TAVR4D: TEMPORAL AUDIO-VISUAL TRANSFORMERS FOR RELIABLE 4D HEAD RECONSTRUCTION

Anonymous authors
Paper under double-blind review

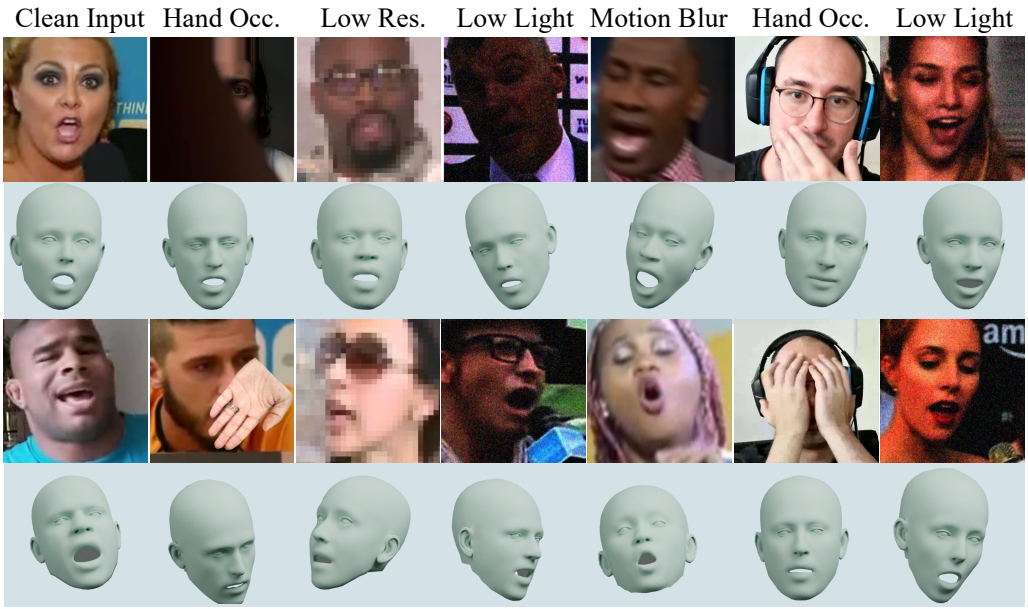


Figure 1: We present TAVR4D, a temporal multi-modal transformer method that can regress 4D head motion and facial expressions under different degradations. Here, we demonstrate the robustness of our model to five different degradations while keeping the clean prediction as plausible as possible.

ABSTRACT

Monocular 3D head reconstruction is accurate when visual evidence is reliable, but real videos often contain occlusions, illumination changes, motion blur, and other image degradations. Under these conditions, frame-wise estimators can interpret corrupted observations as motion, leading to large errors and jitter in pose and expression prediction. Temporal smoothing can reduce jitter but does not recover motion when the current observation is missing or unreliable. We introduce **TAVR4D**, a long-context multimodal model for robust 4D head reconstruction from degraded video. We formulate reconstruction as a temporal inference problem in which different motion components benefit from different complementary sources of information. The model uses long-range visual context to estimate head rotation, neck pose and translation when individual frames are unreliable. It also uses synchronized speech to help estimate facial expression and jaw motion. A learned visual reliability gate controls the balance between visual and audio features. We evaluate TAVR4D under 14 visual degradation conditions, including real-hand and full-face occlusions as well as dropped frames, which are held out as explicit image-space corruption classes during training. Across conditions shared by all baselines, TAVR4D remains close to its clean-input reconstruction accuracy, while remaining competitive on clean input. Our results show that combining long-range temporal inference with reliability-gated audio-visual fusion improves motion recovery and reconstruction robustness under the evaluated visual degradations.

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

Monocular 3D head reconstruction is used in virtual telepresence, remote communication, AR/VR, and performance capture. Current methods Giebenhain et al. (2025); Zollhöfer et al. (2018); Zielonka et al. (2026) achieve high accuracy when the face is clearly visible, but this assumption does not hold throughout many real applications and general videos. A hand may cover the mouth, frames may freeze, illumination may change abruptly, or compression and motion blur may remove facial detail. The underlying head and facial motion remain continuous, while a frame-wise estimator only observes the corrupted frame. As a result, temporary visual failures can be interpreted as changes in 3D pose or expression, producing unstable head motion and articulation.

Many existing reconstruction methods Thies et al. (2016); Grassal et al. (2022); Qian (2024); Giebenhain et al. (2025) operate frame-wise and therefore cannot exploit motion information from surrounding observations when the current frame becomes unreliable. Optimization-based pipelines may additionally depend on visually observable cues from auxiliary components such as surface-normal estimators or landmark detectors, which can themselves become unreliable under severe degradation. Likewise, frame-wise regressors such as DECA Feng et al. (2021), SMIRK Retsinas et al. (2024), and SHeaP Schoneveld et al. (2025) have no temporal information available when visual evidence in the current frame is insufficient.

We introduce **TAVR4D**, a long-context audio-visual model for robust 4D head reconstruction under unreliable visual observations. Instead of reconstructing each frame independently, TAVR4D uses information over time. Global head motion typically evolves coherently over time, such that reliable observations before and after a degraded frame provide strong constraints on missing orientation and position. Facial expression and jaw articulation, in contrast, can change rapidly and are only weakly determined by distant frames; during speech, however, they are strongly correlated with the synchronized acoustic signal. Thus, when the face is visually unreliable, temporal visual context is particularly useful for recovering global motion, while speech provides complementary evidence for recovering expression and jaw articulation.

For facial articulation, TAVR4D predicts a *visual reliability gate* that controls the balance between visual evidence and synchronized speech. During training, we supervise this gate using the feature distance between clean and synthetically corrupted versions of the same frame under a frozen visual encoder. This provides a shared reliability target across degradation types without requiring corruption-specific labels. At test time, the model only uses the visual and audio streams that are available. It does not need a clean reference frame or a hand-designed corruption detector. When the visual observation is reliable, the model mainly uses the visual features. When visual reliability decreases, audio can contribute more to expression and jaw prediction. The same reliability estimate is also used to estimate the *identity frame*, from which the static FLAME Li et al. (2017) identity parameters are estimated with MICA Zielonka et al. (2022) and kept fixed throughout the sequence. Global head motion is estimated separately and does not use this gate.

We evaluate TAVR4D under 14 controlled visual degradation conditions, including occlusion and dropped frames, low resolution, compression, illumination changes, camera shake, and motion blur. Several corruption types are held out as explicit image-space classes during training and are used only for evaluation. Across these conditions, TAVR4D stays close to its clean-input performance (Section 4.3). We additionally evaluate sequences with large head and facial motion to verify that the model recovers motion instead of only producing smooth or constant predictions (Section 4.4). Long-range temporal context helps preserve global head motion when local visual observations become unreliable, while synchronized audio helps recover speech-related facial motion under occlusion. We further test mismatched audio and analyze the contributions of temporal context, audio, and reliability-aware fusion in Section 4.5. To summarize:

- We introduce **TAVR4D**, a long-context audio-visual model that uses temporal visual context for global head motion and synchronized speech for facial expression and jaw motion.
- We introduce a visual reliability gate for adaptive audio-visual fusion and reliability-based identity enrollment.
- We conduct a controlled evaluation of reconstruction accuracy, motion recovery, robustness to visual degradation, and sensitivity to mismatched audio.

2 Related Work

Monocular 3D Face Reconstruction. 3D face reconstruction from images has two main directions: (1) optimization-based fitting of a morphable model, and (2) feed-forward regression of its parameters. We refer to recent surveys (Zollhöfer et al., 2018; Zielonka et al., 2026) for more details, including model-free methods that regress geometry directly. Optimization-based methods fit a 3D morphable model (Blanz & Vetter, 2023; Li et al., 2017; Egger et al., 2020) to cues such as landmarks (Wood et al., 2022) or photometric appearance (Thies et al., 2016; Garrido et al., 2016; Qian, 2024). Pixel3DMM (Giebenhain et al., 2025) leverages dense surface normals and UV correspondences predicted in screen-space for the fitting. The optimization-based methods heavily rely on the initialization and are sensitive to corrupted data. They require optimization for every image, but often use the previous frame as initialization for faster convergence.

Regression-based methods use similar cues as optimization-based methods during train time, however, during inference time, they predict the model parameters in a (single) forward pass. MICA (Zielonka et al., 2022) reconstructs a metrical 3D face from a single image leveraging training on a 3D training corpus. DECA (Feng et al., 2021) learns FLAME identity and expression parameters from in-the-wild images, and EMOCA (Daněček et al., 2022) adds an emotion-consistency objective to recover expressive detail. SMIRK (Retsinas et al., 2024) replaces the differentiable renderer with neural image synthesis to sharpen expression estimates. SHeaP (Schoneveld et al., 2025) supervises head geometry through 2D Gaussian splatting. These methods are fast, however, estimate each frame independently. They therefore interpret an occluded, or photometrically degraded observation as a change in pose or expression, which produces incorrect predictions and jitter over a video. In contrast, we keep the same FLAME output space but condition every frame on the entire sequence, so that degraded frames are explained by temporal context.

3D Face Reconstruction under Occlusions Compared to the clean-image setting, reconstruction under occlusion and other degradations is less explored. Some occlusion-aware fitting methods (Hsieh et al., 2015; Egger et al., 2018) split the face into visible and occluded regions and use only the visible part in the data terms (RGB and/or depth). ROGUE (Tiwari et al., 2021) enforces consistency at the image and feature level between the coefficients predicted from occluded and clean versions of the same image. OFER (Selvaraju et al., 2025) uses a diffusion model to generate several shape and expression hypotheses from one image and then ranks them, which recovers plausible geometry behind an occluder. All of these methods work within a single image. The remaining visible pixels must carry the whole reconstruction. OFER's hypotheses also differ from frame to frame, so identity and expression drift over a video. Unlike these methods, we treat a degraded frame as a missing observation that the sequence fills in, and we predict how much to trust the current frame rather than choosing among single-image hypotheses.

Audio-Visual 4D Face Reconstruction Liu et al. Liu et al. (2015) introduced a method for real-time tracking of facial expressions based on Kinect data including audio, RGB and depth observations. The audio signal is used to retrieve motion from a dataset which is then combined with the visual tracking using a heuristic that takes the fitting error and silence into consideration. In Chen et al. (2018) additional phoneme alignment terms are introduced to support the visual tracking. In Hussen Abdelaziz et al. (2020), a neural network that jointly uses audio and video features is learned to predict blendshape coefficients of a 3DMM. During training time, modality dropout is applied to ensure that the model uses the information from both, audio and video. AVFace Chatziagapi & Samaras (2023) combines audio and visual features and uses synthetic occlusions during training. At test time, it processes videos using overlapping 16-frame windows. Conceptually AVFace is the closest to our work, however no code is provided making a fair comparison impossible. In contrast to AVFace, TAVR4D conditions each prediction on the full 100-frame sequence and explicitly separates long-range visual inference for global head motion from reliability-gated audio-visual inference for facial articulation. This allows reliable observations outside a prolonged corrupted interval to constrain the reconstruction; we isolate this effect through controlled context ablations in Appendix 7.9.2.

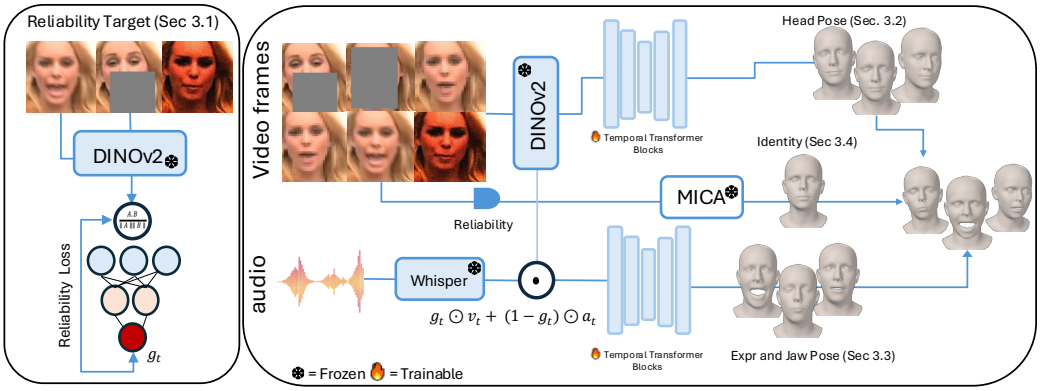


Figure 2: Overview of TAVR4D. A visual reliability gate predicts a scalar gate value from DINOv2 visual features. The gate controls the fusion of visual and audio features for facial articulation. Separate temporal branches estimate global head motion and facial articulation, while static identity is estimated with MICA Zielonka et al. (2022) and kept fixed over the sequence.

3 METHOD

As shown in Figure 2, each frame is encoded with frozen DINOv2 ViT-B/14 Oquab et al. (2023) and pooled into a frame-level representation. Separate temporal branches estimate global head motion and facial articulation, with reliability-gated audio-visual fusion used for the latter. Static identity is estimated with MICA Zielonka et al. (2022) and kept fixed over the sequence. Both temporal branches operate bidirectionally over the full 100-frame sequence, allowing each prediction to use observations before and after the target frame.

Spatial Attention Pooling. Given DINOv2 patch features $P_t \in \mathbb{R}^{N \times d}$ Oquab et al. (2023), we obtain the frame-level visual representation

$$v_t = \mathrm{MHA}\left(q, \mathrm{LN}(WP_t), \mathrm{LN}(WP_t)\right), \tag{1}$$

where q is a learned query and W projects the patch features to the temporal embedding dimension. This pooling step allows different image regions to receive different weights when forming the frame-level representation.

3.1 VISUAL RELIABILITY ESTIMATION

For each frame, TAVR4D predicts a scalar visual reliability gate $g_t \in (0, 1)$ from frozen DINOv2 Oquab et al. (2023) features. A two-layer MLP followed by five-frame temporal smoothing and a sigmoid predicts g_t.

During training, let $\tilde{x}_t$ and x_t denote the corrupted and clean versions of the same frame. Using the frozen DINOv2 Oquab et al. (2023) encoder ϕ, we define the soft reliability target

$$r_t = \exp\left(-\frac{1 - \cos\left(\phi(\tilde{x}_t), \phi(x_t)\right)}{\tau}\right), \tag{2}$$

where we use $\tau = 0.15$. A high value of r_t means that the corrupted frame remains close to the clean frame in DINOv2 feature space. The clean frame is used only to construct this target during training.

We supervise g_t against r_t using soft-label binary cross-entropy,

$$\mathcal{L}_{\text{gate}} = -\frac{1}{\sum_t m_t} \sum_t m_t \left[r_t \log g_t + (1 - r_t) \log(1 - g_t)\right], \tag{3}$$

where $m_t \in \{0, 1\}$ denotes the valid-frame mask.

The visual reliability gate is trained separately from the reconstruction model. During this stage, all other model parameters are frozen and only the gate is optimized. At inference time, the gate uses only the observed visual input; no clean reference frame or corruption-class label is required.

3.2 VISUAL BRANCH FOR HEAD POSE ESTIMATION

The pose branch predicts head rotation, translation, and neck pose from the visual sequence without using audio or reliability-gated fusion. The frame-level visual features are processed by a temporal transformer, allowing the prediction at each frame to use information from other observations in the sequence. This is particularly useful when the current visual frame is degraded or missing useful information. During training, we randomly drop visual features at selected frames to encourage the model to infer motion from temporal context instead of relying only on the current observation. Head and neck rotations are represented using continuous 6D rotations and converted to rotation matrices for supervision. We measure rotation error using the geodesic distance

$$d_{\mathrm{geo}}(R, \hat{R}) = \arccos\left(\frac{\mathrm{tr}\left(R^{\top}\hat{R}\right) - 1}{2}\right). \tag{4}$$

Translation is supervised with

$$\mathcal{L}_{\mathrm{trans}} = \frac{1}{\sum_t m_t} \sum_t m_t \left\|\hat{t}_t - t_t\right\|_2^2. \tag{5}$$

The pose objective is

$$\mathcal{L}_{\mathrm{pose}} = \mathcal{L}_{\mathrm{rot}} + 0.5\mathcal{L}_{\mathrm{neck}} + \mathcal{L}_{\mathrm{trans}}, \tag{6}$$

with

$$\mathcal{L}_{\mathrm{rot}} = d_{\mathrm{geo}}\left(R^{\mathrm{head}}, \hat{R}^{\mathrm{head}}\right), \qquad \mathcal{L}_{\mathrm{neck}} = d_{\mathrm{geo}}\left(R^{\mathrm{neck}}, \hat{R}^{\mathrm{neck}}\right). \tag{7}$$

3.3 AUDIO-VISUAL BRANCH FOR FACIAL EXPRESSION ESTIMATION

The multimodal branch predicts the 100-dimensional FLAME Li et al. (2017) expression vector and the 3-dimensional axis-angle jaw rotation. Audio is encoded using frozen Whisper-large-v3 Radford et al. (2022) features. The audio features are aligned with the video frames and projected to the same 256-dimensional embedding space as the visual representation.

We combine the visual and audio features using the predicted visual reliability gate:

$$f_t = g_t v_t + (1 - g_t)\mathrm{Proj}_a(a_t). \tag{8}$$

A high gate value gives more weight to the visual features, while a low gate value increases the contribution of synchronized audio. The fused sequence is then processed by a three-layer temporal transformer with embedding dimension 256 and eight attention heads.

We denote the dynamic facial parameters as

$$\psi_t = \left[\psi_t^{\mathrm{exp}}; \psi_t^{\mathrm{jaw}}\right] \in \mathbb{R}^{103}. \tag{9}$$

We predict these parameters as a clip-level mean together with a frame-dependent residual. This separates the average facial state of the sequence from its temporal variation. Given unconstrained residuals $\hat{\delta}_t$, we center them over valid frames:

$$\delta_t = \hat{\delta}_t - \frac{\sum_j m_j \hat{\delta}_j}{\sum_j m_j}, \tag{10}$$

and reconstruct

$$\hat{\psi}_t = \mu + \delta_t, \qquad \sum_t m_t \delta_t = 0. \tag{11}$$

The audio-visual branch is trained with the following objective, containing three groups of losses:

$$\mathcal{L}_{\text{expr-jaw}} = \mathcal{L}_{\mathrm{pos}} + \mathcal{L}_{\mathrm{motion}} + \mathcal{L}_{\mathrm{art}}, \tag{12}$$

where

$$\mathcal{L}_{\text{pos}} = \lambda_{\text{exp}}\mathcal{L}_{\text{exp}} + \lambda_{\text{jaw}}\mathcal{L}_{\text{jaw}} + \lambda_{\text{mean}}\mathcal{L}_{\text{mean}} + \lambda_{\text{dev}}\mathcal{L}_{\text{dev}}, \tag{13}$$

$$\mathcal{L}_{\text{motion}} = \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{acc}}\mathcal{L}_{\text{acc}} + \lambda_{\text{var}}\mathcal{L}_{\text{var}}, \tag{14}$$

$$\mathcal{L}_{\text{art}} = \lambda_{\text{clo}}\mathcal{L}_{\text{clo}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}} + \lambda_{\text{rng}}\mathcal{L}_{\text{rng}}. \tag{15}$$

$\mathcal{L}_{\text{pos}}$ supervises the predicted expression and jaw parameters together with their clip-level statistics. $\mathcal{L}_{\text{motion}}$ supervises temporal changes in the prediction using velocity, acceleration, and motion variation. $\mathcal{L}_{\text{art}}$ adds articulation-specific constraints related to mouth closure, geometry, and motion range. Exact definitions and weights are provided in Appendix 7.1.

The pose and expression branches are trained with their respective objectives. The visual reliability gate is trained separately while all other model parameters are frozen. Additional architecture, hyperparameter, and training details are provided in Appendices 7.2 and 7.3.

3.4 Reliability-Based Identity Estimation

Static identity is estimated with MICA Zielonka et al. (2022) and kept fixed throughout the sequence. Instead of always using the first frame, we select the valid frame with the highest gate value:

$$t^* = \arg\max_{t \in \mathcal{V}} g_t, \tag{16}$$

where $\mathcal{V}$ denotes the set of valid frames. MICA Zielonka et al. (2022) is applied to the selected identity enrollment frame, and the resulting static FLAME Li et al. (2017) identity parameters are used for all frames in the sequence.

4 Experimental Results

4.1 Evaluation Protocol and Implementation

Evaluation data. Our primary evaluation uses 54 100-frame Multiface clips Wuu et al. (2022). Additional dataset details are provided in Appendix 7.4. All methods receive identical inputs under 14 controlled degradation settings, detailed in Appendix 7.5.

Training and supervision. We train the audio-visual model for 30,000 steps on Pixel3DMM-fitted VoxCeleb2 Chung et al. (2018), followed by 6,000 steps on FaMoS Bolkart et al. (2023). DINOv2 ViT-B/14 Oquab et al. (2023) and Whisper-large-v3 Radford et al. (2022) remain frozen. Full details are in Appendix 7.3.

4.2 Clean Reconstruction Accuracy

Table 1: Clean reconstruction on 54 Multiface clips. Errors are in mm except rotation (°).

Method	Vertex ↓	Rotation ↓	Nose ↓	Mouth ↓	Lower face ↓
Ours	**1.140**	0.70	**0.976**	2.034	**1.487**
Pixel3DMM	1.177	**0.68**	1.697	**1.979**	1.905
DECA	1.293	0.85	1.466	2.110	2.324
SMIRK	1.318	0.72	1.460	2.179	2.111
SHeaP	1.382	0.86	1.458	2.018	2.205
EMOCA	1.390	0.86	1.597	2.610	2.830

Table 2: Shape error under representative visual degradations. Vertex error is measured in mm on affected frames. A dash denotes no usable fit.

Condition	**Ours**	Pixel3DMM	SHeaP	SMIRK	DECA	EMOCA
Lower face occ. .9	**1.132**	1.850	1.548	1.534	1.493	1.626
Real hand	**1.133**	1.677	1.490	1.463	1.384	1.501
Full face .9	**1.150**	–	–	1.417	1.518	2.028
Low light	**1.167**	–	–	1.392	1.461	1.601

We first verify that the robustness of TAVR4D does not come at the expense of reconstruction accuracy on clean videos, see Table 1. TAVR4D achieves the lowest whole-face vertex error of 1.140 mm and the second-lowest head rotation error of $0.70°$. It remains competitive across the other metrics, although it does not outperform the strongest baselines on mouth reconstruction.

One difference between the methods is identity estimation. TAVR4D estimates one identity for the full sequence using MICA Zielonka et al. (2022), while the compared methods normally estimate identity independently for each frame. For a fairer identity comparison, we fix each baseline's identity prediction from its corresponding clean first frame, while TAVR4D uses one MICA Zielonka et al. (2022) identity for the sequence. Thus, part of our clean-input gain may arise from sequence-level identity enrollment; our primary focus is robustness while retaining competitive clean accuracy.

4.3 Reconstruction Accuracy Under Degradations

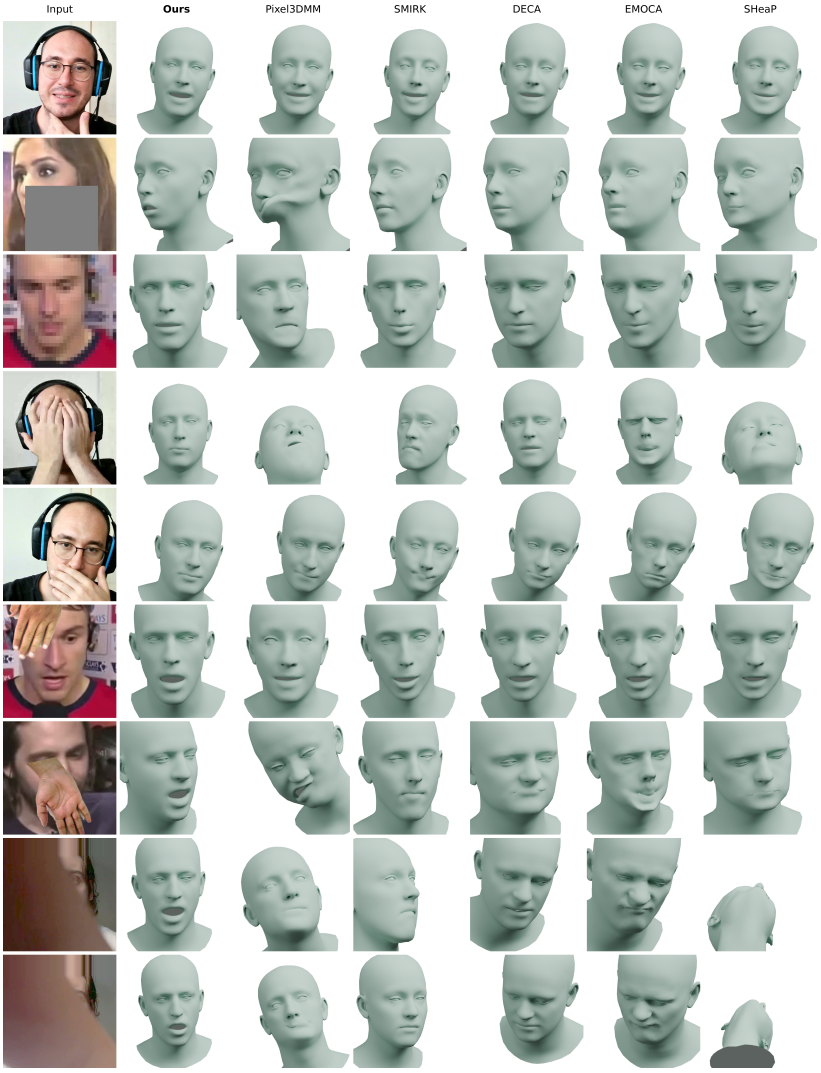


Figure 3: One clip per column (clean, occlusion, occlusion, lower face occluded, low light, motion blur, low exposure, quality drop, low light) and one method per row. Real-hand occlusions are composited from segmented hand images from 11k Hands Afifi (2019) dataset; construction details are provided in Appendix 7.5.

We next evaluate whether temporary loss of visual evidence changes the reconstructed 3D shape. Across the shared corruption conditions, TAVR4D remains close to its reconstruction on the corresponding clean input. The representative conditions in Table 2 show the same trend across occlusion, and photometric degradation. Full results across all degradation conditions are provided in Appendix 7.7. Qualitative results are shown in Figure 3.

Under severe full-face occlusion and low light, Pixel3DMM Giebenhain et al. (2025) and SHeaP Schoneveld et al. (2025) do not provide a consistent full-set evaluation because of fit failures, whereas TAVR4D continues to return valid reconstructions. A model could obtain a stable geometric error by producing an approximately correct but nearly static face. Therefore, we evaluate the motion recovery separately.

4.4 Motion Recovery

Head motion. Because the Multiface Wuu et al. (2022) clips contain limited rigid head motion, we additionally evaluate on 14 held-out FaMoS Bolkart et al. (2023) clips with large head turns, see

Table 3: Motion recovery under large motion and visual degradation. Head rotation is evaluated on 14 held-out FaMoS clips and reported as geodesic rotation error (°); each entry gives the mean over all frames / the high-motion interval. Mouth motion reports Pearson r/RMSE (mm). A dash indicates that a consistent full-set evaluation is unavailable because of fit failures.

(a) Head rotation

Method	Clean	Lower face occ. 0.6	Lower face occ. 0.9
DECA	**5.0 / 5.4**	11.3 / 22.3	13.5 / 27.8
SMIRK	7.5 / 8.0	21.1 / 42.6	21.5 / 43.8
Ours	10.2 / 13.5	**10.7 / 15.2**	**11.3 / 16.7**

Values are mean error over all frames / mean error during the high-motion interval.

(b) Mouth motion: r/RMSE

Condition	**Ours**	P3DMM	SMIRK	SHeaP	DECA	EMOCA
Clean	0.84/1.75	**0.92**/1.53	0.84/1.75	0.84/1.83	0.82/1.92	0.78/1.94
Lower face .9	**0.56/2.98**	–	0.04/3.82	0.03/4.50	0.02/4.09	0.04/4.19
Real hand	**0.74/2.44**	–	0.36/3.48	0.44/3.48	0.24/3.84	0.36/3.52
Full face .9	**0.50/3.21**	–	0.04/3.58	0.02/3.58	0.04/3.59	−0.00/3.61
Low light	**0.63/2.73**	–	0.01/3.60	−0.06/8.41	0.03/3.76	0.18/3.54

Table 3. The clips reach an average peak rotation of $59.7°$, with an average rotation of $46.2°$ during the high-motion interval.

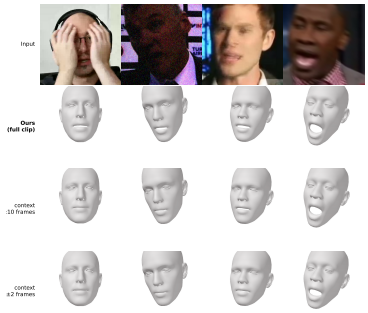
On clean input, DECA Feng et al. (2021) and SMIRK Retsinas et al. (2024) are more accurate, with high-motion errors of $5.4°$ and $8.0°$, compared with $13.5°$ for TAVR4D. This exposes a trade-off on clean sequences with unusually large head motion. Under severe lower face occlusion, however, TAVR4D degrades substantially less: at severity 0.9, its error is $16.7°$, compared with $27.8°$ for DECA and $43.8°$ for SMIRK. Thus, the advantage in this setting is robustness to unreliable observations rather than superior clean large-motion accuracy.

Mouth motion. We measure non-rigid facial motion directly against the Multiface Wuu et al. (2022) scans by computing the vertical extent of the lip region on the aligned prediction and ground-truth scan. We report Pearson correlation and RMSE; full metric definitions are provided in Appendix 7.6. On clean input, the methods perform similarly. Under strong visual degradations, the correlations of frame-wise methods drop sharply, while TAVR4D better preserves the ground-truth motion trajectory. Results across all degradation conditions are provided in Appendix 7.8.

4.5 Ablations



(a) Reliability gate and audio ablation.



(b) Temporal context ablation.

Figure 4: Ablation analysis of TAVR4D. (a): reliability-gated audio-visual fusion under clean, mismatched-audio, and occluded inputs. (b): restricting visual context reduces temporal motion stability.

Table 4: Gate and audio ablations under visual occlusion. r is correlation with scan mouth motion and a is predicted-to-scan motion amplitude.

	Lower face occ. .9		Real hand	
Variant	$r\uparrow$	a	$r\uparrow$	a
Full	**0.56**	0.76	**0.74**	1.00
w/o gate	0.54	0.76	0.72	1.03
w/o audio	0.21	0.40	0.67	0.87

Table 5: Video × audio stress ablation. Wrong audio is another subject's other sentence; occluded denotes lower-face occlusion at severity .9.

		Full		w/o gate	
Video	Audio	$r\uparrow$	a	$r\uparrow$	a
Clean	Clean	**0.84**	1.05	0.86	1.00
Clean	Wrong	**0.82**	0.95	0.61	0.85
Occluded	Clean	**0.56**	0.76	0.54	0.76
Occluded	Wrong	**0.13**	0.79	0.11	0.77

Table 6: Temporal-context and corruption-training ablations. Excess acceleration measures inter-frame jitter. Lower is better.

Temporal context		Corruption training	
Context	Excess acc. ↓	Training	Excess acc. ↓
Self	1.224	With	**1.050**
±2	1.180	Without	1.213
±10	1.167		
Full clip	**1.050**		

Ablations. Reliability gate and audio. We conducted the ablations under same degraded settings for 54 Multiface Wuu et al. (2022) clips. Under severe lower face occlusion, removing audio reduces mouth-motion correlation from 0.56 to 0.21, whereas removing the reliability gate has a smaller effect (0.54), shown in Table 4. Thus, synchronized audio provides the main complementary cue when articulation is visually hidden. The gate is most beneficial when audio is mismatched: with clean video and wrong audio, correlation decreases from 0.82 to 0.61 without reliability-aware fusion (Table 5). Qualitative results of the ablation study are shown in Figure 4 (a).

Temporal stability. Table 6 shows that both long temporal context and corruption training reduce excess acceleration, which we use as a measure of inter-frame jitter. The full-context model is the most stable.

Long-range visual context. To directly test whether distant visual observations contribute to head-motion recovery, we apply centered full-face occlusions of 10–60 frames to the 54 Multiface clips and evaluate the same central 10 frames in every condition. Using the same trained pose model, context is restricted only at inference time. Full-sequence rotation error remains nearly constant ($6.03^\circ \rightarrow 6.37^\circ$), whereas a physical ± 10-frame window degrades from 5.53° to 12.57° once the occlusion outlasts the local context, approaching frame-wise inference (12.75°). The paired degradation interaction is $+6.69^\circ$ (95% CI $[5.57, 7.85]$). As a mechanism check, perturbing visible frames outside the occluded interval changes the full-context prediction but leaves the ± 10 and frame-wise predictions unchanged, confirming that the model uses distant visual evidence. Full results and qualitative examples are in Appendix 7.9.2.

5 Limitations and Future Work

TAVR4D still inherits limitations from its training supervision. Large-scale video training relies partly on pseudo-labels from Pixel3DMM Giebenhain et al. (2025), which have limited dynamic mouth articulation. This limitation can constrain the recovered expressions. Although the finetuning stage reduces this effect, clean mouth reconstruction remains weaker than the strongest image-based methods. More diverse paired audio-visual data with ground-truth 4D geometry could further improve expression fidelity and generalization.

Although it is trained with audio dropout, TAVR4D still assumes that the available speech signal is approximately synchronized with the video and informative about the observed articulation. Moreover, the current reliability gate estimates only visual reliability and does not explicitly detect unreliable or mismatched audio. Extending reliability estimation to both modalities is therefore a natural direction for future work. Also, the current temporal model uses information from the full input sequence and therefore operates offline. A causal or windowed variant would be required for real-time applications.

6 Conclusion

We introduced TAVR4D, a temporal audio-visual model for robust 4D reconstruction under unreliable visual observations. Rather than treating every frame independently, TAVR4D combines long-range visual context and synchronized speech according to the information required by different motion components. Global head motion is recovered primarily from temporal visual context, while speech provides complementary information for expression and jaw articulation. A learned visual reliability score controls the contribution of the current visual observation and is additionally used for reliability-based identity enrollment.

Across a broad range of visual degradations, TAVR4D remains close to its clean-input reconstruction accuracy while preserving substantially more head and facial motion than frame-wise reconstruction methods under severe corruption. Our ablations show that this robustness arises from the combination of long-range temporal context, audio conditioning, corruption-aware training, and reliability-based fusion rather than from temporal smoothing alone. These results suggest that treating 4D reconstruction as reliability-conditioned temporal inference provides a practical way to maintain coherent head and facial motion when individual video frames become unreliable.

AI USE STATEMENT

In this work, we used generative AI tools for editing the scripts and for technical assistance such as code writing. We have not used generative AI tools for any other tasks requiring disclosure, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for editing the paper to improve readability. We have reviewed all AI-assisted work. AI-assisted code was manually reviewed and tested for correctness by the authors, and all AI-edited text was reviewed to ensure accuracy and consistency with the authors' intended meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the details required to reproduce our method and experiments in the main paper and Appendix. The Appendix describes the training procedure, architecture and hyperparameters, dataset splits, visual corruption protocol, and evaluation protocol in detail. We also specify the datasets, pretrained models, training stages, and evaluation metrics used throughout our experiments.

REFERENCES

Mahmoud Afifi. 11k hands: gender recognition and biometric identification using a large dataset of hand images. *Multimedia Tools and Applications*, 2019. doi: 10.1007/s11042-019-7424-8. URL https://doi.org/10.1007/s11042-019-7424-8.

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 157–164. 2023.

Timo Bolkart, Tianye Li, and Michael J. Black. Instant multi-view head capture through learnable registration. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 768–779, 2023.

Aggelina Chatziagapi and Dimitris Samaras. Avface: Towards detailed audio-visual 4d face reconstruction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16878–16889. IEEE, 2023.

Xin Chen, Chen Cao, Zehao Xue, and Wei Chu. Joint audio-video driven facial animation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3046–3050, 2018. doi: 10.1109/ICASSP.2018.8461502.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

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.

Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision*, 126(12):1269–1287, 2018.

Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. 39(5):1–38, 2020.

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

Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. Reconstruction of personalized 3d face rigs from monocular video. *ACM Transactions on Graphics (TOG)*, 35(3):1–15, 2016.

Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Pixel3dmm: Versatile screen-space priors for single-image 3d face reconstruction. *arXiv preprint arXiv:2505.00615*, 2025.

Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. pp. 18653–18664, 2022.

Pei-Lun Hsieh, Chongyang Ma, Jihun Yu, and Hao Li. Unconstrained realtime facial performance capture. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

Ahmed Hussen Abdelaziz, Barry-John Theobald, Paul Dixon, Reinhard Knothe, Nicholas Apostoloff, and Sachin Kajareker. Modality dropout for improved performance-driven talking faces. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, ICMI '20, pp. 378–386, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375818. doi: 10.1145/3382507.3418840. URL https://doi.org/10.1145/3382507.3418840.

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

Yilong Liu, Feng Xu, Jinxiang Chai, Xin Tong, Lijuan Wang, and Qiang Huo. Video-audio driven real-time facial animation. *ACM Trans. Graph.*, 34(6), November 2015. ISSN 0730-0301. doi: 10.1145/2816795.2818122. URL https://doi.org/10.1145/2816795.2818122.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. URL https://api.semanticscholar.org/CorpusID:53592270.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Shenhan Qian. Vhap: Versatile head alignment with adaptive appearance priors, sep 2024. URL https://github.com/ShenhanQian/VHAP.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

George Retsinas, Panagiotis P. Filntisis, Radek Danecek, Victoria F. Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3d facial expressions through analysis-by-neural-synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Liam Schoneveld, Zhe Chen, Davide Davoli, Jiapeng Tang, Saimon Terazawa, Ko Nishino, and Matthias Nießner. Sheap: Self-supervised head geometry predictor learned via 2d gaussians. 2025.

Pratheba Selvaraju, Victoria Fernández Abrevaya, Timo Bolkart, Rick Akkerman, Tianyu Ding, Faezeh Amjadi, and Ilya Zharkov. Ofer: Occluded face expression reconstruction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26985–26995. IEEE, 2025.

J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. 2016.

Hitika Tiwari, Min-Hung Chen, Yi-Min Tsai, Hsien-Kai Kuo, Hung-Jen Chen, Kevin Jou, KS Venkatesh, and Yong-Sheng Chen. Self-supervised robustifying guidance for monocular 3d face reconstruction. *arXiv preprint arXiv:2112.14382*, 2021.

Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, et al. 3d face reconstruction with dense landmarks. In *European Conference on Computer Vision*, pp. 160–177. Springer, 2022.

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. In *arXiv*, 2022. doi: 10.48550/ARXIV.2207.11243. URL https://arxiv.org/abs/2207.11243.

Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pp. 250–269. Springer, 2022.

Wojciech Zielonka, Tobias Kirschstein, Timo Bolkart, Simon Giebenhain, Vanessa Sklyarova, Xiang Deng, Donglai Xiang, Shunsuke Saito, Yebin Liu, Matthias Nießner, and Justus Thies. How to build digital humans? from priors to photorealistic avatars. *Computer Graphics Forum (Eurographics State-of-the-Art Report)*, 45(2), 2026.

Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. 37(2):523–550, 2018.

7 APPENDIX

7.1 LOSS FUNCTIONS

We predict the 100-dimensional FLAME Li et al. (2017) expression parameters and the 3-dimensional jaw rotation for each frame. The expression targets are lightly smoothed over time, while the jaw targets are kept unchanged.

The training objective consists of three groups of losses:

$$\mathcal{L}_{\text{expr-jaw}} = \mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{motion}} + \mathcal{L}_{\text{art}}. \tag{17}$$

The position loss supervises the expression and jaw parameters, together with their clip-level mean and frame-wise deviations:

$$\mathcal{L}_{\text{pos}} = \lambda_{\text{exp}}\mathcal{L}_{\text{exp}} + \lambda_{\text{jaw}}\mathcal{L}_{\text{jaw}} + \lambda_{\text{mean}}\mathcal{L}_{\text{mean}} + \lambda_{\text{dev}}\mathcal{L}_{\text{dev}}. \tag{18}$$

The motion loss supervises temporal changes using velocity, acceleration, and motion variation:

$$\mathcal{L}_{\text{motion}} = \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{acc}}\mathcal{L}_{\text{acc}} + \lambda_{\text{var}}\mathcal{L}_{\text{var}}. \tag{19}$$

The articulation loss constrains jaw rotation and mouth motion:

$$\mathcal{L}_{\text{art}} = \lambda_{\text{clo}}\mathcal{L}_{\text{clo}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}} + \lambda_{\text{rng}}\mathcal{L}_{\text{rng}}. \tag{20}$$

We use $\lambda_{\text{exp}} = 1$, $\lambda_{\text{jaw}} = 1$, $\lambda_{\text{mean}} = 2$, $\lambda_{\text{dev}} = 1$, $\lambda_{\text{vel}} = 1$, $\lambda_{\text{acc}} = 1$, $\lambda_{\text{var}} = 0.5$, $\lambda_{\text{clo}} = 1.5$, $\lambda_{\text{geo}} = 2$, and $\lambda_{\text{rng}} = 1$.

7.2 ARCHITECTURE AND HYPERPARAMETERS

Our model uses separate temporal branches for facial articulation and head motion. Both branches process sequences of $T = 100$ frames using bidirectional transformers with learned absolute positional embeddings.

Reliability gate. The reliability features are processed by a two-layer MLP followed by a depthwise temporal convolution with kernel size 5. The resulting gate is used for audio-visual feature fusion. The expression branch uses dropout of 0.1 and stochastic depth across its transformer blocks. The pose branch does not use audio or reliability-based fusion.

7.3 TRAINING DETAILS

We train the expression branch on VoxCeleb2 Chung et al. (2018) and FaMoS Bolkart et al. (2023) clips to increase the range of facial articulation. The reliability gate is trained afterwards while all other model parameters are frozen.

All stages use AdamW Loshchilov & Hutter (2017) with weight decay 0.05, a batch size of 8 clips of 100 frames, and an input resolution of 224×224. We use gradient clipping at 1.0 and a fixed learning rate without scheduling. The learning rate is 10^{-4} for the initial VoxCeleb2 Chung et al. (2018) training and 3×10^{-5} when continuing training with FaMoS Bolkart et al. (2023). Visual corruption is applied with probability 0.8 and affects at most 75% of the temporal window while preserving clean context before and after the corrupted interval. We use visual and audio modality dropout probabilities of 0.3 and 0.15, respectively.

The reliability gate is trained separately with binary cross-entropy using a learning rate of 3×10^{-4}. Training stops when the validation objective no longer improves.

Pose branch. The pose branch is trained separately with AdamW and a batch size of 2. During robustness training, we use synthetic occlusions and a learning rate of 5×10^{-5} for 6,000 optimization steps.

7.4 DATASETS

The expression model is trained on a mixture of VoxCeleb2 Chung et al. (2018) and FaMoS Bolkart et al. (2023). Our training set contains 1,360 clips in total, consisting of 780 VoxCeleb2 clips and 580 FaMoS clips from 850 subjects. VoxCeleb2 Chung et al. (2018) training samples contain 100 consecutive frames, while FaMoS Bolkart et al. (2023) clips contain 8–59 frames.

For evaluation, we use 54 Multiface Wuu et al. (2022) clips of 100 frames each. We additionally use 14 FaMoS Bolkart et al. (2023) clips with large head motion for the head-motion experiments.

7.5 VISUAL CORRUPTION PROTOCOL

Each corruption is applied to a single contiguous interval of a 100-frame clip, affecting at most 75 frames and leaving clean context before and after the corrupted interval. Evaluation metrics are computed only on the affected frames, and the same corrupted frames are used for all methods.

We evaluate 14 degradation conditions covering 10 degradation types. Lower face occlusion and full-face occlusion are evaluated at severities 0.3, 0.6, and 0.9, while all other degradations are evaluated at severity 0.9. The remaining conditions include real-hand occlusion, dropped frames, low resolution, JPEG compression, low light, exposure changes, camera shake, and motion blur. At severity 0.9, low resolution corresponds to a $7\times$ downsampling, compression to JPEG quality 8, low light to gamma 2.8 with Gaussian noise, and motion blur to a 13-pixel linear blur kernel. Real-hand occlusion, full-face occlusion, and dropped frames are not used during training. In Figure 5, all degradation types used in training and testing are shown.

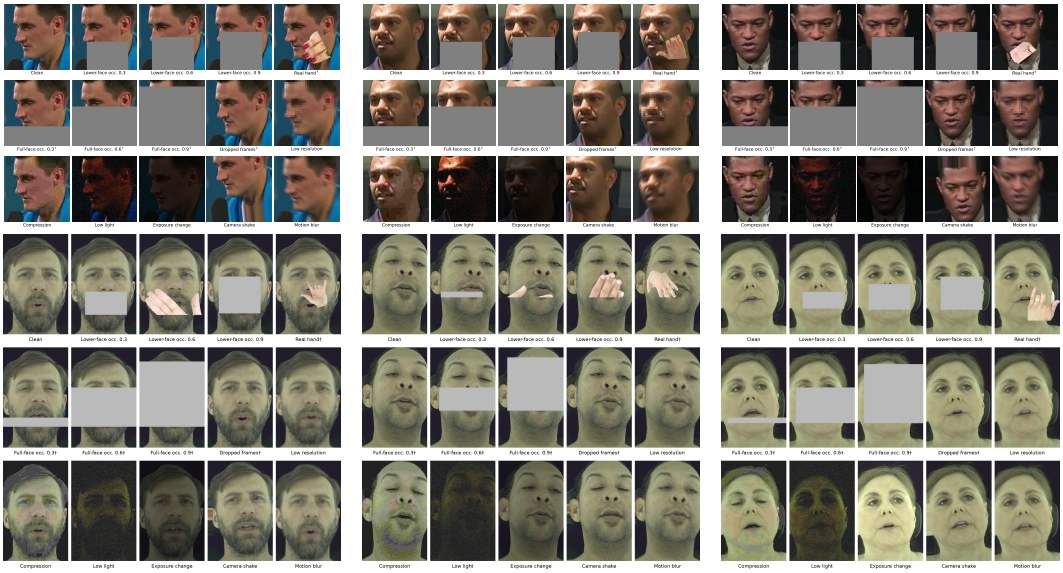


Figure 5: Visual corruption types used for evaluation and training, shown on VoxCeleb2 Chung et al. (2018) and Multiface Wuu et al. (2022)examples. Occluders move coherently over time rather than being sampled independently for each frame. † indicates corruption types that are not used during training.

For real-hand occlusion, we sample hand images from the 11k Hands dataset Afifi (2019), segment them using an off-the-shelf hand segmentation model, and composite the resulting masks onto the video. The hand motion is propagated coherently over time rather than resampled independently for each frame.

7.6 EVALUATION PROTOCOL

For evaluation, we align each predicted mesh to the corresponding ground-truth scan independently for every frame using a similarity transformation estimated from the upper-face region. We exclude

the mouth and jaw area from the alignment so that expression- and articulation-related deformations are not absorbed by the registration. Alignment is performed with 25 iterations of trimmed ICP, discarding the 20% largest residual correspondences at each iteration to reduce the influence of outliers. All methods are evaluated using the same alignment procedure and the same corrupted-frame indices. Vertex errors are then computed as distances from ground-truth scan vertices to the aligned predicted surface and aggregated over frames and clips.

For mouth-motion evaluation, we measure the vertical extent of the lip region on the aligned prediction and ground-truth scan. We report Pearson correlation and mean-centered RMSE over the affected frames.

For head rotation, we recover the relative rotation using the same Procrustes-based procedure for all methods and use a fixed per-clip offset estimated from the corresponding clean prediction. Rotation error is reported as geodesic distance in degrees.

7.7 Full Robustness Results

Table 7: Shape robustness over all degradation conditions on 54 clips. Vertex error is in mm on the affected frames; lower is better. A superscript gives the number of clips excluded from that cell as fit failures, and a dash marks a condition in which no clip yields a usable fit.

Condition	**Ours**	Pixel3DMM	SHeaP	SMIRK	DECA	EMOCA
Lower face occ. .3	**1.128**	1.259	1.438	1.402	1.349	1.540
Lower face occ. .6	**1.130**	1.345	1.447	1.440	1.474	1.645
Lower face occ. .9	**1.132**	1.850^{16}	1.548^{1}	1.534	1.493	1.626
Real hand	**1.133**	1.677^{3}	1.490	1.463	1.384	1.501
Full face .3	**1.135**	1.217	1.404	1.363	1.346	1.429
Full face .6	**1.145**	2.089^{26}	1.650^{7}	1.529	1.444	1.569
Full face .9	**1.150**	–	–	1.417	1.518	2.028
Dropped	**1.153**	1.239	1.385	1.329	1.297	1.413
Low resolution	**1.172**	1.311	1.368	1.304	1.289	1.397
Compression	**1.150**	1.268^{10}	1.364^{8}	1.361	1.315	1.371
Low light	**1.167**	–	–	1.392	1.461	1.601
Exposure	**1.167**	1.576^{1}	1.363	1.308	1.288	1.393
Camera shake	**1.141**	1.315	1.440	1.332	1.330	1.425
Motion blur	**1.140**	1.196	1.382	1.289	1.278	1.394
Fit failures	**0**	164	124	0	0	0

Table 7 reports the complete shape-reconstruction results over all 14 degradation conditions. TAVR4D remains close to its clean-input reconstruction error across degradation types and severities, with errors ranging from 1.128 to 1.172 mm. The frame-wise baselines generally degrade more strongly, particularly under severe occlusion and low-light conditions. TAVR4D produces a valid reconstruction for all evaluated clips, whereas Pixel3DMM Giebenhain et al. (2025) and SHeaP Schoneveld et al. (2025) exhibit fit failures under several severe degradations. Since low shape error alone does not imply that the underlying facial motion is preserved, we evaluate motion recovery separately in Appendix 7.8.

7.8 Mouth Motion under Degradation

A low per-frame shape error does not necessarily imply that the reconstructed sequence preserves the underlying facial motion. We therefore evaluate mouth motion directly over all degradation conditions. As described in Appendix 7.6, we compare the predicted and ground-truth mouth-opening trajectories using Pearson correlation and mean-centered RMSE. Correlation measures how well the temporal variation is preserved, while RMSE measures the discrepancy between the motion trajectories in millimeters.

Table 8: Mouth-motion tracking against the ground-truth scan on 54 clips. Each entry reports Pearson correlation r and the RMSE of the mouth-opening trajectory in millimeters.

Condition	**Ours**	SMIRK	SHeaP	DECA	EMOCA
Clean	**0.84/1.75**	**0.84/1.75**	**0.84**/1.83	0.82/1.92	0.78/1.94
Lower face occ. .3	**0.66/2.52**	0.26/3.25	0.30/3.24	0.18/3.31	0.30/3.06
Lower face occ. .6	**0.61/2.81**	0.13/3.58	0.18/3.70	0.01/3.79	0.09/3.74
Lower face occ. .9	**0.56/2.98**	0.04/3.82	0.03/4.50	0.02/4.09	0.04/4.19
Real hand	**0.74/2.44**	0.36/3.48	0.44/3.48	0.24/3.84	0.36/3.52
Full face .3	**0.77**/1.97	0.70/2.03	0.73/1.97	0.64/2.16	0.76/**1.91**
Full face .6	**0.51/2.97**	−0.00/3.53	−0.02/4.59	0.01/3.97	0.02/3.65
Full face .9	**0.50/3.21**	0.04/3.58	0.02/3.58	0.04/3.59	−0.00/3.61
Dropped	**0.64**/2.78	0.61/2.73	0.62/2.83	0.56/2.83	0.63/**2.71**
Low resolution	**0.84**/1.95	0.75/1.53	0.75/1.70	0.72/1.65	0.76/**1.43**
Compression	**0.79**/2.14	0.76/1.92	0.64/2.99	0.76/2.00	0.74/**1.62**
Low light	**0.63/2.73**	0.01/3.60	−0.06/8.41	0.03/3.76	0.18/3.54
Exposure	**0.83**/1.95	0.80/1.62	0.74/1.75	0.80/1.78	0.76/**1.51**
Motion blur	0.80/1.90	0.84/1.47	0.84/1.64	0.82/1.53	**0.85/1.39**

As shown in Table 8, the largest differences appear when visual evidence around the mouth is strongly degraded. Under severe lower face occlusion, real-hand occlusion, full-face occlusion, and low light, TAVR4D maintains substantially higher correlation with the ground-truth motion than the frame-wise baselines. Under degradations that retain more usable facial information, including low resolution, compression, exposure changes, and motion blur, the differences are smaller and some frame-wise methods obtain lower RMSE. These results indicate that the main advantage of the temporal audio-visual model emerges when the current visual observation provides insufficient evidence for recovering facial articulation.

7.9 Additional Ablations

7.9.1 Reliability Gate and Audio

To disentangle the contribution of synchronized speech from the reliability gate, we compare the full model with variants retrained without the gate and without the audio stream. All variants are evaluated on the same Multiface Wuu et al. (2022) clips, using identical corrupted inputs and affected-frame indices.

Table 9: Ablation of reliability-aware fusion and synchronized audio. We report Pearson correlation r and motion amplitude ratio a.

	Full		w/o gate		w/o audio	
Condition	$r\uparrow$	a	$r\uparrow$	a	$r\uparrow$	a
Clean	**0.84**	1.05	0.76	1.00	**0.86**	1.03
Lower face occ. .3	**0.66**	0.93	0.63	0.98	0.46	0.74
Lower face occ. .6	**0.61**	0.80	0.59	0.85	0.32	0.51
Lower face occ. .9	**0.56**	0.76	0.54	0.76	0.21	0.40
Real hand	**0.74**	1.00	0.72	1.03	0.67	0.87

Table 10: Response of the visual-reliability gate. Lower values indicate reduced visual reliability.

Condition	Clean	Corrupted
Full face .9	0.94	0.09
Low light	0.94	0.13
Full face .6	0.94	0.21
Lower face occ. .9	0.94	0.34
Lower face occ. .6	0.94	0.50
Lower face occ. .3	0.94	0.65
Real hand	0.94	0.68
Compression	0.94	0.72

As shown in Table 9, removing the audio stream increasingly reduces mouth-motion recovery as occlusion severity increases. At lower face occlusion severity 0.9, the full model reaches a correlation of 0.56, compared with 0.21 without audio. In contrast, the model without the reliability gate remains close to the full model under matched clean audio, reaching 0.54 at the same severity. This indicates that synchronized audio provides the main complementary signal for recovering articulation when the mouth is visually occluded.

The reliability gate instead controls when the audio stream should influence the reconstruction. On clean input, removing the gate reduces correlation from 0.84 to 0.76, while under real-hand occlusion the difference is smaller (0.74 vs. 0.72). Together with the mismatched-audio experiment in the main

paper, these results indicate that the gate primarily protects the reconstruction from unnecessary or unreliable audio conditioning, whereas audio itself provides the missing articulation signal under visual occlusion.

Table 10 further shows that the learned gate responds systematically to degradations that reduce the reliability of the current visual observation. Severe full-face occlusion and low light produce the largest decrease in gate value, while moderate lower face occlusion and real-hand occlusion induce smaller reductions.

Across the 14 degradation conditions, the decrease in visual reliability is strongly associated with the improvement over the best frame-wise baseline (Spearman $\rho = 0.81$, $p < 0.001$). This relationship is consistent with the intended behavior of the fusion mechanism: the advantage of temporal audio-visual conditioning becomes larger as the current visual evidence becomes less reliable.

7.9.2 Visual Temporal Context and Fusion Design

AVFace Chatziagapi & Samaras (2023) is the closest prior audio-visual 4D face reconstruction method. However, we could not identify publicly available code or pretrained models, which prevents a controlled direct comparison under our evaluation protocol. AVFace processes test sequences using overlapping 16-frame windows, whereas TAVR4D conditions each prediction on the full 100-frame sequence.

Rather than treating our context ablation as a surrogate evaluation of AVFace, we use it to isolate the effect of this architectural difference under a common model and evaluation protocol. We additionally ablate reliability-aware audio-visual fusion separately.

Long-range visual context for head motion. Our model is designed such that long-range visual context primarily supports global head motion, whereas synchronized speech provides complementary information for expression and jaw articulation. We therefore test the visual-context design directly on head pose.

We construct full-face occlusions of 10, 20, 40, and 60 consecutive frames, centered at the same temporal location in each of the 54 Multiface Wuu et al. (2022) clips. For all occlusion lengths, we evaluate the same central 10 frames. Shared frames are pixel-identical across conditions, so that only the temporal extent of the surrounding occlusion changes.

We compare full-sequence context against two inference-time restrictions: a physical ± 10-frame crop centered on the evaluated frame and a frame-wise (*Self*) variant. All variants use the same pose network, trained on 100-frame sequences; only the available context is changed at inference time.

Table 11: Head-rotation error ($^\circ$, lower is better) as the duration of a contiguous full-face occlusion increases. All values are evaluated on the same central 10 frames of 54 Multiface Wuu et al. (2022) clips. The final column reports the change from 10 to 60 occluded frames; brackets denote 95% clip-level bootstrap confidence intervals. Context restriction is applied only at inference time.

Context	10	20	40	60	Δ_{60-10}
Full clip	6.03	5.96	6.07	6.37	$+0.35$ [-0.03, $+0.77$]
± 10 frames	5.53	5.68	12.57	12.57	$+7.04$ [$+6.06$, $+8.01$]
Self	12.75	12.75	12.75	12.75	0.00

Table 11 summarizes the quantitative effect of restricting temporal context as occlusion duration increases. The paired degradation interaction between the ± 10 and full-context conditions is $+6.69^\circ$ (95% bootstrap CI $[+5.57, +7.85]$), confirming that the benefit of full context increases as the occlusion extends beyond the local window.

For short occlusions, the local ± 10 context is sufficient and performs comparably to the full context. Once the occlusion outlasts the local window, however, the local model no longer has access to reliable pose observations: its error increases to 12.57° at 40–60 frames, close to the frame-wise variant (12.75°). In contrast, the full-context model remains stable, reaching 6.37° at 60 frames.

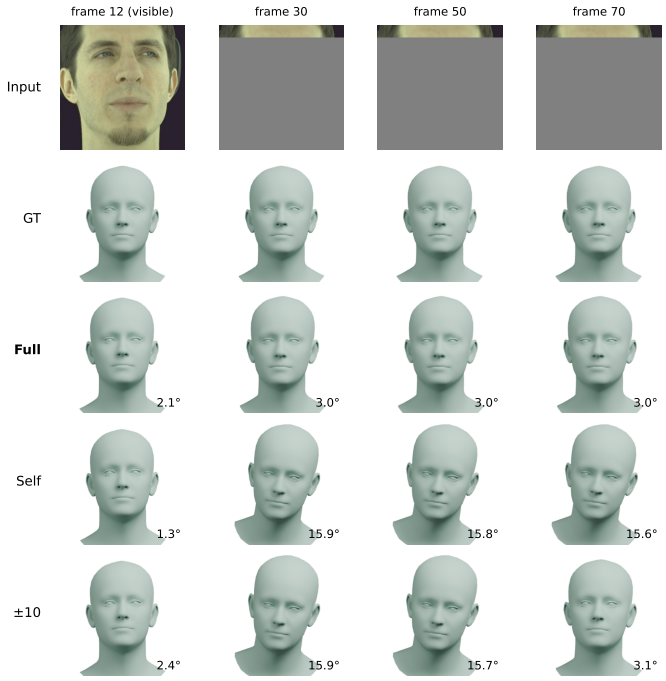


Figure 6: Qualitative effect of visual context during a 60-frame full-face occlusion. Per-frame geodesic head-rotation errors are shown below each prediction. When the target frame is visible (frame 12), all context variants recover the pose accurately. Deep inside the occlusion (frames 30 and 50), the ± 10-frame and frame-wise variants have no access to reliable visual observations and drift to an incorrect pose, whereas the full-context model uses visible observations outside the corrupted interval and remains close to the ground truth.

Figure 6 shows the same mechanism qualitatively. The local-context model is accurate while reliable frames are available within its temporal window, fails in the interior of the long occlusion, and recovers near the boundary once visible frames become accessible again. In contrast, full-clip context maintains the correct head orientation throughout the corrupted interval.