Real-Time Facial Animation of Gaussian Head Avatars via Mocap-to-Parametric Expression Mapping

I-Hsin Chen
National Taiwan University
Taipei City, Taiwan
r12944005@csie.ntu.edu.tw

Sheng-Yen Huang
National Taiwan University
Taipei City, Taiwan
d12944001@csie.ntu.edu.tw

Yi-Ping Hung
National Taiwan University
Taipei City, Taiwan
hung@csie.ntu.edu.tw

Abstract

We present a real-time system for animating photorealistic 3D Gaussian head avatars driven by motion data from consumer facial mocap systems. At the core of our framework is an efficient regression model that maps mocap-derived blendshape parameters to the expression space of parametric face models, enabling direct control over Gaussian avatars without iterative optimization. Our design incorporates a lightweight expression regularization mechanism that improves stability and expressiveness by encouraging semantically disentangled, identity-specific deformations. Through extensive evaluations—implemented using ARKit as the mocap source and FLAME as the target parametric model—we show that our method outperforms both fitting-based and regression-based baselines in animation quality and latency. The system supports both live streaming and offline reenactment, enabling efficient, real-time avatar control for virtual meetings and social telepresence.

1. Introduction

Recent advances in 3D Gaussian Splatting [11] have led to highly realistic neural head avatars with efficient real-time rendering. To enable animation control, many recent systems leverage parametric face models as a structural prior, encoding facial motion into interpretable low-dimensional parameters, providing a controllable abstraction that supports dynamic avatar synthesis and facilitates cross-identity reenactment.

In this setting, expressions from a source actor can be estimated by fitting the parametric model to monocular input, and the resulting facial parameters can be applied to a target identity to animate a distinct avatar. However, these fitting-based methods often incur significant computational overhead and temporal latency, limiting their suitability for real-time applications such as live avatar streaming or interactive communication. Moreover, they typically ignore the expression distribution of the target avatar, making direct parameter transfer prone to identity mismatch or deformation artifacts.

To address this limitation, we present a real-time system that controls Gaussian head avatars directly from facial motion capture (mocap). Our method learns a lightweight regression model that maps mocap-derived blendshape parameters to the expression space of a parametric model, enabling fast and stable animation without iterative fitting. In addition, we introduce an expression regularization strategy that constrains the learned mapping toward semantically disentangled and identity-consistent deformations.

In our implementation, we adopt Apple's ARKit [2] as the mocap source and FLAME [15] as the target model due to their wide availability and compatibility. The system supports both live streaming and offline playback, running at up to 60 FPS on standard consumer GPUs. We evaluate our method on the NeRSemble [12] dataset and benchmark against both fitting-based and regression-based alternatives. Results demonstrate that our expression-regularized model consistently outperforms existing baselines in terms of geometric accuracy, temporal stability, and perceptual quality.

In summary, this paper contributes a real-time avatar control framework that combines facial motion capture input with Gaussian rendering, featuring a novel expression mapping strategy enhanced by expression regularization, along with a comprehensive evaluation across realistic animation scenarios.

2. Related Work

2.1. 3D Morphable Models

3D Morphable Models [6] (3DMMs) are a class of parametric models that represent the geometry and appearance of human faces using a set of low-dimensional latent parameters. By decomposing facial geometry into distinct subspaces—typically shape, expression, and pose—3DMMs enable controllable facial reconstruction and animation.

Early 3DMMs, such as the Basel Face Model (BFM) [20], focused on modeling identity through PCA over 3D face scans. Later extensions, including FaceWarehouse [3], introduced expression modeling by incorporating RGB-D facial sequences across multiple subjects, thus expanding 3DMMs beyond static identity modeling. Among recent models, FLAME [15] has become a widely adopted standard, offering detailed control over identity $(\beta, \Delta v)$, expression (ϕ), jaw, neck and eye pose (θ). With PCA-based shape and expression spaces, its compatibility with mesh-based deformation makes it a powerful backbone for tasks involving dynamic face tracking and controllable avatar animation.

2.2. Gaussian Head Creation

Recent methods have integrated 3D Gaussian Splatting [11] with parametric face models to create photorealistic and controllable head avatars. These hybrid approaches leverage the structural prior of models like FLAME to guide Gaussian deformation in response to facial motion.

One line of research [14, 19] models expression changes as weighted Gaussian blendshapes. They interpolate Gaussian attributes—such as position, color, and opacity—based on FLAME expression coefficients, enabling smooth and expressive control. Another line of work [22, 24, 27] establishes a direct binding between Gaussians and mesh surfaces by associating each Gaussian with a specific triangle. This surface-aware strategy enhances geometric consistency and offers improved controllability during animation.

As shown in Figure 1, these systems typically perform cross-identity reenactment by extracting expression and pose parameters (ϕ_S, θ_S) from a source video using optimization-based fitting, and transferring them to a target identity $(\beta_T, \Delta v_T)$ for rendering. High-fidelity pipelines such as VHAP [21] and INSTA [30] optimize over entire sequences to achieve temporally consistent reenactment, but are computationally expensive and inherently offline.

Rather than optimizing over entire video sequences, several methods estimate facial parameters from individual frames using both FLAME [4, 8, 23] and BFM [5, 13] models. While these approaches are more efficient than multi-frame optimization, they still fall short in responsiveness and temporal stability, limiting their suitability for real-time avatar control.

2.3. Mocap to Parametric Expression Mapping

EMAGE [18] addresses the challenge of translating mocap-derived blendshape parameters into a parametric expression space by proposing a linear projection model that maps large-scale ARKit sequences from the BEAT [17] dataset into FLAME parameters, despite lacking ground-truth supervision. It constructs 52 reference meshes—one neutral and 51 with isolated blendshape activations—and fits

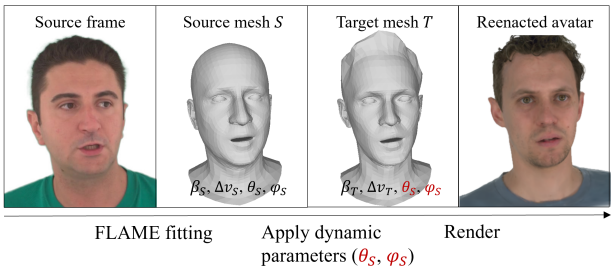


Figure 1. Traditional Cross-identity reenactment pipeline. Source dynamic parameters (ϕ_S, θ_S) are estimated via FLAME fitting and applied to a target identity $(\beta_T, \Delta v_T)$.

FLAME parameters by minimizing vertex distances to these handcrafted targets.

While EMAGE uses these meshes as direct regression targets, our approach leverages them in a different role to guide the projection process. We further adopt a supervision strategy tailored to each subject, as detailed in Section 3.

3. Method

3.1. System Overview

Our system enables real-time animation of Gaussian head avatars, driven by facial motion capture signals. It builds upon GaussianAvatars [22], a framework that binds 3D Gaussians to the FLAME mesh for photorealistic head reconstruction and expression control. On top of this rendering backbone, we introduce an efficient regression-based expression mapping module and a dual-mode rendering pipeline that supports both live streaming and offline playback. We adopt ARKit as a representative mocap source due to its widespread use and robust tracking capabilities on consumer devices. Beyond ARKit, the framework generalizes naturally to other mocap systems inspired by the Facial Action Coding System (FACS) [7] that output interpretable blendshape coefficients.

As illustrated in Figure 2, the pipeline comprises three main components: (1) Expression Mapper, (2) Gaussian Head Renderer, and (3) a graphical user interface (GUI).

In live mode, ARKit blendshape data is streamed from an iPhone via Open Sound Control (OSC) using the *Face Cap* [1] iOS application. These inputs are processed in real time by our pretrained subject-specific expression mapper, which projects them into the FLAME expression space. The resulting dynamic parameters are passed to the Gaussian head renderer, where they are combined with static identity parameters to update the avatar at 60 FPS under the desired camera pose. Head and eye poses are extracted directly from ARKit and remapped to the pose parameter space to ensure complete articulation.

In recording mode, the blendshape data stream is cap-

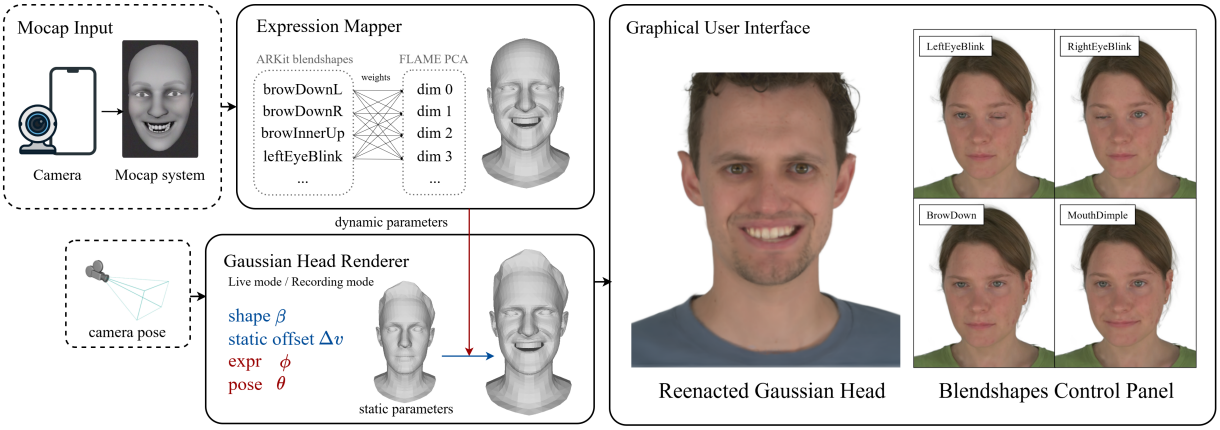


Figure 2. System overview. Solid boxes denote system modules, while dashed boxes indicate user inputs. Motion capture input is processed through a learned expression mapper to estimate FLAME dynamic parameters. These are passed to a Gaussian head renderer, where they are combined with static identity parameters to animate head avatars in either live or recording mode. The GUI supports both streaming and manual expression control.

tured and saved as timestamped sequences. These can be replayed and applied to arbitrary avatars, facilitating consistent evaluation across different identities. Figure 3 demonstrates cross-identity reenactment results using the same expression sequence mapped to multiple Gaussian avatars.

To support intuitive inspection and debugging, the system includes an interactive GUI adapted from the official GaussianAvatars [22] viewer. In addition to real-time streaming, the interface provides a dedicated control panel for manually activating blendshapes. Compared to the original PCA component sliders, this adjustment method offers clearer semantic interpretability, as each blendshape coefficient corresponds to a well-defined facial action unit.

This modular architecture allows for flexible operation across real-time and offline scenarios, while preserving high-fidelity animation.

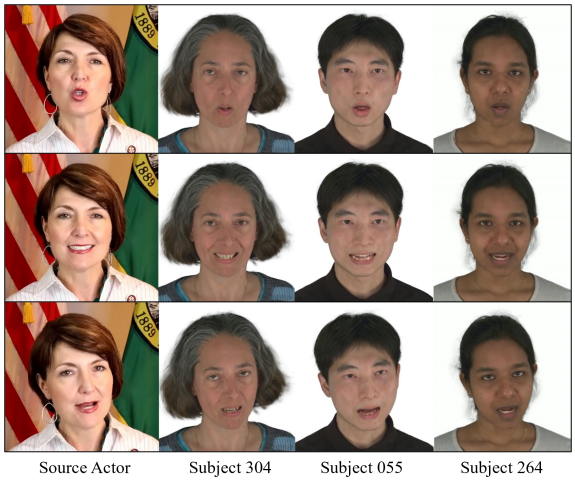


Figure 3. Recording mode allows the same blendshape sequence to be applied to different Gaussian avatars.

3.2. Expression Mapper

3.2.1 Problem Formulation

ARKit provides a 51-dimensional blendshape vector $x \in \mathbb{R}^{51}$ to represent facial expressions in a semantically meaningful format. While this representation is effective for real-time motion capture, it is defined in a proprietary space and not directly compatible with FLAME expression space.

FLAME represents expressions using a 100-dimensional PCA-based expression parameter $\phi \in \mathbb{R}^{100}$, and models jaw movement as part of the pose parameters, specifically as a 3-dimensional rotation vector $\theta_{\text{jaw}} \in \mathbb{R}^3$. Therefore, in order to animate a FLAME-based Gaussian avatar using ARKit inputs, we aim to learn a mapping function:

$$f : \mathbb{R}^{51} \rightarrow \mathbb{R}^{103}, \tag{1}$$

which projects the blendshape vector x to the FLAME expression and jaw pose vector $y = [\phi, \theta_{\text{jaw}}] \in \mathbb{R}^{103}$.

This task is formulated as a supervised regression problem, where the training data consists of temporally aligned pairs of blendshape vectors and corresponding FLAME expression and jaw pose parameters. The goal is to find a mapping function f that generalizes well across expressive behaviors and captures the nuanced relationships between the two representation domains.

3.2.2 Data Preparation

We conduct all experiments on the NeRSemble dataset [12], a multi-view facial performance corpus designed for neural

head reconstruction. Each sequence is captured simultaneously by 16 synchronized RGB cameras with calibrated parameters, enabling high-fidelity expression modeling.

To supervise our expression mapper, we collect paired sequences of ARKit blendshapes and FLAME expression parameters. FLAME parameters are extracted using a multi-view fitting pipeline based on VHAP [21], which jointly optimizes shape, expression, and pose across views and time using photometric and landmark-based losses. The implementation uses the 2023 FLAME model with manually added rigid teeth geometry for improved visual realism.

For ARKit blendshape capture, we use the same front-view camera video frames from the multi-view sequences employed in the previous FLAME-fitting step to ensure frame-level synchronization. These frames are played back as the capture stimulus and recorded using the Face Cap app [1] on an iPhone's TrueDepth camera. This procedure yields well-aligned training data, with each frame represented by a 51-dimensional blendshape vector and a set of FLAME expression and pose parameters.

3.2.3 Model Selection

We evaluate three types of regression models with increasing complexity: k-Nearest Neighbors (KNN), ridge regression, and a shallow multi-layer perceptron (MLP). These models are chosen for their simplicity and suitability for real-time applications.

KNN serves as an instance-based baseline but suffers from poor generalization due to its reliance on sample density. MLPs offer greater expressive power but are less stable and harder to interpret in low-data, subject-specific settings.

Ridge regression offers the best trade-off between accuracy, robustness, and real-time performance. Its linearity provides interpretability and stability, while ℓ_2 regularization prevents overfitting and promotes generalization. As a result, we adopt ridge regression as the core of our expression mapping module.

3.2.4 Expression Regularization

While ridge regression provides a good balance between accuracy and efficiency, it often entangles unrelated blendshapes and results in non-localized deformations (e.g., eye closure causing unintended jaw drop).

To address this, we introduce an *expression regularization* (ER) prior, implemented as a matrix $W_{\text{ER}} \in \mathbb{R}^{51\times 103}$ that encodes the ideal FLAME response when each ARKit blendshape is activated in isolation. This prior constrains the mapping function toward semantically disentangled and spatially localized deformations.

As shown in Figure 4, the ER prior is constructed using 52 canonical meshes from the EMAGE framework [18],

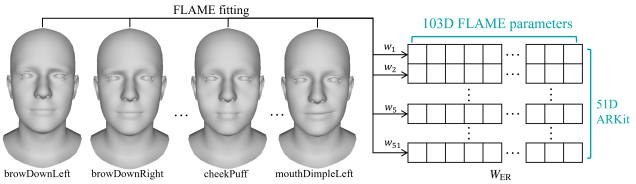


Figure 4. Illustration of the expression regularization (ER) prior. Each row represents the FLAME response to activating a single ARKit blendshape.

including one neutral face and 51 expressions with isolated blendshapes activated. We refit each mesh using the FLAME 2023 model to extract consistent expression ϕ and θ_{jaw} parameters, holding shape fixed. Stacking the resulting 51 vectors row-wise yields W_{ER}.

We incorporate W_{ER} as a soft prior in the training loss, encouraging the model to associate each blendshape with its corresponding facial region, thereby promoting semantic disentanglement while preserving learning flexibility.

3.2.5 Training Objective

The expression mapper learns a projection matrix $W \in \mathbb{R}^{51\times 103}$ that maps blendshape inputs $x_i \in \mathbb{R}^{51}$ to FLAME parameters $y_i \in \mathbb{R}^{103}$, comprising 100 expression coefficients and 3 jaw pose values. The training objective is a weighted combination of data loss, ridge regularization (RR), and expression regularization (ER):

$$L = \underbrace{\frac{1}{T}\sum_{i=1}^{T}\|x_i^\top W - y_i\|_2^2}_{L_{\text{MSE}}} + \underbrace{\alpha\,\|W\|_F^2}_{L_{\text{RR}}} + \underbrace{\lambda_{\text{ER}}\,\|W - W_{\text{ER}}\|_F^2}_{L_{\text{ER}}}. \tag{2}$$

T denotes the number of training samples and α, λ_{ER} are hyperparameters controlling the RR and ER weights.

L_{MSE} enforces accurate reconstruction, L_{RR} stabilizes the solution under limited data, and L_{ER} encourages disentanglement by penalizing deviation from the precomputed W_{ER}. This formulation balances accuracy and stability, while the ER prior further enhances the semantic separation of individual blendshape coefficients.

Training the Gaussian Avatar To enable photorealistic reenactment, we adopt the training pipeline from GaussianAvatars [22] to reconstruct a Gaussian head avatar. The same video frames and FLAME parameters used for training the expression mapper are reused to supervise the avatar reconstruction, ensuring consistency in expression representation. Figure 5 illustrates the joint pipeline comprising both the expression mapping module and avatar training.

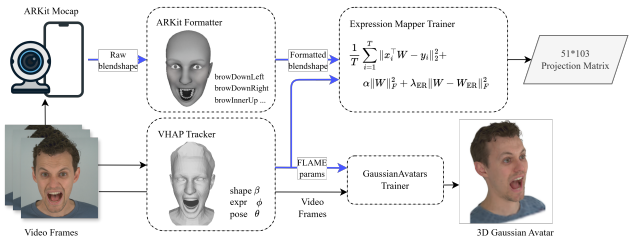


Figure 5. Overview of the joint training pipeline. The top branch optimizes a projection matrix $W \in \mathbb{R}^{51\times103}$ using temporally aligned ARKit and FLAME sequences. The bottom branch shows the training of the Gaussian Avatar model using the same FLAME sequences and video frames.

3.3. Performance Analysis

Our system is designed for real-time streaming applications, where low-latency response is critical. To assess its practical performance, we benchmark the full pipeline—from mocap input to Gaussian rendering—under live streaming conditions on an RTX 3090 GPU. The system includes:

1. Blendshape acquisition via iPhone (~60 FPS)
2. Expression mapping using ridge regression with expression regularization (matrix multiplication, ~40,000 FPS)
3. Gaussian avatar rendering (~60 FPS)

The end-to-end pipeline achieves **60 FPS**, with rendering being the most computationally demanding stage.

In contrast, while widely used for facial parameter estimation, existing fitting-based methods such as DECA [8] and SMIRK [23] achieve only 2 FPS and 20 FPS during expression estimation, respectively, when benchmarked under identical hardware conditions (excluding Gaussian rendering). This substantial gap underscores the efficiency of our lightweight mapping design, making it more suitable for interactive scenarios such as live telepresence, meetings, or virtual media.

4. Experiment

4.1. Experimental Setup

We conduct experiments on a customized subset of the NeRSemble dataset [12], comprising over 10,000 frames from nine subjects with varied demographics. Data acquisition and preprocessing follow the procedure described in Section 3.2.2. Gaussian Head Avatars are trained per subject using 1,000–1,500 frames.

4.2. Evaluation Metrics

We evaluate the generated animations using five metrics categorized into two types: vertex-based and image-based. These metrics jointly assess mesh-level precision, motion smoothness, and perceptual consistency.

Vertex-based Metrics.

- **Mean Vertex Error (MVE)** [25] is the average, over all frames, of the maximum per-vertex L2 distance between predicted and ground-truth meshes, reflecting geometric fidelity.
- **Facial Dynamics Deviation (FDD)** [25, 28] evaluates how well the predicted facial motion matches the temporal dynamics of the ground truth by comparing standard deviations of vertex displacement.
- **Motion Stability Index (MSI)** [16] quantifies temporal smoothness by computing the inverse variance of vertex acceleration.

Image-based Metrics.

- **Inter-frame Similarity Index (ISI)** [9, 10] computes SSIM [26] between adjacent rendered frames to measure structural consistency.
- **Inter-frame Transformation Fidelity (ITF)** [9, 10] measures PSNR between neighboring frames, capturing low-level frame-to-frame fidelity and flicker.

4.3. Quantitative Evaluation

We quantitatively evaluate our animation framework in terms of accuracy, temporal consistency, and generalization. Experiments cover three settings—self reenactment, cross-identity reenactment, and subject generalization—each targeting a distinct challenge in facial animation.

4.3.1 Self Reenactment

This experiment assesses the accuracy and stability of different regression models on held-out sequences from the same subject. We adopt a 90/10 train-test split for each of the nine subjects. Four regressors are compared: MLP, KNN, ridge regression (RR), and our proposed regression with additional expression regularization (RR+ER).

All models share fixed hyperparameters: RR and RR+ER use $\alpha = 0.1$, with the latter additionally applying $\lambda_{ER} = 1$; MLP has two hidden layers (32 and 64 units) with ReLU activations; KNN uses $k = 3$.

This setting focuses on subject-specific learning, excluding fitting-based baselines (e.g., DECA [8], SMIRK [23]) that cannot leverage training data from the same individual. The experiment thus serves as an ablation to isolate the effect of regression strategies under strong priors.

Table 1. Comparison of regression models on all metrics (averaged across 9 subjects). VHAP results are omitted from vertex-based metrics, as it serves as the reference for supervision.

Model	MVE ($\times10^3$) ↓	FDD ($\times10^4$) ↓	MSI ($\times10^{-4}$) ↑	ISI ↑	ITF ↑
RR	5.1969	0.6770	8.8959	0.9867	42.4660
RR+ER	6.1654	0.2762	8.9847	0.9879	43.1688
KNN	6.9952	2.5272	8.2742	0.9848	41.9741
MLP	6.9252	1.3253	8.7088	0.9853	41.7221
VHAP	-	-	-	0.9878	42.7841

Table 1 shows that RR achieves the best geometric accuracy (lowest MVE), while RR+ER slightly increases reconstruction error due to regularization, but yields significant gains in temporal metrics (FDD, MSI).

Notably, RR+ER achieves the highest scores in both image-based metrics (ISI, ITF), outperforming all baselines, even including VHAP, which consistently ranks second. Although VHAP serves as our pseudo ground truth, its tracking may retain subtle micro-movements inherent in real-world facial motion. In contrast, the regularized regression suppresses such motion-level jitter, resulting in more temporally stable outputs.

These results validate the expression regularization as an effective structural prior that promotes temporal stability and smoother motion. Detailed per-subject metrics are included in the supplementary material.

4.3.2 Cross-identity Reenactment

To evaluate generalization to unseen source actors and real-world videos, we test our system on monocular sequences from actors not present in the training set, using them to drive Gaussian Head Avatars trained on different identities. This one-shot reenactment setup enables fair comparison with optimization-based methods. We select three test sequences with diverse motion complexity:

- **A (monocular capture)**: Calm expressions (e.g., blinking, smiling), recorded via iPhone.
- **B (NeRSemble [12])**: High-strain expressive motions from an unseen dataset subject.
- **C (HDTF [29])**: Speech-driven facial motion with rapid articulation.

We compare our RR+ER model against DECA [8] and SMIRK [23], two fitting-based baselines. While DECA and SMIRK extract expression and pose from video via fitting, our method regresses ARKit blendshapes into the FLAME space. The resulting parameters drive a target avatar, with all methods equally unfamiliar with the source actor.

Table 2. Cross-identity reenactment results on metrics MSI, ISI, and ITF. MVE and FDD are excluded, as no ground-truth FLAME parameters exist for these test subjects.

Sequence	Metric	DECA	SMIRK	Ours
A	MSI ($\times 10^{-4}$) ↑	7.50	8.79	**9.74**
	ISI ↑	0.973	0.978	**0.987**
	ITF ↑	36.67	36.80	**44.48**
B	MSI ($\times 10^{-4}$) ↑	2.74	9.16	**9.56**
	ISI ↑	0.913	0.971	**0.986**
	ITF ↑	27.89	34.32	**41.85**
C	MSI ($\times 10^{-4}$) ↑	8.50	8.81	**9.51**
	ISI ↑	0.960	0.963	**0.971**
	ITF ↑	33.01	32.48	**35.97**

Table 2 shows that our method consistently achieves higher temporal stability and perceptual smoothness across all sequences. Improvements are especially pronounced in Sequence B, where high-intensity expressions challenge temporal coherence.

Qualitative analyses in Section 4.4.2 further illustrate how our subject-specific mapping preserves expression semantics from unseen source actors while maintaining plausible target-specific deformations, outperforming fitting-based baselines without access to actor-specific priors.

4.3.3 Subject Generalization

While the previous section 4.3.2 evaluates generalization to *unseen source actors*, here we assess the complementary case of *unseen target avatars*. Specifically, we train a single RR+ER model on six subjects to form a generalized model, and evaluate it on the sequences from three unseen subjects.

Table 3. Comparison of subject-specific and generalized models on unseen subjects (excluded from training). Results are averaged across 3 subjects.

Metric	Subject-specific	Generalized
MVE ($\times 10^{3}$) ↓	**6.0718**	9.1246
FDD ($\times 10^{4}$) ↓	0.3574	**0.2292**
MSI ($\times 10^{-4}$) ↑	**9.4858**	9.4437
ISI ↑	**0.9841**	0.9831
ITF ↑	**42.6810**	42.2762

Table 3 compares this generalized model against a subject-specific model trained on each unseen subject. Compared to the subject-specific model, the generalized model suffers from a noticeable drop in reconstruction accuracy (MVE), indicating the benefit of personalized priors. However, temporal smoothness (MSI, ISI, ITF) remains comparable, and even improves in FDD, suggesting that identity-agnostic regressors can still produce perceptually stable animations. These results highlight a trade-off between reconstruction fidelity and general applicability.

4.4. Qualitative Evaluation

While the quantitative analysis provided numerical insights into accuracy and stability, qualitative evaluation is necessary to further assess the perceptual realism, facial coherence, and failure modes of the generated animations.

4.4.1 Self Reenactment

We visualize representative reenactment results corresponding to the self-reenactment experiment in Section 4.3.1, focusing on two subjects—038 and 253—that best illustrate the behavioral differences between baseline and regularized regressors. We include only RR and RR+ER models for

clarity, as KNN and MLP exhibit consistently inferior visual quality and are omitted here.

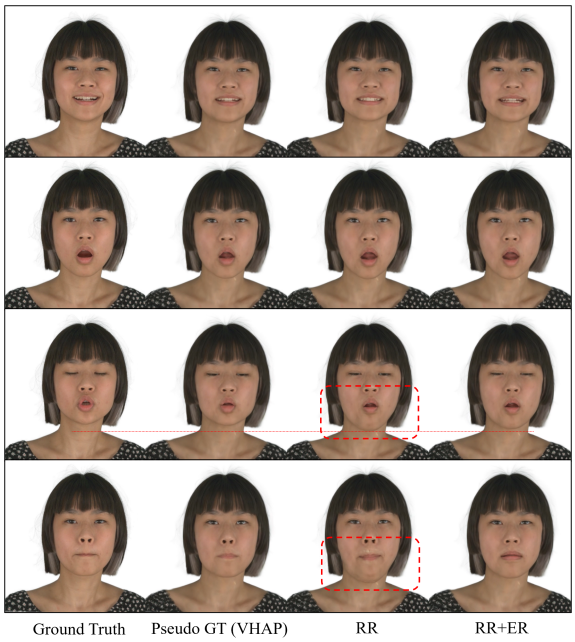


Figure 6. Self reenactment results for Subject 038. Top: both models closely match the pseudo-GT. Bottom: RR model exhibits lower-face instability during eye blinks and lip press (highlighted in the red box), while the RR+ER model maintains more localized and consistent deformations.

Figure 6 presents four rows of results for Subject 038. The top two rows demonstrate that both models faithfully reconstruct typical expressions, closely aligning with VHAP. The bottom rows highlight critical differences: (1) the RR model incorrectly couples eye-blink motion with jaw displacement, and (2) when facing rare expressions like lip press (unseen in training), it produces unstable deformations. These artifacts suggest that the RR baseline tends to entangle unrelated blendshapes and is highly dependent on the coverage of the training data. In contrast, the RR+ER model maintains structural locality, constraining motion to stable regions even under unfamiliar inputs. For unseen expressions like lip press, it shows slight deviations and minor entanglement, though far less than the RR baseline.

In Figure 7, RR+ER exhibits a slightly more conservative bias in reproducing large jaw movements than RR. While it enhances motion stability, this may come at the cost of reduced expression intensity in high-strain scenarios.

Together, the observations highlight a core trade-off: RR+ER offers better semantic control and robustness to out-of-distribution inputs, whereas RR may better preserve dynamic range but is prone to entangled or unstable outputs.

For completeness, supplementary material includes video versions of the results shown in Figures 6 and 7, along with qualitative comparisons to other fitting-based methods, which are excluded from the main experiment for fairness.

Figure 7. Self reenactment results for Subject 253. RR+ER yields a slightly more conservative mouth-opening amplitude, with red lines marking the lower lip for reference.

4.4.2 Cross-identity Reenactment

We visualize representative cross-identity reenactment results corresponding to the experiment in Section 4.3.2, perceptually comparing our approach against two per-frame FLAME fitting baselines: DECA [8] and SMIRK [23].

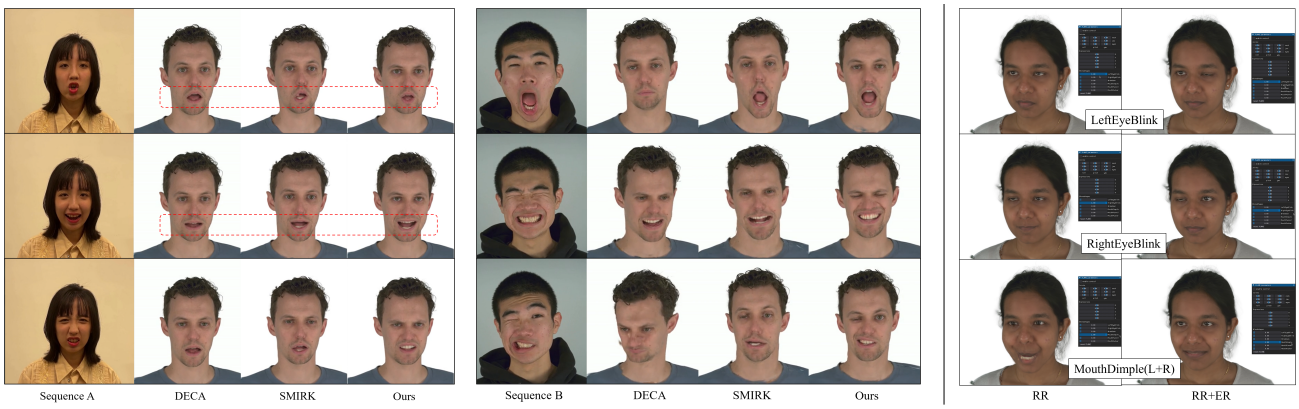
As shown in Figure 8a, our method better preserves the structural fidelity than DECA and SMIRK. In these frames, the baseline methods often exhibit smoothing artifacts in the mouth region (red box) and lack expressive richness. Our approach, by contrast, maintains sharper and more plausible facial deformations that better match the input intent while remaining visually stable.

Figure 8b further highlights the performance gap under extreme expressions. DECA exhibits frequent fitting failures in multiple frames, misaligning basic facial landmarks. SMIRK, while better aligned, suffers from severe mesh artifacts such as tearing and unnatural compression, particularly under exaggerated facial deformations. In contrast, our method consistently generates coherent outputs that preserve avatar integrity.

We also observe that temporal jittering—common in per-frame fitting approaches—is mitigated in our method. These observations are consistent with our earlier quantitative findings, reinforcing the effectiveness of a target-aware reenactment strategy—particularly in scenarios where animation stability and identity preservation are prioritized over exact frame-wise matching.

4.4.3 Blendshape Disentanglement

To qualitatively evaluate the semantic disentanglement capacity of our expression mapping, we manually activate individual blendshapes using the GUI interface and observe the resulting avatar deformations. Figure 8c compares the



(a) Visualization of Sequence A on Subject 306. (b) Visualization of Sequence B on Subject 306. (c) Blendshape disentanglement test.

Figure 8. Cross-identity reenactment visualization (8a, 8b) shows that our method preserves sharper geometry, structural consistency, and emotional accuracy, whereas baseline methods exhibit varying degrees of artifacts or loss of detail. Blendshape disentanglement test (8c) indicates that, under single-blendshape activation, RR+ER yields more localized and semantically coherent deformations compared to the plain RR model.

output of two models—RR and RR+ER—under identical single-blendshape activation.

Across multiple expression types, the RR+ER model consistently yields more localized and coherent responses. It avoids common entanglement artifacts observed in the RR baseline, such as bilateral eye motion during unilateral blinks, jaw deformation during eyelid activation, or over-amplified mouth dimples. These entanglements likely stem from uncontrolled extrapolation in sparsely represented regions of the training distribution, where the baseline projection matrix fails to preserve semantic independence between expression dimensions.

These results highlight the effectiveness of the expression regularization as a structural prior that encourages interpretable and stable projection behavior.

5. Conclusion

5.1. Limitations and Future Work

1. Expression Diversity Depends on Training Data Our model achieves stable reenactment by avoiding extrapolation into unfamiliar expression regions. However, this conservative behavior also limits expressiveness when faced with out-of-distribution or high-strain facial inputs. To improve fidelity in emotionally rich or exaggerated performances, future work should incorporate training data with greater variation in speaking styles and facial dynamics.

2. Input Modality and Cross-Platform Generalization While our experiments leverage ARKit due to its accessibility and high-quality facial tracking, our method is not restricted to this specific input modality. The regression-based mapping framework is compatible with other mocap systems that provide FACS [7]-like blendshape representations, including open-source alternatives. Future work may systematically explore cross-device performance and generalization, broadening deployment scenarios across platforms and capture conditions.

3. Integration of Dynamic Gaussian Rendering in VR Despite recent advances, real-time rendering of dynamic Gaussian scenes on VR headsets remains challenging. The current ecosystem for dynamic neural rendering in immersive environments is still immature. Future work may explore efficient rendering strategies or hybrid representations to enable high-fidelity avatar control in VR settings.

5.2. Summary

We present a real-time facial animation framework that maps facial motion capture inputs to parametric expression representations via efficient regression, enhanced by an expression regularization mechanism for semantic disentanglement and temporal stability. Evaluations show that lightweight subject-specific mappings can outperform competitive fitting-based pipelines and baseline regression models in real-time scenarios, suggesting a shift toward personalized control systems in avatar animation.

The proposed method proves effective not only for subject-specific control, but also generalizes well across unseen actors in cross-identity reenactment, accurately preserving expression semantics. The system supports live and offline modes, enabling flexible deployment for real-time and post-processed applications using commodity devices.

These findings position our system as a promising step toward real-time, controllable virtual humans with photo-realistim, laying a foundation for broader applications in telepresence, virtual communication, and interactive media.

References

[1] Face cap - motion capture for ios. https://www.bannaflak.com/face-cap/. Accessed: 2025-08-06. 2, 4

[2] Apple Inc. ARKit: Apple's Augmented Reality Platform. https://developer.apple.com/augmented-reality/arkit/, 2024. Accessed: 2025-05-30. 1

[3] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014. 2

[4] Radek Danecek, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20311–20322, 2022. 2

[5] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set, 2020. 2

[6] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3d morphable face models – past, present and future, 2020. 1

[7] Paul Ekman and Wallace V. Friesen. Facial action coding system: A technique for the measurement of facial movement. Technical report, Consulting Psychologists Press, Palo Alto, CA, 1978. 2, 8

[8] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. 2021. 2, 5, 6, 7

[9] Wilko Guilluy, Azeddine Beghdadi, and Laurent Oudre. A performance evaluation framework for video stabilization methods. In *2018 7th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6, 2018. 5

[10] Jerin Geo James, Devansh Jain, and Ajit Rajwade. Globalflownet: Video stabilization using deep distilled global motion estimates, 2022. 5

[11] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1, 2

[12] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 42(4), 2023. 1, 3, 5, 6

[13] Chunlu Li, Andreas Morel-Forster, Thomas Vetter, Bernhard Egger, and Adam Kortylewski. Robust model-based face reconstruction through weakly-supervised outlier segmentation, 2023. 2

[14] Linzhou Li, Yumeng Li, Yanlin Weng, Youyi Zheng, and Kun Zhou. Rgbavatar: Reduced gaussian blendshapes for online modeling of head avatars. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 10747–10757, 2025. 2

[15] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 1, 2

[16] Jun Ling, Xu Tan, Liyang Chen, Runnan Li, Yuchao Zhang, Sheng Zhao, and Li Song. Stableface: Analyzing and improving motion stability for talking face generation, 2022. 5

[17] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. *arXiv preprint arXiv:2203.05297*, 2022. 2

[18] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J. Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling, 2024. 2, 4

[19] Shengjie Ma, Yanlin Weng, Tianjia Shao, and Kun Zhou. 3d gaussian blendshapes for head avatar animation. In *ACM SIGGRAPH Conference Proceedings, Denver, CO, United States, July 28 - August 1, 2024*, 2024. 2

[20] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. Genova, Italy, 2009. IEEE. 2

[21] Shenhan Qian. Vhap: Versatile head alignment with adaptive appearance priors, 2024. 2, 4

[22] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 2, 3, 4

[23] George Retsinas, Panagiotis P. Filntisis, Radek Danecek, Victoria F. Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. Smirk: 3d facial expressions through analysis-by-neural-synthesis, 2025. 2, 5, 6, 7

[24] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[25] Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. Facediffuser: Speech-driven 3d facial animation synthesis using diffusion, 2023. 5

[26] Zhou Wang and Alan C. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009. 5

[27] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[28] 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*, pages 12780–12790, 2023. 5

[29] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with

a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 6

[30] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. 2023. 2