VIDEOFACE: LEARNING A DISCRETE MOTION SPACE FOR SPARSE-TO-DENSE FACIAL MOTION RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing facial motion from sparse visual keyframes requires visual estimation under appearance changes and temporal completion between observations. Independently estimated states and subsequent interpolation do not explicitly model how an expression develops, including its relative intensity, regional coordination, and timing. We present VideoFace, a framework that combines visual conditioning from a pretrained vision-language model with a learned discrete motion prior. Motivated by the potential of semantic representations to support robustness across appearances, the model uses keyframe images, their temporal indices, and motion history to predict a complete motion-token sequence of a specified length. A dual-branch tokenizer connects ARKit and FLAME through shared finite scalar quantization, assigning one token per frame. Self- and cross-reconstruction preserve motion content, while window-level code matching and frame-level alignment encourage correspondence between the two representations. After tokenizer training, we freeze its parameters and fine-tune the vision-language model for autoregressive token prediction; either decoder then produces the corresponding continuous trajectory. We also construct MAMC-Face, comprising 10 hours of facial motion at 30 Hz rendered across 65 avatars under moving camera views. Its paired visual–motion supervision supports learning complete motion trajectories from sparse visual observations across varied appearances and viewpoints.

1 INTRODUCTION

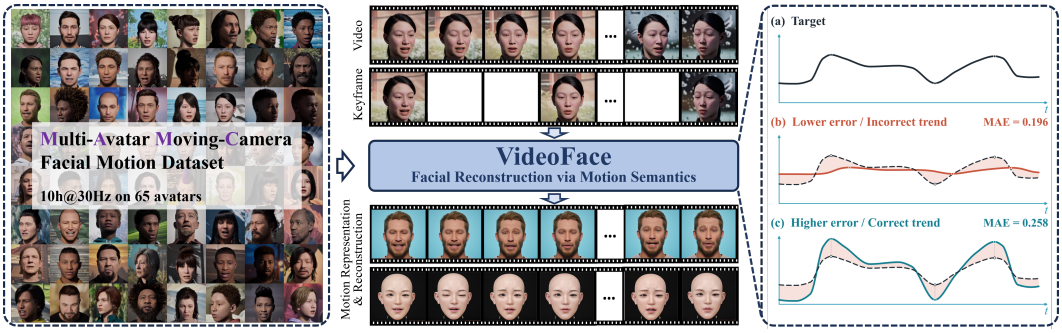

Figure 1: VideoFace reconstructs complete facial motion from sparse visual keyframes, using MAMC-Face for visual–motion supervision. Expression quality depends on motion amplitude, onset–hold–release timing, and coordinated relative changes across facial regions.

Recovering 3D facial motion from video supports digital humans, character animation, and facial interaction. These applications require trajectories that capture visible facial configurations together with the way expressions unfold over time. A smile, for example, involves changes in intensity and coordination across facial regions as it emerges, is held, and subsides. We study the recovery of such trajectories from sparse visual keyframes: given facial images sampled from a video, their frame

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.

indices, and a target sequence length, predict a complete sequence of facial motion parameters for subsequent animation. Using every frame as an observation is a special case of this setting. Under sparse sampling, however, the images provide only partial evidence about the intervening motion, making temporal reconstruction central to the task.

A straightforward pipeline first estimates facial parameters independently at the observed frames, interpolates the missing positions, and applies temporal filtering to reduce jitter before animation. Image-based methods such as DECA (Feng et al., 2021) and EMOCA (Daněček et al., 2022) provide useful building blocks for the estimation step, recovering animatable geometry and improving expression fidelity through perceptual supervision. Even accurate estimates at the observed positions leave open how the face should move between them.

The distinction is between matching sampled facial states and preserving the dynamics of a performance. Expression quality depends on motion amplitude as well as the timing of onset, sustained activation, and release, including the relative changes and coordination of different facial regions. The same initial and final configurations can be connected by trajectories with different speeds, peak timings, and coordination between the eyes and mouth. Interpolation connects the estimated states according to a prescribed rule, without learning which transitions are consistent with the visible expression and motion context. Filtering can suppress fluctuations while also attenuating rapid movements and blurring their timing. Thus, low error at the observed frames or a smooth output trajectory is insufficient by itself to establish dynamic fidelity. The task calls for both consistency with the observations and a learned model of plausible expression development in the missing intervals, where multiple trajectories may fit the same evidence.

Existing video methods improve facial reconstruction through cross-frame constraints, temporal context, and perceptual supervision (Tewari et al., 2019; Filntisis et al., 2023; Kuang et al., 2024). SemanticFace (Kang et al., 2026b) suggests that semantic supervision can support facial-action estimation across appearances. Meanwhile, CodeTalker (Xing et al., 2023) and TDMM-LM (Song et al., 2026) establish discrete motion representations as an interface for sequence modeling. Building on these directions, we combine visual conditioning from a pretrained vision-language model (VLM) with motion-token history to predict expression development between sparse observations. This formulation aims to preserve relative intensity, regional coordination, and timing throughout the reconstructed trajectory.

Learning this mapping requires continuous motion trajectories paired with their visual realizations. We construct MAMC-Face (Multi-Avatar Moving-Camera Facial Motion Dataset) by using motion extracted from human performances to drive multiple avatars and rendering image sequences under moving camera views, with variation in background and lighting. The corpus comprises 10 hours of facial motion at 30 Hz rendered across 65 avatars. We retain the actual driving ARKit controls as labels, allowing the same motion to appear under different visual conditions with corresponding control-space supervision. The resulting visual–motion pairs support training on sparse image observations with complete target trajectories across varied appearances and viewpoints.

We propose **VideoFace**, which connects this supervision to a VLM through a unified motion-token interface. Downstream animation uses both interpretable ARKit controls and geometric FLAME parameters (Li et al., 2017), motivating an interface that supports either output without separate token generators. A dual-branch tokenizer learns from paired trajectories using separate encoders and decoders with shared finite scalar quantization (FSQ). Self- and cross-reconstruction preserve decodable motion content, while window-level code matching (CMCM) and frame-level alignment (MMS) encourage correspondence across representations. The resulting discrete sequences retain one token per frame. We then freeze the tokenizer and train the VLM on sparse rendered keyframes paired with complete target motion-token sequences. Conditioned on all supplied images, their temporal positions, the requested length, and preceding motion tokens, the VLM autoregressively predicts the full sequence. The corresponding decoder recovers continuous ARKit or FLAME parameters. This design brings visual context and learned motion structure together to model how a facial performance unfolds between observations.

Our contributions are as follows:

- We introduce VideoFace, a framework for recovering complete facial motion trajectories from sparse visual keyframes by conditioning sequence generation on visual context, temporal positions, and motion history.

- We develop a dual-branch ARKit–FLAME motion tokenizer that provides a shared discrete interface for visual-conditioned generation and decoding into either representation through cross-reconstruction, window-level code matching, and frame-level alignment.
- We construct MAMC-Face, pairing multi-avatar renderings under moving camera views and varied visual conditions with their driving trajectories to support learning complete facial motion from sparse visual observations.

2 Related Work

Facial Motion Representations. Parametric representations provide compact interfaces for facial reconstruction and animation. Classical 3D Morphable Models (3DMMs) (Blanz & Vetter, 1999) describe shape and appearance through statistical bases, while FLAME (Li et al., 2017) combines identity and expression spaces with articulated jaw, neck, and eye motion. Complementing these geometric models, FACS (Ekman & Friesen, 1978) describes anatomically grounded facial action units, and ARKit (Apple, 2023) exposes named blendshape controls for localized movements. We connect ARKit's interpretable controls and FLAME's geometric representation through a shared motion-token interface.

3D Facial Reconstruction. Single-image methods recover facial states from visual appearance: DECA (Feng et al., 2021) estimates FLAME parameters and animatable surface details, while EMOCA (Daněček et al., 2022) introduces emotion-feature consistency to improve expression fidelity. Multi-frame methods exploit complementary temporal cues. FML (Tewari et al., 2019) shares identity-related shape and reflectance across frames; SPECTRE (Filntisis et al., 2023) combines temporal context with lipreading and expression supervision; and AU-aware reconstruction (Kuang et al., 2024) models temporal dependencies and spatial action-unit correlations. These methods motivate incorporating cross-frame context and perceptual information. Our focus is complete motion reconstruction from sparse visual keyframes, which additionally requires inferring expression development between observations.

Language Models for Motion Modeling. Discrete motion vocabularies connect continuous movement with language-model sequence prediction. CodeTalker (Xing et al., 2023) predicts facial motion codes from speech, and MotionGPT (Jiang et al., 2023) unifies motion and language for generation and understanding. The Language of Motion (Chen et al., 2025) integrates text, audio, and compositional face, hand, and body tokens, while TDMM-LM (Song et al., 2026) connects FLAME motion with language through geometry-aware quantization. Semantic conditioning provides complementary expressive context: SemanticFace (Kang et al., 2026b) distills global expression and local action descriptions into an ARKit predictor, and the listening-head framework of Lai et al. (2025) combines conversational cues with motion history to generate facial responses. Building on these ideas, VideoFace conditions shared ARKit–FLAME motion tokens on keyframe images, temporal indices, and preceding motion, using visual-semantic context and a learned motion prior to reconstruct complete trajectories.

3 Method

To train a vision-language model that reconstructs complete 3D facial motion from sparse keyframes, we construct MAMC-Face by rendering multiple avatars under moving camera views and pairing the resulting image sequences with their driving motion coefficients. We then train a dual-branch tokenizer on these motion trajectories to represent ARKit and FLAME motion in a shared, compact token space. Using the resulting visual–token pairs, we fine-tune the VLM to predict complete motion-token sequences from sparse keyframes and preceding motion.

3.1 Unified Motion Tokenizer

Shared temporal vocabulary. To provide the VLM with a common output vocabulary for facial motion, we map paired ARKit and FLAME trajectories $x^A_{1:N}$ and $x^F_{1:N}$ into a shared discrete space. Representation-specific causal encoders E_r and decoders D_s are coupled through a shared finite

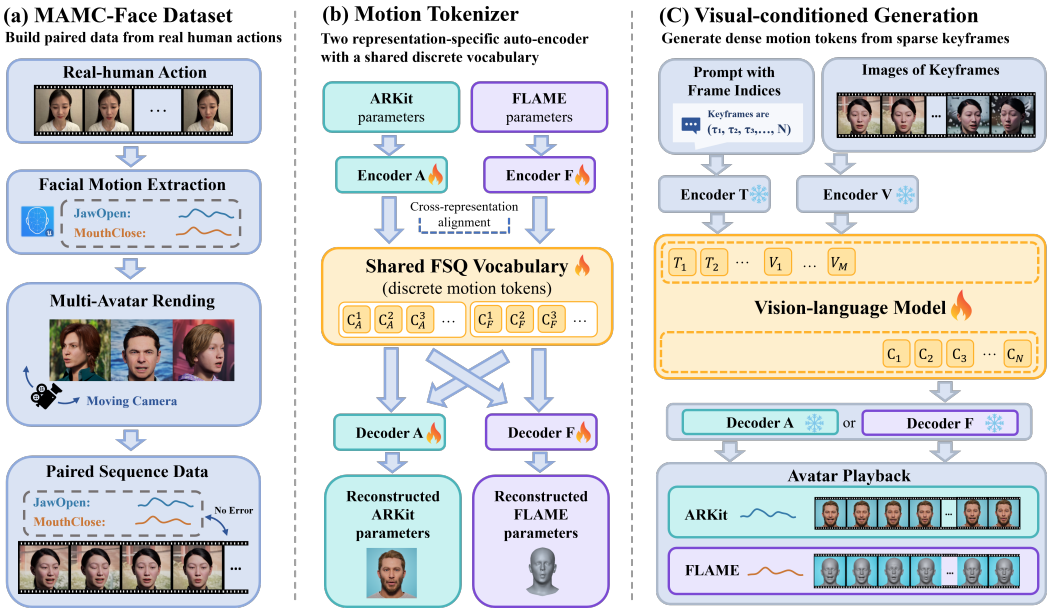

Figure 2: Overview of VideoFace. (a) MAMC-Face pairs multi-avatar videos rendered under moving camera views with their driving facial motion. (b) A dual-branch tokenizer learns a shared discrete motion space for ARKit and FLAME through finite scalar quantization and cross-representation alignment. (c) A vision-language model predicts dense motion-token sequences from sparse visual keyframes and their temporal indices, which are decoded into continuous ARKit or FLAME motion trajectories.

scalar quantizer (FSQ) (Mentzer et al., 2023):

$$z_t^r = E_r(x_{1:t}^r), \quad q_t^r = Q(z_t^r) = e_{c_t^r}, \quad \hat{x}_t^{r\to s} = D_s(q_{1:t}^r), \quad r, s \in \{A, F\}. \tag{1}$$

Here, c_t^r denotes a discrete motion token index, and $q_t^r = e_{c_t^r}$ is its corresponding quantized vector on the shared FSQ grid of K codes. Encoding preserves the temporal resolution, producing one token per frame. Each token can therefore encode preceding motion while retaining an explicit position on the output timeline, allowing the generator to predict a complete trajectory of a specified length.

Cross-representation learning. We train self- and cross-reconstruction so that tokens from either representation preserve motion content for both decoders, using Smooth-L1 losses on motion coefficients and their first-order temporal differences. Reconstruction alone, however, does not directly enforce compatible code assignments across representations. We therefore align the two branches at complementary scales: cross-modal code matching (CMCM) (Liu et al., 2022) encourages paired windows to share code-usage patterns, while frame-level matching, additionally constrains correspondence at paired time steps. Both operate on bounded continuous coordinates $u_t^r = \mathcal{B}(z_t^r)$ in the FSQ grid's coordinate system, linking alignment to the space that determines code selection. For a batch of B paired motion windows, the alignment losses are

$$\begin{aligned}
\mathcal{L}_{\text{CMCM}} &= -\frac{1}{B}\sum_{b=1}^{B} \log \frac{\exp(C_{bb}/\tau_c)}{\sum_{j=1}^{B}\exp(C_{bj}/\tau_c)}, \\
\mathcal{L}_{\text{MMS}} &= -\frac{1}{2|\mathcal{I}|}\sum_{(b,t)\in\mathcal{I}} \left[\log \frac{\exp(G_{bb,t})}{\sum_{j\in\mathcal{V}_t}\exp(G_{bj,t})} + \log \frac{\exp(G_{bb,t})}{\sum_{j\in\mathcal{V}_t}\exp(G_{jb,t})}\right].
\end{aligned} \tag{2}$$

Here, C_{bj} compares the window-averaged soft code distributions of ARKit window b and FLAME window j using negative symmetric cross-entropy, with contrastive temperature τ_c. The frame-matching logit $G_{bj,t}$ uses temperature-scaled negative squared distance between their bounded FSQ coordinates, with a margin subtracted for positive pairs. Paired clips ($b = j$) form positives, and

other clips form negatives. The sets $\mathcal{V}_t$ and $\mathcal{I}$ contain valid candidate clips at position t and eligible paired frame anchors, respectively. The two MMS terms match ARKit to FLAME and vice versa. Complete score constructions and validity criteria are given in Appendix B.

Training and generation interface. Combining the self- and cross-reconstruction losses with both alignment objectives gives

$$\mathcal{L}_{\text{tok}} = \mathcal{L}_{\text{self}} + \lambda_c \mathcal{L}_{\text{cross}} + \lambda_{\text{CMCM}} \mathcal{L}_{\text{CMCM}} + \lambda_{\text{MMS}} \mathcal{L}_{\text{MMS}}. \tag{3}$$

The resulting ARKit token sequences supervise the visual-conditioned generator in Sec. 3.2; generated indices select shared grid vectors that either decoder maps to its corresponding motion trajectory.

3.2 Visual-Conditioned Motion Generation

Given a clip of N frames at 30 Hz, we observe M keyframe images $I_{1:M}$ at frame indices $1 \leq \tau_1 < \cdots < \tau_M \leq N$. The target is the complete N-frame motion trajectory, including the unobserved intervals.

The facial generator uses a pretrained vision-language model to predict motion in the shared token space. The motion tokens are registered as new special tokens in VLM. The keyframe images are encoded into visual tokens $v_{1:M}$ and the input prompt specifies the keyframe indices $\tau_{1:M}$ and target sequence length N, allowing the model to associate each visual observation with its position in the output trajectory. The VLM uses the keyframes to form visual context of facial movements and combines this context with previously generated motion token to predict the entire temporal progression of expression sequence.

Writing $\mathcal{O} = \{(I_i, \tau_i)\}_{i=1}^{M}$ for the observations, we model

$$p_\theta(c_{1:N} \mid \mathcal{O}, N) = \prod_{t=1}^{N} p_\theta(c_t \mid v_{1:M}, \tau_{1:M}, N, c_{1:t-1}) \,. \tag{4}$$

The model conditions on all supplied keyframes, including those after the current output position. When $M = N$ and $\tau_i = i$, every target frame is observed. When $M < N$, the same formulation also generates the unobserved intervals. Visual-semantic context is intended to guide how an expression develops across these intervals, while the token history provides the evolving motion state.

During VLM training, the encoder and decoder in motion tokenizer is frozen. Teacher forcing supplies the preceding target tokens, and the loss is applied only to motion-token predictions:

$$\mathcal{L}_{\text{VLM}} = -\frac{1}{N} \sum_{t=1}^{N} \log p_\theta\big(c_t^A \mid v_{1:M}, \tau_{1:M}, N, c_{1:t-1}^A\big) \,. \tag{5}$$

At inference, N generated motion tokens can be passed to either frozen decoder:

$$\hat{x}_t^A = D_A(e_{\hat{c}_1}, \ldots, e_{\hat{c}_t}), \qquad \hat{x}_t^F = D_F(e_{\hat{c}_1}, \ldots, e_{\hat{c}_t}). \tag{6}$$

The decoder reconstructs a complete facial motion trajectory at 30 Hz, which is used to render digital avatars and control robots without additional temporal filtering or smoothing.

3.3 MAMC-Face Dataset

Learning visual-conditioned facial motion requires reliable image–parameter correspondence and complete temporal trajectories. Motion estimated from captured videos provides pseudo-labels that may deviate from visible expressions (Kang et al., 2026a), while high-fidelity 3D capture can require specialized equipment and substantial acquisition effort. SemanticFace demonstrates robust VLM-based facial action estimation under challenging imaging conditions and generalization to stylized faces (Kang et al., 2026b), motivating our exploration of rendered avatars as a source of visual–motion supervision. As summarized in Table 1, our task calls for temporal sequences with known driving controls and visual variation across identities and viewpoints.

Table 1: Dataset comparison for visual-conditioned facial motion learning.

Dataset	IDs	Scale	Temporal data	View diversity	ARKit controls	Render labels	Cross-avatar trajectories
MEAD(Wang et al., 2020)	60	281,400 clips	✓	✓	×	×	×
CelebV-HQ(Zhu et al., 2022)	15,653	35,666 clips	✓	✓	×	×	×
Multiface(Wuu et al., 2022)	13	12.2k / 23k frames per subject	✓	✓	×	×	×
MultiREX(Josi et al., 2025)	8	49k images	✓	✓	×	×	×
Express4D(Aloni et al., 2025)	18	1,205 sequences	✓	×	✓	×	×
SEREP synthetic(Josi et al., 2025)	NR	135k images	×	✓	×	✓	×
MAMC-Face (ours)	**65**	**864K motion frames 5.70M images**	✓	✓	✓	✓	✓

We therefore construct MAMC-Face (Multi-Avatar Moving-Camera Facial Motion Dataset), comprising 10 hours of facial motion at 30 Hz rendered across 65 avatars. We extract ARKit motion parameters from human performances and refine the driving trajectories to improve the naturalness and temporal coherence of the resulting animations. Each trajectory is reused across multiple avatars and rendered under moving camera views, providing different visual realizations of the same motion. We retain the final parameters supplied to the rendering rig as labels, establishing direct image–control correspondence without re-estimating motion from the rendered images. Complete trajectories support motion tokenizer training, while sparse keyframes sampled from the rendered sequences are paired with complete target trajectories to train the visual-conditioned generator.

4 Experiments

4.1 Experimental Setup

Dataset We conduct our experiments on our MAMC-Face dataset, which contains 10 hours of facial motion sampled at 30 Hz. Each motion sequence is rendered using at least six avatars, with variations in camera viewpoint and environmental lighting. We use 8 hours of motion for training, 1 hour for validation, and 1 hour for testing.

Baseline We compare our method with five advanced 3D facial reconstruction approaches: MediaPipe (Lugaresi et al., 2019), SemanticFace (Kang et al., 2026b), SPECTRE (Filntisis et al., 2023), TEASER (Liu et al., 2025), and OFER (Selvaraju et al., 2025). Among these methods, MediaPipe and SemanticFace support the ARKit blendshape representation, whereas SPECTRE, TEASER, and OFER do not natively support it. To enable comparison in the same representation space, we retrain SPECTRE, TEASER, and OFER on MAMC-Face Dataset to predict ARKit blendshape coefficients.

Metric We evaluate facial motion generation at three complementary levels. *Frame-wise reconstruction accuracy* is measured by MAE, MSE, and Regional Mean Error (RME). *Sequence-level statistical consistency* is assessed using FFD, which compares per-coefficient temporal standard deviations, and Fréchet Distance (FD), which compares Gaussian fits to the predicted and ground-truth coefficient distributions. *Image–motion correspondence* is evaluated using R-Precision at rank 1 (R@1) with 32 candidates and Multimodal Distance (MMD), the mean Euclidean distance between paired, normalized image and motion embeddings. Detailed definitions of all metrics are provided in Appendix A. For the user study, ten participants independently rate reconstruction accuracy (RA), smoothness (SM), and motion diversity (MD) on a 1–5 scale, with higher scores indicating better quality. Responses marked as unable to judge are excluded from the mean scores.

Implementation Our motion tokenizer uses a dual-branch causal encoder–decoder with a hidden width of 256 and shared finite scalar quantization (FSQ), yielding 1,000 motion codes and one token per frame at 30 Hz. We adopt Qwen3.5-4B as the backbone and use LoRA (rank = 16) for all VLM training stages. The pretrained tokenizer remains frozen, while its motion codes are added as special tokens with trainable VLM embeddings and an output head. The keyframe-training stage is configured for 5 epochs in BF16 with a learning rate of 10^{-4}.

Table 2: Quantitative and qualitative comparison under dense continuous frames.

	Frame-wise accuracy			Sequence-level consistency		Image–motion correspondence		User Study (Qualitative)		
Method	MAE ↓	MSE ↓	RME ↓	FFD ↓	FD ↓	R@1 ↑	MMD ↓	RA ↑	SM ↑	MD ↑
MediaPipe(Lugaresi et al., 2019)	0.1007	0.0301	0.6379	0.0453	0.9053	0.0257	1.4282	2.269	2.265	2.735
SemanticFace(Kang et al., 2026b)	0.0567	0.0114	0.2483	0.0314	0.4776	0.1170	1.3780	3.359	2.524	3.273
SPECTRE(Filntisis et al., 2023)	0.0667	0.0123	0.3639	0.0290	0.5186	0.1146	1.3831	3.065	2.666	3.125
TEASER(Liu et al., 2025)	0.0590	0.0102	0.3597	0.0263	0.4494	0.2499	1.3064	3.562	3.814	3.618
OFER(Selvaraju et al., 2025)	0.0734	0.0133	0.4353	0.0298	0.5523	0.0572	1.4187	3.052	2.614	2.151
Ours	**0.0309**	**0.0025**	**0.2278**	**0.0166**	**0.1034**	**0.4310**	**1.1594**	**3.967**	**4.472**	**4.215**

Table 3: Quantitative and qualitative comparison under sparse keyframes.

	Frame-wise accuracy			Sequence-level consistency		Image–motion correspondence		User Study (Qualitative)		
Method	MAE ↓	MSE ↓	RME ↓	FFD ↓	FD ↓	R@1 ↑	MMD ↓	RA ↑	SM ↑	MD ↑
MediaPipe(Lugaresi et al., 2019)	0.0963	0.0252	0.7431	0.0511	1.0208	0.0090	1.4603	1.836	1.614	1.628
SemanticFace(Kang et al., 2026b)	0.0676	0.0126	0.2805	0.0355	0.4974	0.0156	1.3921	2.307	1.451	2.735
SPECTRE(Filntisis et al., 2023)	0.0686	0.0135	0.4584	0.0373	0.5343	0.0325	1.4171	1.925	1.623	1.925
TEASER(Liu et al., 2025)	0.0687	0.0117	0.4535	0.0342	0.4738	0.0762	1.3377	2.696	2.661	2.265
OFER(Selvaraju et al., 2025)	0.0840	0.0149	0.5142	0.0328	0.5754	0.0312	1.4861	1.664	1.315	1.625
Ours	**0.0355**	**0.0036**	**0.2470**	**0.0141**	**0.1141**	**0.3906**	**1.1855**	**4.004**	**4.271**	**4.018**

4.2 Visual-Conditioned Facial Motion Reconstruction

Under dense visual observations, VideoFace models inter-frame relationships by conditioning motion prediction on visual context and preceding motion tokens. This formulation allows the model to use information across frames when reconstructing facial motion. As shown in Table 2, VideoFace achieves lower frame-wise reconstruction errors, together with better sequence-level statistical consistency and stronger image–motion correspondence. These results support the effectiveness of incorporating temporal context into facial motion reconstruction.

Under sparse visual observations, several baseline methods show substantial deterioration, particularly in image–motion correspondence and perceptual ratings, as reported in Table 3. VideoFace maintains comparatively stable performance and continues to achieve strong results across reconstruction accuracy, sequence-level consistency, and image–motion correspondence. Its high user-study scores further suggest that the reconstructed sequences remain perceptually smooth and expressive despite the reduced visual input.

Qualitative comparison. Figure 3 compares reconstructed expressions at three aligned time instants. VideoFace captures the open-mouth smile in row (a) and the raised-brow, open-mouth configurations in rows (b) and (c), while the baseline outputs exhibit differences in mouth opening, brow position, and eyelid configuration. All predictions are displayed on the same target avatar, enabling comparison of local facial actions across methods despite the appearance and viewpoint differences from the reference.

4.3 Motion Tokenizer Evaluation

Cross-representation alignment. Our dual-branch tokenizer aligns ARKit and FLAME facial representations in a shared motion space through the CMCM and MMS alignment losses. Figure 4(a) visualizes the bounded pre-quantization FSQ features with and without alignment. Without these losses, the two representations tend to occupy distinct regions, indicating a tendency toward modality-specific codebook partitioning. With alignment, their feature distributions exhibit greater overlap. Consistently, the cross-modal code-usage overlap increases from 0.258 to 0.675 on the held-out validation windows, indicating that the two branches increasingly share the discrete vocabulary. These results support more compatible code usage across facial representations.

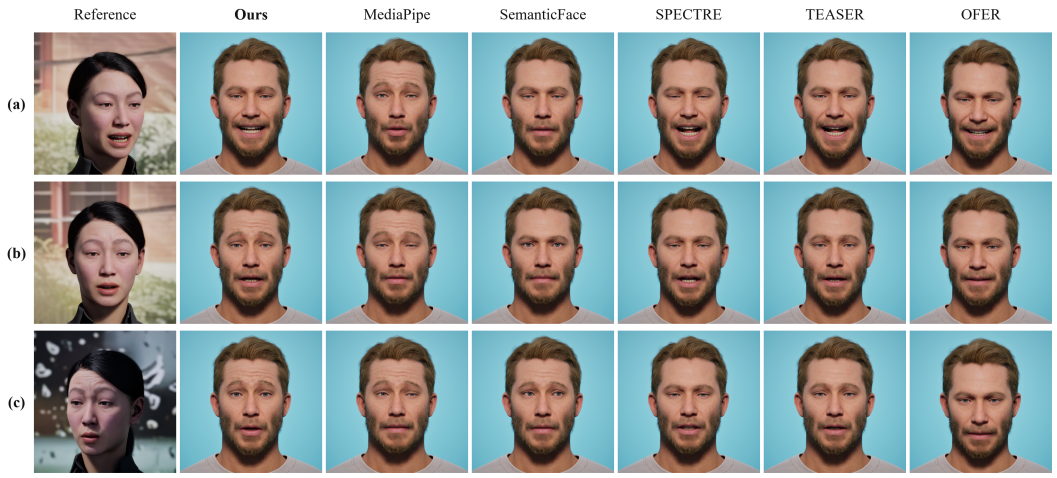

Figure 3: Qualitative comparison of facial motion reconstruction. Columns show the reference, VideoFace (Ours), MediaPipe, SemanticFace, SPECTRE, TEASER, and OFER. Predictions are aligned to the same source frames and rendered on a common target avatar; the reference retains the source avatar, viewpoint, and background.

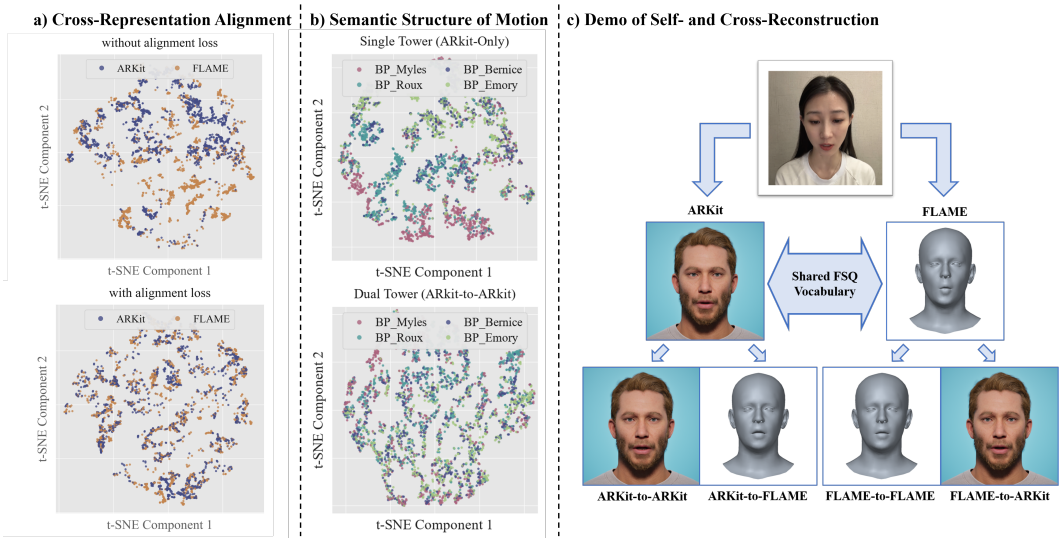

Figure 4: Motion tokenizer evaluation. (a) Cross-representation alignment of ARKit and FLAME features with and without the alignment losses. (b) Semantic structure of motion in the ARKit-only and dual-branch tokenizers, visualized using four sequences of similar facial motions. (c) Self- and cross-reconstruction between ARKit and FLAME through the shared FSQ vocabulary.

Motion semantics. By aligning ARKit and FLAME through cross-reconstruction, our tokenizer enables motion conversion between datasets using these representations and encourages the shared codebook to better capture motion semantics. To examine this semantic structure, Fig. 4(b) compares four sequences of similar facial motions. The ARKit-only tokenizer exhibits visible identity-dependent grouping, consistent with systematic differences in their motion coefficients. With ARKit–FLAME alignment, similar motions from different avatars show greater local overlap, suggesting that the token space increasingly reflects shared motion semantics. This semantic organization encourages compatible token representations for similar motions and can provide more consistent supervision for the VLM to learn the correspondence between visual observations and motion semantics across avatars. Figure 4(c) illustrates self- and cross-reconstruction through all

Table 4: Comparison of direct numerical text prediction and motion-token prediction for facial motion sequences under dense and sparse visual inputs. Incorrect output lengths and collapse to a constant mean configuration are counted as invalid inference.

Method	Dense frames			Sparse frames		
	Valid inference (%) ↑	MAE ↓	MSE ↓	Valid inference (%) ↑	MAE ↓	MSE ↓
Text-based	96.0	0.047	0.0172	10.5	0.115	0.0320
Text-based + linear interpolation	96.0	0.047	0.0172	87.0	0.078	0.0189
Motion token-based	**100.0**	**0.031**	**0.0025**	**100.0**	**0.035**	**0.0036**

four decoding paths, demonstrating how the shared vocabulary supports motion reconstruction in either parameterization.

4.4 Ablation Studies

Text-based vs. motion-token prediction. Encoding and decoding motion through a tokenizer introduces reconstruction error, motivating a comparison with direct parameter prediction. Our previous work explored predicting motion parameters as numerical text under semantic guidance. We extend this approach to video inputs and compare it with motion-token prediction under dense and sparse visual observations. The text-based variant directly generates sequences of numerical motion parameters, while the motion-token variant predicts discrete codes that are decoded into continuous trajectories. Table 4 reports the valid inference rate and mean absolute error (MAE), together with auxiliary mean squared error (MSE) results from separate archived runs.

Text-based prediction remains usable with dense frames but deteriorates sharply under sparse observations, producing invalid outputs with incorrect frame counts or trajectories that collapse to a constant mean configuration. We hypothesize that visually similar frames introduce ambiguity in temporal correspondence, while the LLM struggles to preserve fine-grained numerical relationships when generating long parameter sequences. Motion tokens offer a compact representation of facial configurations and their temporal context. These results support their use for motion reconstruction, particularly with sparse observations, despite the reconstruction error introduced by the tokenizer.

5 Conclusion

We presented VideoFace, a framework for sparse-to-dense facial motion reconstruction through a learned discrete motion space. We first constructed MAMC-Face by pairing human-derived motion trajectories with multi-avatar videos rendered under moving camera views. We then trained a dual-branch ARKit–FLAME tokenizer that encourages motion tokens to preserve shared facial action semantics while reducing identity-dependent variation. With the tokenizer frozen, we fine-tuned a vision-language model to learn temporal dependencies in this space and predict complete motion-token sequences from sparse visual keyframes. Decoding these tokens produces dense, continuous trajectories that directly drive facial animation without additional temporal filtering or smoothing. Experiments on MAMC-Face demonstrate the lowest reconstruction errors among the evaluated baselines, together with stronger sequence-level consistency and higher perceptual ratings for smoothness and motion diversity. These results highlight the value of learning motion structure across representations for recovering coherent expression dynamics and preserving the expressive information conveyed by motion trends under sparse visual observations.

AI USE STATEMENT

We used generative AI tools to assist with the design and refinement of research methodology and experiments, method implementation, result analysis and interpretation, and data cleaning or reformatting. We also used these tools to summarize and analyze existing literature, improve manuscript wording and organization, and prepare or revise pseudocode, citation formatting, and figure captions.

We did not use generative AI tools to develop theoretical models or conceptual frameworks, formulate or refine research hypotheses, generate synthetic datasets, translate text, or conduct qualitative or thematic data analysis. We also did not use these tools to identify research topics or gaps or for brainstorming. AI assistance in formulating mathematical claims, developing key proof arguments, or writing proofs was not applicable to this work.

We manually reviewed and revised AI-assisted code and checked it through execution, debugging, or testing. We inspected processed data and checked result interpretations against experimental outputs, metrics, or figures. For literature and manuscript preparation, we checked AI-assisted literature summaries against the original sources, reviewed and revised manuscript wording and organization, checked citation formatting and sources, and verified figure captions against the figures. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Yaron Aloni, Rotem Shalev-Arkushin, Yonatan Shafir, Guy Tevet, Ohad Fried, and Amit Haim Bermano. Express4D: Expressive, friendly, and extensible 4D facial motion generation benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 485–494, 2025.

Apple. Arkit. `https://developer.apple.com/documentation/arkit`, 2023. URL `https://developer.apple.com/documentation/arkit`. Apple Developer Documentation.

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.

Changan Chen, Juze Zhang, Shrinidhi K. Lakshmikanth, Yusu Fang, Ruizhi Shao, Gordon Wetzstein, Li Fei-Fei, and Ehsan Adeli. The language of motion: Unifying verbal and non-verbal language of 3D human motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6200–6211, 2025. URL `https://openaccess.thecvf.com/content/CVPR2025/html/Chen_The_Language_of_Motion_Unifying_Verbal_and_Non-verbal_Language_of_CVPR_2025_paper.html`.

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.

Paul Ekman and Wallace V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.

Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.

Panagiotis P Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. SPECTRE: Visual speech-informed perceptual 3D facial expression reconstruction from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 5745–5755, 2023. URL `https://openaccess.thecvf.com/content/CVPR2023W/ABAW/html/Filntisis_SPECTRE_Visual_Speech-Informed_Perceptual_3D_Facial_Expression_Reconstruction_From_Videos_CVPRW_2023_paper.html`.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. MotionGPT: Human motion as a foreign language. In *Advances in Neural Information Processing Systems*, volume 36, pp. 20067–20079, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3fbf0c1ea0716c03dea93bb6be78dd6f-Paper-Conference.pdf.

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

Zejian Kang, Xuanyang Xu, Wentao Yang, Kai Zheng, Yuanchen Fei, Hongyuan Zou, Hui Shan, Shuo Yang, and Xiangru Huang. Superface: Preference-aligned facial expression estimation beyond pseudo supervision. *arXiv preprint arXiv:2605.06179*, 2026a.

Zejian Kang, Kai Zheng, Yuanchen Fei, Wentao Yang, Hongyuan Zou, and Xiangru Huang. Semanticface: Semantic facial action estimation via semantic distillation in interpretable space. *arXiv preprint arXiv:2603.14827*, 2026b. URL https://arxiv.org/abs/2603.14827.

Chenyi Kuang, Jeffrey O Kephart, and Qiang Ji. Au-aware dynamic 3d face reconstruction from videos with transformer. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 6237–6247, 2024.

Peiwen Lai, Weizhi Zhong, Yipeng Qin, Xiaohang Ren, Baoyuan Wang, and Guanbin Li. Llm-driven multimodal and multi-identity listening head generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10656–10666, 2025.

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

Alexander Liu, SouYoung Jin, Cheng-I Lai, Andrew Rouditchenko, Aude Oliva, and James Glass. Cross-modal discrete representation learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3013–3035. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.215. URL https://aclanthology.org/2022.acl-long.215/.

Yunfei Liu, Lei Zhu, Lijian Lin, Ye Zhu, Ailing Zhang, and Yu Li. TEASER: Token enhanced spatial modeling for expressions reconstruction. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pTeOOKnjGM.

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

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. *arXiv preprint arXiv:2309.15505*, 2023. URL https://arxiv.org/abs/2309.15505.

Pratheba Selvaraju, Victoria Fernandez Abrevaya, Timo Bolkart, Rick Akkerman, Tianyu Ding, Faezeh Amjadi, and Ilya Zharkov. OFER: Occluded face expression reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26985–26995, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Selvaraju_OFER_Occluded_Face_Expression_Reconstruction_CVPR_2025_paper.html.

Luchuan Song, Pinxin Liu, Haiyang Liu, Zhenchao Jin, Yolo Yunlong Tang, Zichong Xu, Susan Liang, Jing Bi, Jason J Corso, and Chenliang Xu. Tdmm-lm: Bridging facial understanding and animation via language models. *arXiv preprint arXiv:2603.16936*, 2026.

Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. FML: Face model learning from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://vcai.mpi-inf.mpg.de/projects/FML19/.

Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. MEAD: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, 2020. URL https://github.com/uniBruce/Mead.

Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Xuhua Huang, et al. Multiface: A dataset for neural face rendering. *arXiv preprint arXiv:2207.11243*, 2022. URL https://github.com/facebookresearch/multiface.

Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12780–12790, 2023.

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *European Conference on Computer Vision*, 2022. URL https://celebv-hq.github.io/.

A Evaluation Metrics

Let $\hat{\mathbf{b}}_t, \mathbf{b}_t \in \mathbb{R}^D$ denote the predicted and ground-truth expression coefficients at frame t, respectively and T is the number of aligned evaluation frames.

A.1 Frame-wise Reconstruction Accuracy

MAE, MSE, and Regional Mean Error. MAE and MSE measure coefficient reconstruction errors across all frames and expression dimensions:

$$\mathrm{MAE} = \frac{1}{TD}\sum_{t=1}^{T}\|\hat{\mathbf{b}}_t - \mathbf{b}_t\|_1, \qquad \mathrm{MSE} = \frac{1}{TD}\sum_{t=1}^{T}\|\hat{\mathbf{b}}_t - \mathbf{b}_t\|_2^2. \tag{7}$$

Regional Mean Error specifically focuses on the mouth region. Let $\mathcal{M}$ contain the coefficients associated with the jaw and lips. We compute the mean per-frame L2 error over this subset:

$$\text{Regional Mean Error} = \frac{1}{T}\sum_{t=1}^{T}\|\hat{\mathbf{b}}_{t,\mathcal{M}} - \mathbf{b}_{t,\mathcal{M}}\|_2. \tag{8}$$

A.2 Sequence-level Statistical Consistency

FFD. FFD measures the discrepancy in temporal variation amplitude by comparing the standard deviation of each predicted coefficient with that of the ground truth:

$$\mathrm{FFD} = \frac{1}{D}\sum_{d=1}^{D}|\hat{\sigma}_d - \sigma_d|. \tag{9}$$

Here, $\hat{\sigma}_d$ and σ_d are the temporal standard deviations of coefficient d, computed over the T evaluation frames using normalization $1/T$.

Fréchet Distance. Fréchet Distance compares Gaussian fits to the predicted and ground-truth coefficient distributions, treating each frame as a D-dimensional sample. Let $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ and $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote their empirical means and sample covariance matrices, respectively. The distance is

$$\mathrm{FD} = \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2 + \mathrm{tr}\left(\hat{\boldsymbol{\Sigma}} + \boldsymbol{\Sigma} - 2\left(\hat{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{1/2}\right)^{1/2}\right). \tag{10}$$

A.3 Image–Motion Correspondence

R-Precision (Top-1). R-Precision measures image-to-motion retrieval. Let $\mathbf{u}_t$ and $\mathbf{v}_t$ denote the L2-normalized image and predicted-motion embeddings from the fixed evaluation dual encoder. Each query t is compared against a candidate set $\mathcal{C}_t$ containing its paired motion and 31 other motions. For the query set $\mathcal{Q}$,

$$\mathrm{R@1} = \frac{1}{|\mathcal{Q}|}\sum_{t\in\mathcal{Q}}\mathbb{I}\left[t = \arg\max_{j\in\mathcal{C}_t}\mathbf{u}_t^\top\mathbf{v}_j\right], \tag{11}$$

where $\mathbb{I}[\cdot]$ is the indicator function. We use one random partition into groups of 32 with seed 42, discard the incomplete final group, and resolve ties by candidate order.

Multimodal Distance (MMD). MMD measures the mean Euclidean distance between paired, normalized image and predicted-motion embeddings over all aligned frames:

$$\mathrm{MMD} = \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{u}_t - \mathbf{v}_t\|_2. \tag{12}$$

B MOTION TOKENIZER TRAINING OBJECTIVES

This section specifies the reconstruction and alignment objectives used in Sec. 3.1. The ARKit and FLAME trajectories are paired at corresponding motion times. For each representation $r \in \{A, F\}$, the causal encoder produces $z_t^r = E_r(x_{1:t}^r)$. The transform $\mathcal{B}$ maps this output to bounded continuous coordinates $u_t^r = \mathcal{B}(z_t^r)$ in the same coordinate system as the fixed Cartesian FSQ grid $\mathcal{C} = \{e_v\}_{v=0}^{K-1}$; quantization yields $q_t^r = Q(z_t^r) = e_{c_t^r}$.

B.1 SELF- AND CROSS-RECONSTRUCTION

For source representation r and target representation s, the causal decoder predicts $\hat{x}_t^{r\to s} = D_s(q_{1:t}^r)$. Each of the four reconstruction paths minimizes

$$\begin{aligned} \ell_{r\to s} &= \langle \rho(\hat{x}^{r\to s} - x^s) \rangle_{\text{valid}} + \lambda_v \langle \rho(\Delta\hat{x}^{r\to s} - \Delta x^s) \rangle_{\text{adj}}, \\ \Delta x_t &= x_t - x_{t-1}. \end{aligned} \tag{13}$$

Here, ρ is the Smooth-L1 loss. The first average is over valid frames and target channels, excluding padding; the second is over target channels and adjacent frame pairs for which both frames are valid. The coefficient λ_v balances coefficient and temporal-difference supervision. The aggregated losses in Eq. 3 are

$$\mathcal{L}_{\text{self}} = \ell_{A\to A} + \ell_{F\to F}, \qquad \mathcal{L}_{\text{cross}} = \ell_{A\to F} + \ell_{F\to A}. \tag{14}$$

The weight λ_c controls cross-reconstruction relative to self-reconstruction, while λ_{CMCM} and λ_{MMS} weight the two alignment objectives below.

B.2 WINDOW-LEVEL CODE MATCHING

For a batch of B paired motion windows, let $\mathcal{T}_b$ denote the valid time positions in window b. Following CMCM (Liu et al., 2022), we construct soft assignments to the fixed FSQ grid and average them over each window. We then compare the resulting distributions using negative symmetric cross-entropy:

$$\begin{aligned} p_{b,t}^r(v) &= \text{softmax}_v\left(-\frac{\|u_{b,t}^r - e_v\|_2^2}{\tau_d}\right), \\ P_b^r(v) &= \frac{1}{|\mathcal{T}_b|}\sum_{t\in\mathcal{T}_b} p_{b,t}^r(v), \\ C_{bj} &= \sum_{v=0}^{K-1}\left[P_b^A(v)\log P_j^F(v) + P_j^F(v)\log P_b^A(v)\right], \\ \mathcal{L}_{\text{CMCM}} &= -\frac{1}{B}\sum_{b=1}^{B}\log\frac{\exp(C_{bb}/\tau_c)}{\sum_{j=1}^{B}\exp(C_{bj}/\tau_c)}. \end{aligned} \tag{15}$$

The temperatures τ_d and τ_c control soft code assignment and contrastive matching, respectively. Paired windows form positives, while FLAME windows from other clips provide negatives. Averaging is invariant to permutations of the already encoded frame representations, so this objective constrains aggregated code usage without directly specifying correspondence at each time step.

B.3 FRAME-LEVEL ALIGNMENT

We adapt the MMS objective of Liu et al. (2022) to paired time steps in the bounded continuous FSQ coordinates. At window position t, representations of the same clip form a positive pair, and representations from other clips at that position form negatives. Let $\mathcal{V}_t$ contain the clips with valid paired representations and sufficient encoder context at position t, and let $\mathcal{I} = \{(b,t) : b \in \mathcal{V}_t,\ |\mathcal{V}_t| \geq 2\}$ contain the anchors for which contrastive matching is defined. We use negative squared Euclidean distance, subtract a positive margin μ from the matched logit after temperature scaling, and average

both matching directions:

$$G_{bj,t} = -\frac{\|u^A_{b,t} - u^F_{j,t}\|_2^2}{\tau_f} - \mu\,\mathbf{1}[b = j],$$
$$\mathcal{L}_{\mathrm{MMS}} = -\frac{1}{2|\mathcal{I}|}\sum_{(b,t)\in\mathcal{I}}\left[\log\frac{\exp(G_{bb,t})}{\sum_{j\in\mathcal{V}_t}\exp(G_{bj,t})} + \log\frac{\exp(G_{bb,t})}{\sum_{j\in\mathcal{V}_t}\exp(G_{jb,t})}\right]. \tag{16}$$

Here, τ_f is the frame-matching temperature. The two terms match ARKit to FLAME and FLAME to ARKit, respectively. The positive-only margin encourages a larger similarity advantage for the matched pair over competing clips.