(\mathbf{z}), \mathbf{q}), d\right). \tag{23}$$

The *Projector Step* freezes the Rig Discriminator parameters while preserving gradients with respect to $\mathbf{z}$, and minimizes

$$\mathcal{L}_{\text{conf}} = \mathbb{E}_{\mathbf{z},q}\left[-\frac{1}{2}\log p_F - \frac{1}{2}\log p_A - \log 2\right]. \tag{24}$$

The final configuration uses explicit alternating Discriminator and Projector optimization with the uniform-confusion objective above.

Cross-Rig Multi-Positive Semantic Alignment. The alignment loss is evaluated on the same synchronized 256-sample batch. For every anchor there are four positives from the opposite rig with the same (k, s) condition. For every anchor, the denominator contains the four opposite-rig positives together with all 248 different-condition rows, giving 252 candidates. The four same-condition samples from the anchor rig define the within-rig set and remain outside this candidate set. We use $\lambda_{\text{xrig}} = 0.25$ and $\tau_{\text{xrig}} = 0.1$ throughout Step 2.

Projector optimization and selection. Projector optimization first combines expression supervision with conditional rig confusion (Step 1), then initializes from the Validation-selected checkpoint and adds cross-rig alignment (Step 2). Each stage has a fixed budget of 78 epochs with 465 Projector updates per epoch: 36,270 Projector and 181,350 Rig Discriminator updates per stage. Both use AdamW with $\beta = (0.9, 0.95)$, $\epsilon = 10^{-10}$, and gradient clipping at 1. The Projector uses learning rate 2×10^{-4} and weight decay 10^{-2}; the Rig Discriminator uses learning rate 3×10^{-4} and weight decay 10^{-4}. Both use cosine schedules with a minimum learning-rate ratio of 0.1. Projector forward passes use BF16 autocast. The confusion weight is zero through the first 2,000 Projector updates and then increases linearly over 4,000 updates to $\lambda_{\text{conf}}^{\max} = 0.2$.

Checkpoint selection uses only the fixed Validation views. Thirty scheduled checkpoints are considered, with candidates restricted to at least 6,000 Projector updates. The primary selection quantity is the larger of the fresh unconditional and conditional MLP rig-separability scores. Differences within 0.01 are treated as ties and resolved by the fixed retargeting-retrieval diagnostic, strict-triplet satisfaction, and then the earlier update. The Step 1 checkpoint is selected at epoch 73 (33,945 Projector updates); the final Step 2 checkpoint is selected at epoch 68 (31,620 updates within that stage). Projector selection is performed exclusively on the fixed Validation protocol.

A.6 RIG-SPECIFIC PREDICTOR OUTPUTS AND RECONSTRUCTION OBJECTIVES

Predictor architecture. The FLAME and ARKit rig-specific predictors are independent decoders D_F and D_A. Each predictor first applies LayerNorm to the 128-dimensional Expression Latent, followed by a linear projection to hidden dimension 512 and GELU. Three residual blocks then apply

$$\mathrm{LN} \to \mathrm{Linear}(512, 1024) \to \mathrm{GELU} \to \mathrm{Dropout}(0.05) \to \mathrm{Linear}(1024, 512) \tag{25}$$

with a skip connection, followed by a final LayerNorm. FLAME uses one 55-dimensional output head. ARKit uses separate 51-dimensional coefficient and activation-logit heads.

FLAME output and reconstruction. The FLAME branch follows the SMIRK55 control contract. Let $P_F(\mathbf{z})$ be the predictor's raw 55-dimensional output. The first 53 entries parameterize standardized expression and jaw coefficients, and the last two are eyelid logits:

$$D_F(\mathbf{z}) = \left[\boldsymbol{\mu}^c_{F,1:53} + \boldsymbol{\sigma}^c_{F,1:53} \odot P_F(\mathbf{z})_{1:53}; \mathrm{sigmoid}(P_F(\mathbf{z})_{54:55})\right]. \tag{26}$$

These are one-based mathematical indices. In the implementation, the three blocks are `[0:50]`, `[50:53]`, and `[53:55]`. Shape is a separate 300-dimensional input to the frozen SMIRK geometry. Head pose is fixed in the controlled render context; both predicted eyelid controls remain differentiable. The eye-closure deformation fields are part of the SMIRK-compatible geometry (Retsinas et al., 2024).

Let $\mathbf{r} = (D_F(\mathbf{z}) - \mathbf{c}_F) \oslash \boldsymbol{\sigma}^c_F$, with the two eyelid scales floored at 0.05. The per-sample FLAME reconstruction loss is

$$\ell_F(\mathbf{z}, \mathbf{c}_F) = \frac{0.6}{50}\|\mathbf{r}_{1:50}\|_2^2 + \frac{0.2}{3}\|\mathbf{r}_{51:53}\|_2^2 + \frac{0.2}{2}\|\mathbf{r}_{54:55}\|_2^2. \tag{27}$$

This block reduction is used for FLAME Same-Rig Reconstruction.

ARKit output and reconstruction. Let P^c_A and P^a_A denote the coefficient and activation-logit outputs of the ARKit predictor. With a training-derived effective-channel mask $\mathbf{m}_A$, the unclipped physical coefficient output is

$$\mathbf{a}_A(\mathbf{z}) = \mathbf{m}_A \odot \left(\boldsymbol{\mu}^c_A + \boldsymbol{\sigma}^c_A \odot P^c_A(\mathbf{z})\right), \qquad D_A(\mathbf{z}) = \mathrm{clip}_{[0,1]}\left(\mathbf{a}_A(\mathbf{z})\right). \tag{28}$$

Let $\mathcal{J}_A$ be the effective channels, $\mathcal{N}^c_A$ the training coefficient standardization, and $b_j = \mathbf{1}[c_{A,j} > \tau_A]$. Reconstruction supervises the raw standardized output and the auxiliary activation logits:

$$\ell_A(\mathbf{z}, \mathbf{c}_A) = \frac{1}{|\mathcal{J}_A|}\sum_{j \in \mathcal{J}_A}\left[\left(P^c_{A,j}(\mathbf{z}) - \mathcal{N}^c_A(\mathbf{c}_A)_j\right)^2 + \lambda_{\mathrm{act}}\,\mathrm{BCE}_{w^+_j}\left(P^a_{A,j}(\mathbf{z}), b_j\right)\right]. \tag{29}$$

The threshold is $\tau_A = 0.01$ and $\lambda_{\mathrm{act}} = 0.1$. Positive-class weights are estimated from the Train labels. The coefficient head defines the physical ARKit output, while the activation head provides auxiliary supervision during Same-Rig Reconstruction.

Same-Rig Reconstruction Training. After selecting the Expression Projector, its parameters remain fixed while the FLAME and ARKit predictors are trained independently from native coefficient supervision. Both predictors use AdamW with learning rate 3×10^{-4}, $\beta = (0.9, 0.95)$, $\epsilon = 10^{-10}$, weight decay 10^{-4}, gradient clipping at 1, and cosine decay to 3×10^{-5}. FLAME is trained for 17,000 updates using 192 natural and 64 canonical-zero-shape samples per update. ARKit is trained for 12,750 updates using 256 natural samples per update.

FLAME validation is performed every 340 updates and ARKit validation every 255 updates, giving 50 scheduled validation opportunities for each branch. The two predictors are selected independently by the mean nonzero-strength macro NRMSE over the four natural Validation views. Their selected states form the final rig-specific predictors used for cross-rig inference.

A.7 UNSEEN-RIG TRANSFER EVALUATION

Native target learning and source transfer. The DINOv3 Visual Backbone, Train feature normalization, and RetargetingFace Expression Projector remain fixed throughout these evaluations.

For FLAME→GNM/MHR, only the new target predictor is trained, using its own rig's expression latents and native coefficients for same-rig reconstruction. Table 4 compares the GNM and MHR checkpoints at updates 33,500 and 48,500 with direction-specific Direct MLP. The existing FLAME and ARKit predictors remain unchanged. For GNM/MHR→FLAME, new source renders drive the existing representation and FLAME predictor directly, without new source-specific predictor training. GNM↔MHR reuses the same independently trained new-rig predictors without learning a pair-specific mapping. Direct MLP is trained separately for each ordered direction. For Figure 6, GNM/MHR source renders are generated from matched FLAME Test latents with the Validation-selected same-rig predictors, then rendered and re-encoded before transfer.

GNM and MHR predictor training. For each new rig, native expressive and matched-neutral renders are passed through the frozen Visual Backbone and Expression Projector to obtain 128-dimensional latents. A predictor receives its own rig's latent and reconstructs that rig's native controls; the caches contain 76,320 Train and 52,580 Validation samples per rig. Control means and standard deviations are fitted on the 63,600 nonneutral Train samples, and channels with raw standard deviation below 10^{-5} are excluded. The only optimized data loss is the mean squared error between predicted normalized controls and $(\mathbf{c}_r - \boldsymbol{\mu}_r)/\boldsymbol{\sigma}_r$ over the remaining channels. GNM predicts 383 controls with a width-512, three-residual-block predictor (dropout 0.2) and a separate 33-channel tail; its trunk is initialized from the FLAME predictor and updated during training. MHR predicts 72 controls with a newly initialized width-1024, three-residual-block predictor (dropout 0.1) and standardizes its input latents using Train statistics. Neither predictor uses cross-rig coefficient pairs or cross-rig losses for this adaptation.

Both runs use seed 20260925, batch size 256, AdamW ($\beta = (0.9, 0.95)$, $\epsilon = 10^{-10}$), and gradient clipping at 1. GNM/MHR use learning rates $10^{-4}/3 \times 10^{-4}$ and weight decays $10^{-2}/10^{-3}$, respectively. Every 500 updates, Validation macro NRMSE is averaged over four nonneutral seen/unseen state-by-strength views; the minimum selects the checkpoint without Test data. ReduceLROnPlateau halves the rate after six non-improving evaluations (absolute threshold 10^{-4}, one-evaluation cooldown) to a floor of 10^{-6}. Runs last at least 12,000 updates, stop after 12 non-improving evaluations at the floor, and are capped at 120,000 updates.

Scope of rig generalization. GNM supplies the source expression atlas used for native extraction, but its images and controls are not inputs to base representation or predictor training. Its extension therefore tests a new native system within the atlas's expression content. Transfer beyond that constructed content is evaluated separately on captured expressions (Sec. 5.5).

First-hop expression preservation. The source expression and predicted target expression are rendered with their respective matched neutrals and passed through the same frozen visual readout used for the main experiments. Their unit-normalized embeddings define S_{emb}. Each direction uses 9,560 matched Test cases for both methods. The FLAME-source directions test newly attached native predictors; GNM/MHR→FLAME tests reuse of the existing FLAME predictor for new source rigs, while GNM↔MHR tests reuse between the new rigs.

Human comparison. For each reported direction, the two methods share 50 source cases. Twenty participants rate every case on the same method-blinded 1–3 similarity scale as the main study. Table 4 reports the mean human score for each method and direction; no confidence intervals are reported for this comparison. These ratings provide a perceptual evaluation separate from the frozen-projector readout.

Source-space cycle recovery. For RetargetingFace, we render and re-encode the first-hop target prediction, then apply the source-rig predictor to recover the original native controls. Cycle NRMSE follows the source-space normalization and aggregation protocol in Appendix A.4. For FLAME→GNM/MHR, the return leg uses the existing FLAME predictor. For GNM/MHR as source, including GNM↔MHR, the same-rig source predictor is used on the diagnostic return leg but not on the first hop. Direct MLP follows the same Test-case, rendering, readout, and source-space normalization protocol. Since source control spaces and normalization differ, cycle scores characterize recovery within each direction rather than a common geometric-error scale across rigs.

Comparisons hold the target rig and source cases fixed. The shared expression readout is not calibrated as an absolute perceptual scale across different native systems (Appendix A.3).

A.8 Optional Cross-Rig Refinement Comparison

We evaluate whether cross-rig feedback improves the base FLAME–ARKit predictors learned through same-rig reconstruction. Starting from the selected native predictors, this comparison adds expression consistency and source coefficient recovery. All main experiments retain same-rig reconstruction.

Differentiable Expression Re-Projection path. For $t \neq d$, the Target-Rig Predictor is D_t. We denote the fixed composition of target-rig differentiable geometry, PyTorch3D rendering, DINOv3 ViT-B/16 CLS extraction, feature normalization, and the frozen Expression Projector by G_t. It maps predicted target controls back into the shared expression space:

$$\mathbf{z}_{dt} = G_t\left(D_t(\mathbf{z}_d)\right). \tag{30}$$

The SMIRK/FLAME and ARKit geometry, renderer, DINOv3 Visual Backbone, and Expression Projector remain frozen. Gradients propagate through this fixed differentiable path to the predicted target controls and update the two rig-specific predictors.

The comparison uses *Same-Rig Reconstruction* ($\mathcal{L}_{\mathrm{rec}}$), *Expression Consistency* ($\mathcal{L}_{\mathrm{emb}}$), and *Source Coefficient Recovery* ($\mathcal{L}_{\mathrm{cyc}}$):

$$\begin{aligned} \mathcal{L}^d_{\mathrm{rec}} &= \ell_d(\mathbf{z}_d, \mathbf{c}_d), \\ \mathcal{L}^d_{\mathrm{emb}} &= 1 - \cos\left(\mathbf{z}_{dt}, \mathrm{sg}(\mathbf{z}_d)\right), \\ \mathcal{L}^d_{\mathrm{cyc}} &= \ell_d(\mathbf{z}_{dt}, \mathbf{c}_d). \end{aligned} \tag{31}$$

The Expression Consistency reference is the source sample's own Expression Latent. For FLAME-to-ARKit transfer, range regularization is applied to the unclipped first-hop ARKit output:

$$\mathcal{L}_{\mathrm{bnd}} = \mathrm{Mean}_{j \in \mathcal{J}_A}\left([-a_{A,j}]^2_+ + [a_{A,j} - 1]^2_+\right), \qquad \lambda_{\mathrm{bnd}} = 0.01. \tag{32}$$

The comparison objective is

$$\mathcal{L}_D(n) = \frac{1}{2} \sum_{d \in \{F,A\}} \mathbb{E}\left[\mathcal{L}^d_{\mathrm{rec}} + r(n)\mathcal{L}^d_{\mathrm{emb}} + 0.5 r(n)\mathcal{L}^d_{\mathrm{cyc}} + 0.01 \mathbf{1}[t = A]\mathcal{L}_{\mathrm{bnd}}\right], \tag{33}$$

where

$$r(n) = \min\left(1, \frac{n+1}{954}\right). \tag{34}$$

The final Same-Rig Reconstruction, Expression Consistency, and Source Coefficient Recovery weights in this comparison are $1 : 1 : 0.5$, and the two source domains are weighted equally.

Refinement training protocol. The Pure3D Train pools contain 457,920 FLAME records and 76,320 ARKit records over 2,544 conditions, corresponding to 424 training states and six training strengths. We use four GPUs with global FLAME and ARKit batches of 48 and 8. Each rank therefore receives 12 FLAME and 2 ARKit samples; the differentiable rendering pass further processes them in FLAME micro-batches of 6 and ARKit micro-batches of 1. One complete physical-record coverage takes 9,540 successful optimizer updates. The full budget is 19,080 successful updates, so every Train record in both domains is visited exactly twice.

The two predictors are optimized jointly with AdamW, $\beta = (0.9, 0.95)$, $\epsilon = 10^{-10}$, weight decay 10^{-4}, gradient clipping at 1, and cosine learning-rate decay to 10% of the initial value. Mixed-precision optimization uses a GradScaler initialized at 128. The rendering/sampling seed is 20260721 and the Validation seed is 20260719.

Learning-rate and checkpoint selection. We compare learning rates 1×10^{-5} and 3×10^{-5} from the same selected Expression Projector and same-rig reconstruction predictors. Each candidate is first trained to the fixed 25% nonzero physical-coverage point and evaluated on the same frozen quick-Validation protocol. The selected learning rate is 1×10^{-5}, which is then continued under the unchanged 19,080-update schedule.

Validation is evaluated every 250 successful updates and at fixed coverage milestones. The validation score is normalized to the frozen Same-Rig Reconstruction reference:

$$S = 0.5\frac{L_{\text{emb}}^{\text{val}}}{L_{\text{emb,base}}^{\text{val}}} + 0.3\frac{E_{\text{cyc}}^{\text{val}}}{E_{\text{cyc,base}}^{\text{val}}} + 0.2\frac{E_{\text{rec}}^{\text{val}}}{E_{\text{rec,base}}^{\text{val}}}, \tag{35}$$

where $L_{\text{emb}}^{\text{val}}$ is the source-embedding first-hop loss, $E_{\text{cyc}}^{\text{val}}$ is the normalized cycle reconstruction error, and $E_{\text{rec}}^{\text{val}}$ is the same-domain macro NRMSE. During training, we retain the best validation checkpoint over the complete optimization trajectory, the best checkpoint obtained after one full physical-data coverage, and the final checkpoint at the end of the fixed training budget. After training, these candidates are reevaluated on the full Validation set using the same score S. The checkpoint with the lowest validation score is selected, with the earlier update used to break ties. The final selected checkpoint occurs at update 1,250.

Differentiable rendering implementation. The Differentiable Expression Re-Projection bridge uses the same 512×512 PyTorch3D render configuration as dataset construction and the same frozen DINOv3 preprocessing and Train normalization. FLAME coefficients are realized through the SMIRK-compatible FLAME geometry with the current FLAME shape; ARKit vertices are produced from the fixed neutral mesh and weighted blendshape offsets. Native left/right orientation is preserved directly in screen coordinates. The DINOv3 path remains differentiable with respect to the rendered image and therefore to the predicted native controls, although its parameters are frozen.

Effect on first-hop retargeting. Table 10 compares the selected same-rig predictors with the refinement variant on the same nonneutral Test populations and frozen expression-space readout. Adding both cross-rig losses yields no consistent first-hop gain on the base pair: same-rig reconstruction scores slightly higher for FLAME→ARKit, while refinement is higher by 0.0010 for ARKit→FLAME. Figure 8 shows four matched cases under identical target scenes. The base FLAME–ARKit results therefore use native reconstruction; this joint comparison does not assign an effect to either added loss separately.

Table 10: **Optional cross-rig refinement comparison.** Mean source–target cosine on the full nonneutral Test populations, measured by the same frozen DINOv3 Visual Backbone and Retargeting-Face Expression Projector. The refinement variant starts from the selected same-rig predictors and jointly adds expression consistency and source coefficient recovery.

Predictor training	FLAME→ARKit	ARKit→FLAME
Same-Rig Reconstruction	**0.8956**	0.9624
With cross-rig losses	0.8953	**0.9634**

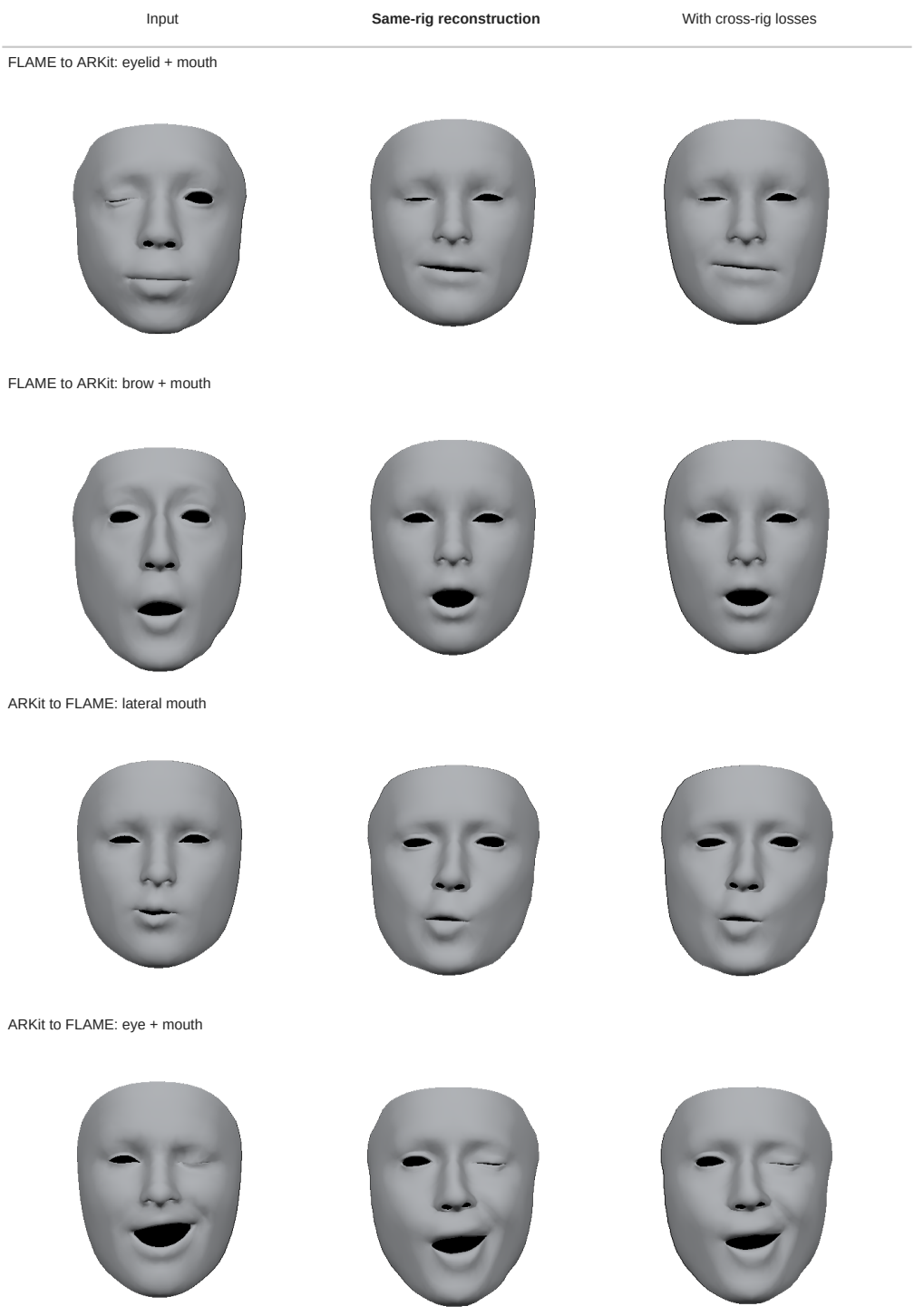

Figure 8: **Selected Test expressions under the two predictor-training strategies.** Each row fixes the source expression and target geometry. Same-Rig Reconstruction is the main model; the comparison adds cross-rig expression-consistency and source-recovery losses during predictor refinement. The upper rows transfer FLAME to ARKit, and the lower rows transfer ARKit to FLAME.

A.9 Additional Qualitative Results

We provide six selected Test cases with the evaluated baselines, including Direct MLP, in each direction. Each row is one expression, and method names appear once as shared column headers. Input is first and RetargetingFace (Ours) is second. Within each comparison, all methods share the target geometry, camera, lighting, material, background, orientation, and crop.

FLAME→ARKit. These cases cover mouth opening, a relatively closed mouth contour, and combined eyelid–mouth motion.

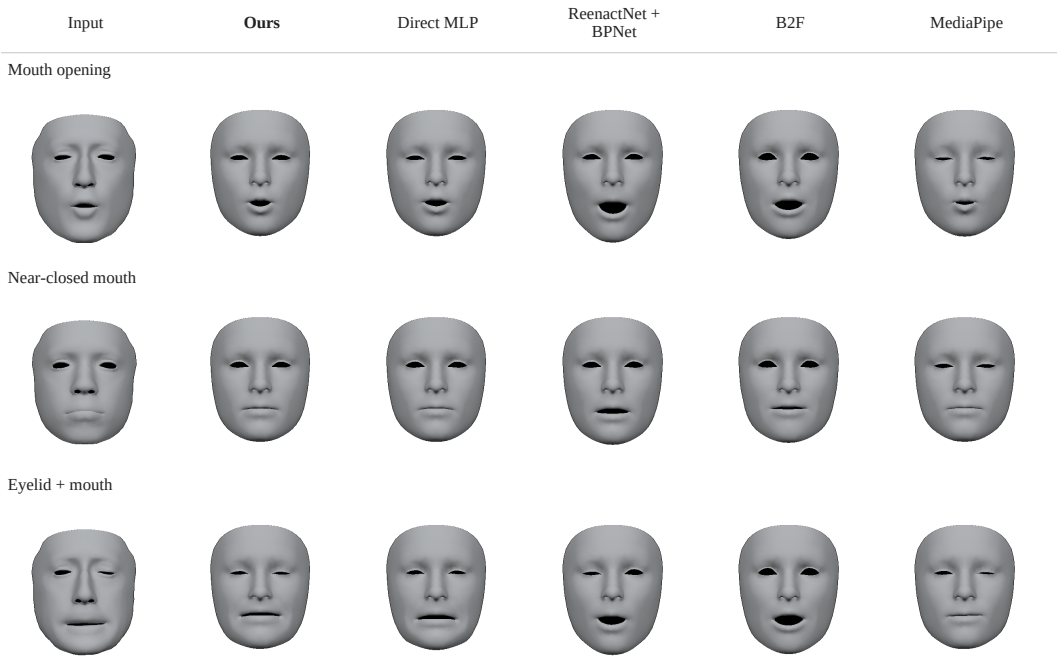

Figure 9: **Additional FLAME→ARKit retargeting.** From top to bottom: near-symmetric mouth opening, a near-closed mouth contour, and combined eyelid–mouth motion. All predictions use the same Said target. The final row uses Test FLAME `shape_09` (state 083, variant 030, strength 1.0). The evaluated baselines, including Direct MLP, are shown.

ARKit→FLAME. These cases cover unilateral eye closure with a near-closed mouth, near-symmetric mouth opening, and a broad open-mouth smile.

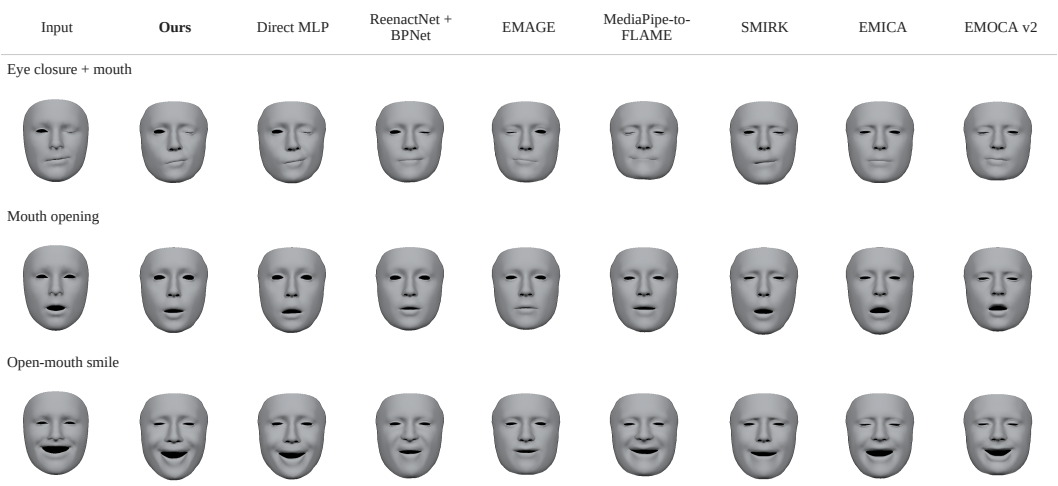

Figure 10: **Additional ARKit→FLAME retargeting.** From top to bottom: unilateral eye closure with a near-closed mouth, near-symmetric mouth opening, and a broad open-mouth smile. All inputs are Said renders; predictions use canonical-zero FLAME and the common 1,787-vertex face mask. The smile uses state 179, variant 002, strength 1.0. The evaluated baselines, including Direct MLP, are shown.