How do diffusion models learn personalized expressions? Global Representation Learning for Talking Portrait Generation

Anonymous authors
Paper under double-blind review

Abstract

In recent years, diffusion models have shown great potential in talking portrait generation. However, limitations still exist in personalized representation, particularly in appearance and motion, which we attribute to the current learning mechanisms. Regarding portrait appearance, the model tends to favor high-probability samples from the training data, leading to averaged results and weakened individual characteristics. In terms of motion, the model tends to produce locally plausible yet independent motion patterns, making it difficult to capture coordinated relationships in cross-identity portrait animation. To address these issues, we propose a talking portrait generation method based on global representation learning, which consists of two core components: alignment from group to individual (AG2I) and fusion of geometry and motion (FG&M). AG2I constructs the global appearance representations and performs group-to-individual alignment, thereby avoiding averaged appearance. FG&M constructs the global motion representations and leverages the fusion of geometric structures with multi-scale motion, enhancing motion coordination. Experimental results demonstrate that the proposed method effectively enhances the personalization and realism of the generated results compared to state-of-the-art methods.

1 Introduction

Audio-driven talking video generation has broad application prospects in digital humans, virtual interaction, and animation production Liu et al. (2026); Wang et al. (2026b); Xu et al. (2026); Nazarieh et al. (2026); Wang et al. (2026a). Although existing methods can generate visually realistic results, they often lack fine-grained personalized representation, particularly in capturing distinctive appearance characteristics and achieving cross-identity motion coordination.

In recent years, the generation of human-centric portrait animations has developed rapidly. Early methods primarily focused on precise lip-syncing Prajwal et al. (2020), while subsequent research introduced motion sources Zhou et al. (2021); Wang et al. (2023a) and emotion-based conditioning Tian et al. (2024); Choi et al. (2026) to enhance interactivity and expressiveness. In addition, multi-view synthesis techniques such as NeRF Li et al. (2024); Dong & Zhang (2025); Guo et al. (2021) and 3DGS Cho et al. (2024); Li et al. (2025); Yu et al. (2024) further improve geometric realism. Although these methods have achieved some progress, they still struggle to meet higher visual quality requirements. In this context, diffusion models have significantly improved the resolution and diversity of generated results Du et al. (2023); Sun et al. (2024); Hogue et al. (2024). However, the step-by-step denoising process in diffusion models often results in over-smoothed details and limited motion coherence, thereby reducing personalization.

We observe that distinctive individual characteristics are typically manifested as low-frequency and unstable signals. Although such signals are crucial for realism, they tend to be weakened and progressively homogenized during the multi-step noise-adding and denoising process in diffusion models Zhou et al. (2015); Ho et al. (2020). Moreover, the dependency between motion and portrait geometry often exhibits multi-granular and structurally diverse variations in continuous sequences Chen et al. (2021; 2020). Therefore, it is necessary to characterize their dynamic relationship from a multi-level manner. Based on the above observations, we focus on the learning mechanisms within

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.

the diffusion process and propose a talking portrait generation method based on global representation learning, introducing two core strategies: alignment from group to individual (AG2I) and fusion of geometry and motion (FG&M). In the AG2I, we extract portrait features to construct a global appearance representation and decouple them into individual and group features. We then use individual features to guide the alignment of group features, mitigating issues such as appearance averaging or over-smoothing. In the FG&M, we capture geometric variations across continuous sequences to construct a global motion representation and introduce multi-scale constraints to reflect the motion characteristics. Subsequently, the motion is fused with the portrait structure to enhance motion coordination in cross-identity generation.

The main contribution of this paper is a global representation learning framework for talking portrait generation. We first propose the AG2I (alignment from group to individual) strategy to enhance the individual characteristics of portrait appearance. We then introduce the FG&M (fusion of geometry and motion) strategy to improve motion coordination. Experimental results demonstrate that this method significantly improves the personalized expression of the generated results, outperforming state-of-the-art approaches.

2 Overview

The framework overview is illustrated in Figure 1, comprising two core strategies: alignment from group to individual (AG2I) and fusion of geometry and motion (FG&M).

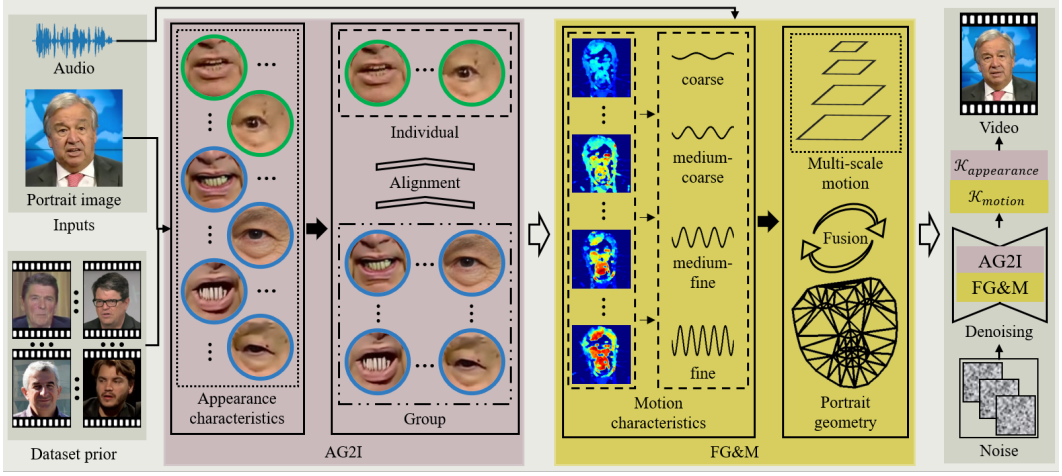

Figure 1: The framework of our method. The method consists of two parts: alignment from group to individual (AG2I) and fusion of geometry and motion (FG&M). AG2I performs group-to-individual characteristics alignment, thereby mitigating appearance homogenization. FG&M leverages the fusion of multi-scale motion and portrait geometry to enhance motion coordination.

In the AG2I module, we define the combination of training data priors and the input portrait as an appearance representation, which is adaptively updated during the generation process. Due to the influence of group-level statistical features, fine-grained regions are more prone to the degradation of individual-specific characteristics. To address this issue, we introduce a group-to-individual alignment mechanism that emphasizes personalized appearance features, effectively alleviating the over-smoothing of fine details. In the FG&M module, we observed significant differences in motion performance among different portrait structures, such as amplitude and frequency. Based on this observation, we define motion sequences as multi-scale representations to capture their dynamic characteristics. By leveraging a fusion mechanism that integrates geometry and motion information, we coordinate geometric variations across different structures under multi-scale constraints, thereby improving overall motion coherence.

Finally, as illustrated in Figure 1, by integrating the AG2I and FG&M modules, we obtain the global appearance alignment representation $\mathcal{K}_{appearance}$ and the global motion fusion representa-

tion $\mathcal{K}_{motion}$. These condition settings are then used to guide the denoising process, enabling the generation of high-quality talking portrait videos.

2.1 Preliminaries: Diffusion Models

The forward diffusion process is defined as:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon. \tag{1}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ and $\epsilon \sim \mathcal{N}(0, I)$. This represents a weighted linear combination of the clean data x_0 and Gaussian noise ϵ at time step t.

During training, the model learns a parameterized function $\epsilon_\theta(x_t, t)$ to predict the noise by minimizing the mean squared error, expressed as follows:

$$\mathcal{L}_{base} = \mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]. \tag{2}$$

which enables accurate estimation of perturbations under different noise levels.

2.2 Condition Settings

To enable personalized generation, a conditional variable c is introduced, leading to the conditional formulation $\epsilon_\theta(x_t, t, c)$, with an objective function expressed as follows:

$$\mathcal{L}(c) = \mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t, c)\|^2\right]. \tag{3}$$

Specifically, we construct a conditional setting based on AG2I and FG&M, and input it into the denoising network to obtain global appearance and motion representations, defined as follows:

$$c = \mathcal{F}_{condition}(\mathcal{K}_{appearance}, \mathcal{K}_{motion}). \tag{4}$$

where $\mathcal{F}_{condition}$ denotes a weighting function. By introducing conditional information, we are able to generate distinctive individual characteristics and capture the coordinated variations of fine-grained dynamics, thereby improving the personalization quality of the generated results.

3 Method

3.1 Alignment from Group to Individual

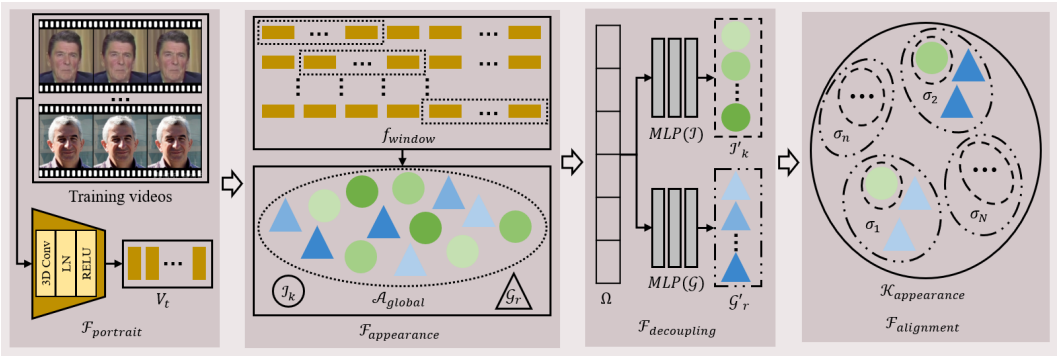

Figure 2: The process of alignment from group to individual (AG2I).

The AG2I process is shown in Figure 2. We first use the portrait encoder $\mathcal{F}_{portrait}$ to extract features from the frames $\{f_t\}_{t=1}^{T}$, as shown below:

$$\mathcal{F}_{portrait} : f_t \rightarrow V_t. \tag{5}$$

where $\mathcal{F}_{portrait}$ consists of lightweight 3D convolutions with channel-wise normalization, producing appearance features V_t that capture structural, texture, and identity information.

To address inconsistencies in temporal variations within the training data, we introduce a temporal gradient loss function to suppress discontinuous changes, defined as follows:

$$\mathcal{L}_{portrait} = \sum_{t} \|(V_t - V_{t-1}) - (V_{t-1} - V_{t-2})\|^2. \tag{6}$$

Next, we construct the global appearance domain representation $\mathcal{A}_{global}$. We designed a network $\mathcal{F}_{appearance}$ that further maps the initial appearance features V_t to individual features $\mathcal{I}_k$ and group features $\mathcal{G}_r$, as shown below:

$$\mathcal{F}_{appearance} : V_t \rightarrow (\mathcal{I}_k, \mathcal{G}_r). \tag{7}$$

where $\mathcal{F}_{appearance}$ utilizes a self-attention-based window structure, f_{window}, to integrate cross-frame information in the feature space while mitigating identity interference. $\mathcal{I}_k$ and $\mathcal{G}_r$ are jointly defined as $\mathcal{A}_{global}$.

Here, we introduce an appearance distribution alignment loss $\mathcal{L}_{appearance}$ to reduce the statistical distribution shift between the global appearance feature and frame-wise features, as follows:

$$\mathcal{L}_{appearance} = \left\| \mathrm{Cov}(\mathcal{A}_{global}) - \frac{1}{T} \sum_{t} \mathrm{Cov}(V_t) \right\|_F. \tag{8}$$

where $\mathrm{Cov}(\cdot)$ denotes the covariance matrix that captures the statistical correlations among feature dimensions, and $\|\cdot\|_F$ denotes the Frobenius norm.

Subsequently, due to the influence of multi-identity features in the training data, individual features and group features remain coupled to some extent. To address this, we introduce a decoupling network $\mathcal{F}_{decoupling}$, as shown in the following formula:

$$\mathcal{F}_{decoupling} : \mathrm{MLP}(\mathcal{I}_k) \rightarrow \mathcal{I}'_k, \quad \mathrm{MLP}(\mathcal{G}_r) \rightarrow \mathcal{G}'_r. \tag{9}$$

where $\mathcal{F}_{decoupling}$ employs a lightweight dual-branch multilayer perceptron (MLP), with each branch independently mapping individual and group features, thereby reducing feature coupling.

To enforce feature independence and information preservation, we introduce the loss function:

$$\mathcal{L}_{decoupling} = \|\mathrm{Cov}(\mathcal{I}'_k, \mathcal{G}'_r)\|_F^2 + (1 - \cos(\mathcal{I}'_k + \mathcal{G}'_r, \mathcal{A}_{global})). \tag{10}$$

where the first term is designed to reduce feature correlation, whereas the second term is introduced to maintain reconstruction consistency and alleviate information loss.

Finally, we introduce an alignment mechanism $\mathcal{F}_{alignment}$ to guide the group features $\mathcal{G}'_j$ driven by individual features $\mathcal{I}'_k$, obtaining the aligned feature σ_n, as follows:

$$\mathcal{F}_{alignment} : \mathcal{G}'_r + \gamma(\mathcal{I}'_k) \odot \mathcal{G}'_r + \beta(\mathcal{I}'_k) \rightarrow \sigma_n. \tag{11}$$

where $\gamma(\cdot)$ and $\beta(\cdot)$ denote channel modulation functions. Then, σ_n is integrated into the global appearance alignment representation $\mathcal{K}_{appearance}$, as follows:

$$\mathcal{K}_{appearance} = \mathcal{T}(\{\sigma_n\}_{n=1}^{N}). \tag{12}$$

where $\mathcal{T}(\cdot)$ denotes an aggregation operation.

We introduce loss $\mathcal{L}_{alignment}$ to enhance the consistency between $\mathcal{K}_{appearance}$ and $\mathcal{I}'_k$, while suppressing their drift toward $\mathcal{G}'_r$, as follows:

$$\mathcal{L}_{alignment} = \|\phi(\mathcal{I}'_k) - \phi(\mathcal{K}_{appearance})\|^2 - \|\phi(\mathcal{G}'_r) - \phi(\mathcal{K}_{appearance})\|^2. \tag{13}$$

where $\phi(\cdot)$ denotes a linear projection used to measure feature differences in a unified space.

3.2 Fusion of Geometry and Motion

The FG&M process is illustrated in Figure 3. First, we introduce the network $\mathcal{F}_{scale}$ to extract joint features $\mathcal{U}_t^q$, including motion variations and portrait geometric information, as shown below:

$$\mathcal{F}_{scale} : f_t \rightarrow \mathcal{U}_t^q. \tag{14}$$

where $\mathcal{F}_{scale}$ is based on multi-branch convolution, and q denotes the scale index.

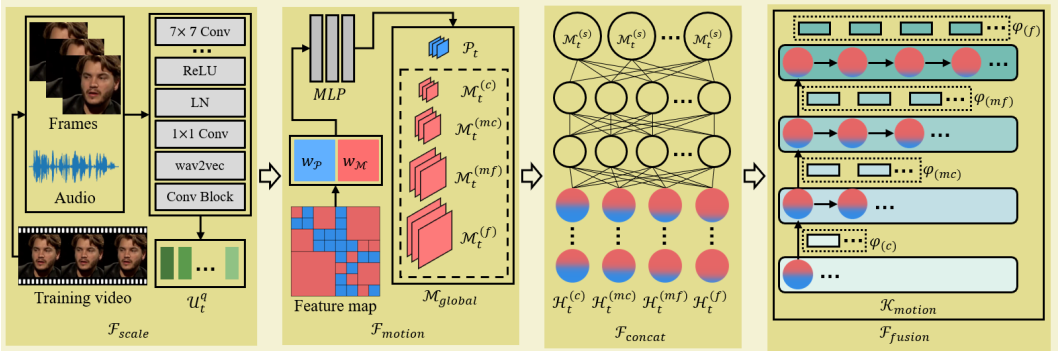

Figure 3: The process of fusion of geometry and motion (FG&M).

We further impose a cross-scale consistency constraint to maintain smooth transitions between adjacent motion scales, as shown below:

$$\mathcal{L}_{scale} = \sum_{q=2}^{Q} \left\| \mathcal{D}(\mathcal{U}_t^q) - \mathcal{D}(\mathcal{U}_t^{q-1}) \right\|_2^2. \tag{15}$$

where $\mathcal{D}(\cdot)$ denotes the scale transformation, and Q represents the total number of motion scales.

Next, we construct the global motion domain representation $\mathcal{M}_{global}$. Specifically, we introduce an attention-weighted network $\mathcal{F}_{motion}$ to compute dynamic weights:

$$\mathcal{F}_{motion} : \mathcal{U}_t^q \rightarrow (\omega_{\mathcal{P}}, \omega_{\mathcal{M}}). \tag{16}$$

where $\omega_{\mathcal{P}}$ and $\omega_{\mathcal{M}}$ denote the weights for structural and multi-scale motion information, respectively. Based on this, we employ an MLP to obtain the structural features $\mathcal{P}_t$ and multi-scale motion features $\{\mathcal{M}_t^q\}_{q=1}^{Q}$, which jointly constitute $\mathcal{M}_{global}$. In our implementation, we set $Q = 4$, corresponding to four levels of granularity: $\mathcal{M}_t^{(c)}$ (coarse-grained), $\mathcal{M}_t^{(mc)}$ (medium-coarse grained), $\mathcal{M}_t^{(mf)}$ (medium-fine grained), and $\mathcal{M}_t^{(f)}$ (fine-grained).

Here, we introduce a decorrelation loss $\mathcal{L}_{motion}$ to reduce feature redundancy by minimizing the correlation among feature dimensions within each motion scale, defined as follows:

$$\mathcal{L}_{motion} = \sum_{q=1}^{Q} \left\| \mathrm{Cov}\left(\mathcal{M}_t^q\right) - \mathbf{I} \right\|_F^2. \tag{17}$$

where $\mathbf{I}$ denotes the identity matrix.

Then, we obtain the scale-aware representation $\mathcal{H}_t^q$ by reorganizing $\mathcal{M}_{global}$ to further refine the interaction between structure and multi-scale motion, as follows:

$$\mathcal{F}_{concat} : \mathcal{M}_{global} \rightarrow \mathcal{H}_t^q, \quad q \in \{c, mc, mf, f\}. \tag{18}$$

where $\mathcal{F}_{concat}$ is implemented with a Transformer to capture cross-scale dependencies.

To avoid scale collapse, we introduce an entropy-regularized loss function, as follows:

$$\mathcal{L}_{concat} = -\sum_{s} w_q \log w_q. \tag{19}$$

where w_q denotes the normalized scale weight.

Finally, for the structure-aware representation $\mathcal{H}_t^q$ at each scale, we introduce a progressive layer-wise fusion mechanism $\mathcal{F}_{fusion}$ to obtain the fused features φ_q, as follows:

$$\mathcal{F}_{fusion} : \{\mathcal{H}_t^q\} \rightarrow \varphi_q, \quad q \in \{c, mc, mf, f\}. \tag{20}$$

The fused motion features at different scales are concatenated using $Concat(\cdot)$ to obtain the final global fused motion representation $\mathcal{K}_{motion}$, as follows:

$$\mathcal{K}_{motion} = Concat(\{\varphi_q\}). \tag{21}$$

We introduce a spatial constraint that enhances motion consistency across different regions:

$$\mathcal{L}_{fusion} = \sum_{i} \sum_{j \in \mathcal{N}(i)} S_{ij} \left\| \nabla E^q(i) - \nabla E^q(j) \right\|^2 . \tag{22}$$

where ∇E denotes the spatial gradient, S_{ij} represents a weight based on structural similarity, and $\mathcal{N}(i)$ denotes the spatial neighborhood.

4 EXPERIMENT

We first evaluate the generation quality through qualitative and quantitative experiments, and analyze the contributions of key components through ablation studies. Furthermore, we discuss the applications of our model. The appendix provides details on related work (Section A), a lightweight version of our method (Section C), user studies (Section D.1), model performance (Section D.2), parameter settings (Section D.3), and limitations (Section D.4). We also recommend viewing the supplementary video (*supplementary.mp4*), which provides a dynamic demonstration.

Implementation details. The model is trained and evaluated on a server equipped with 8 NVIDIA A100 GPUs. To ensure fairness and generalization, we construct a hybrid dataset that combines public datasets (LRS3-TED Afouras et al. (2018), SpeakerVid5M Zhang et al. (2025), and HDTF Zhang et al. (2021)) with self-collected YouTube videos (e.g., interviews and speeches). After filtering, the dataset comprises approximately 1,200 hours of 720p videos, covering diverse scenarios and speakers, and is split into training and testing sets with a 7:3 ratio.

Comparison methods. We compared our method with other diffusion-based approaches: StreamAvatar Sun et al. (2026), EchoMimicV3 Meng et al. (2026), Hallo3 Cui et al. (2025), MoDiTalker Kim et al. (2025), AniPortrait Wei et al. (2024), and V-Express Wang et al. (2024).

4.1 QUALITATIVE COMPARISON

As shown in Figure 4, Figure 4 (a) presents the reconstruction results of tooth details across consecutive frames, while Figure 4 (b) compares the consistency of facial wrinkle details and eye/eyebrow movements between non-adjacent frames. The Ground Truth displays the actual video corresponding to the input audio.

StreamAvatar adopts a two-stage framework with autoregressive distillation and adversarial optimization, but suffers temporal inconsistency and limited motion diversity in long sequences. EchoMimicV3 leverages multimodal cross-attention for unified animation, yet exhibits detail distortion and identity drift over long sequences. Hallo3 builds on a video Transformer with identity preservation, but its diffusion-based design limits efficiency and introduces artifacts. MoDiTalker uses a two-stage diffusion framework to improve lip-sync, but its keypoint-based driving restricts motion expressiveness. AniPortrait employs 3D representations for temporal stability, yet often produces unrealistic details. V-Express supports multi-condition control, but underperforms in lip-sync, motion coordination, and efficiency.

In contrast, our method enhances identity consistency and personalized expression by aligning group-level features under the guidance of individual-specific characteristics. Furthermore, we introduce a multi-level motion constraint mechanism that dynamically adjusts the geometric contribution of different regions, thereby improving the coordinated movement across facial areas such as teeth, wrinkles, and eyes. Thanks to these designs, our method outperforms existing approaches in terms of detail authenticity, temporal consistency, and motion naturalness, significantly elevating the overall sense of realism and immersion.

4.2 QUANTITATIVE COMPARISON

As shown in Table 1, we evaluated the proposed method from multiple perspectives using standard metrics, including image quality (PSNR, SSIM, LPIPS) Wang et al. (2004), lip synchronization (SyncNet, LSE-C) Chung & Zisserman (2017); Prajwal et al. (2020), motion modeling (AUE, LMD) Chen et al. (2018), identity preservation (FID, CSIM) Heusel et al. (2017); Huang et al. (2020), and inference efficiency (FPS) Ye et al. (2023); Chen et al. (2019).

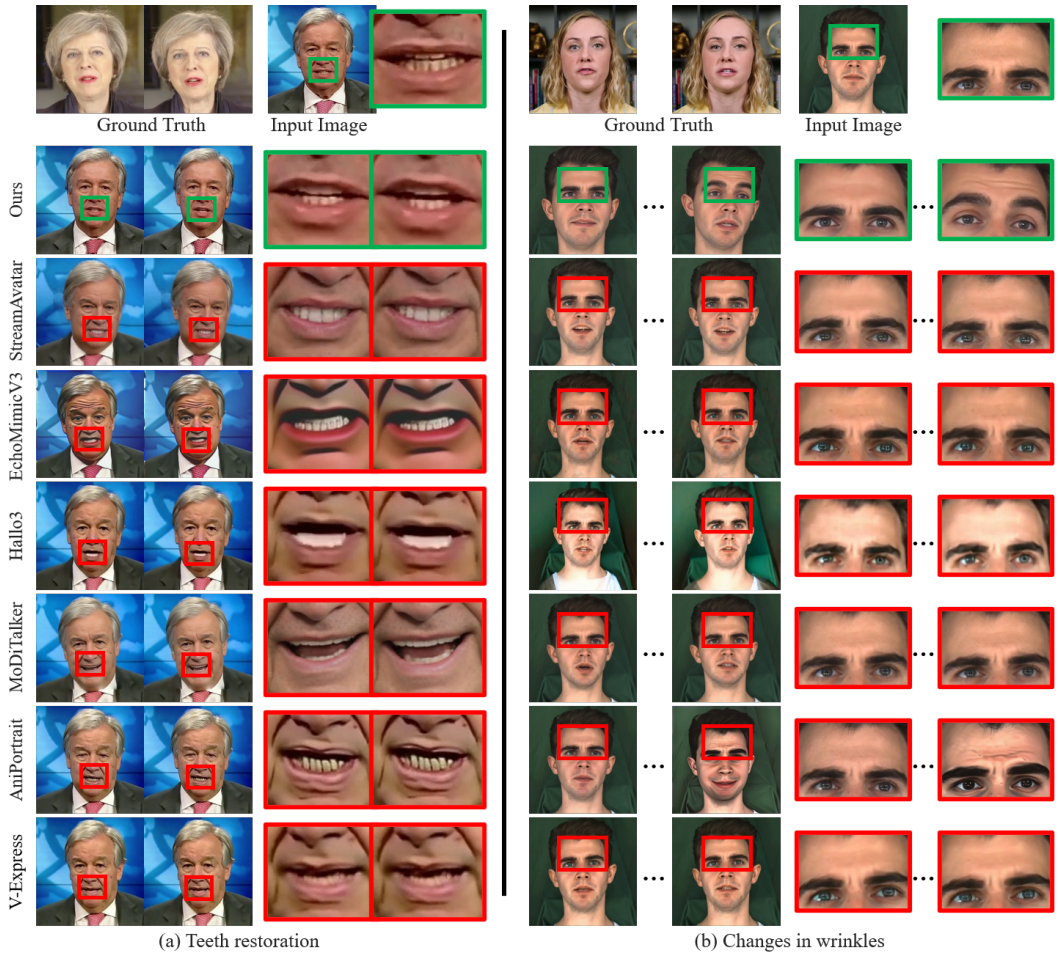

Figure 4: Qualitative results. Green boxes indicate the magnified regions of the Input Image and Ours, while red boxes indicate the corresponding regions of the comparison methods, where varying degrees of personalized detail loss can be observed.

Table 1: Quantitative results.

Methods	PSNR ↑	LPIPS ↓	SyncNet ↑	LSE-C ↑	LMD ↓	CSIM ↑	FPS ↑
Ground Truth	N/A	N/A	8.743	7.937	N/A	1	N/A
Ours	**37.65**	**0.01**	**8.465**	**7.615**	**1.04**	**0.96**	**10.16**
StreamAvatar	32.57	0.07	7.944	7.172	1.74	0.91	5.154
EchoMimicV3	33.71	0.04	8.115	7.049	1.75	0.89	0.297
Hallo3	34.27	0.05	7.031	7.265	1.63	0.89	0.065
MoDiTalker	33.18	0.09	7.149	7.164	1.34	0.84	0.758
AniPortrait	35.29	0.03	7.442	7.067	1.35	0.92	2.047
V-Express	31.45	0.06	7.763	7.265	1.63	0.85	0.746

StreamAvatar is efficient but suffers from motion inconsistency and identity drift in long sequences. EchoMimicV3 and V-Express enable multimodal conditioning but fall short in visual quality, alignment, and efficiency. Hallo3 and AniPortrait maintain structural stability yet exhibit artifacts and limited detail realism. MoDiTalker achieves stable lip-sync but shows limited motion expressiveness. Overall, identity drift, motion inconsistency, and efficiency remain key challenges, especially for long-sequence and real-time generation. In contrast, our method improves motion coordination while preserving structural fidelity and visual realism. It also achieves more stable identity preservation and synchronization performance, while maintaining competitive inference efficiency.

4.3 Ablation Study

To evaluate the contributions of AG2I and FG&M, we conducted ablation experiments, as shown in Figure 5 and Table 2. All experiments were conducted using identical training configurations and datasets to ensure a fair comparison.



Figure 5: Qualitative results of the ablation study.

Table 2: Quantitative ablation study.

Methods	PSNR ↑	LPIPS ↓	SyncNet ↑	LSE-C ↑	LMD ↓	CSIM ↑
Ground Truth	N/A	N/A	8.57	7.628	N/A	1
Ours	**37.81**	**0.01**	**8.531**	**7.533**	**1.46**	**0.96**
Ours w/o $\mathcal{F}_{portrait}$	32.85	0.08	8.526	7.484	1.53	0.95
Ours w/o $\mathcal{F}_{appearance}$	33.62	0.09	8.485	7.526	1.51	0.95
Ours w/o $\mathcal{F}_{decoupling}$	33.61	0.07	8.428	7.522	1.49	0.94
Ours w/o $\mathcal{F}_{alignment}$	32.57	0.09	8.507	7.394	1.53	0.94
Ours w/o $\mathcal{F}_{scale}$	37.74	0.03	8.128	7.194	2.77	0.95
Ours w/o $\mathcal{F}_{motion}$	37.69	0.02	8.083	7.039	2.05	0.94
Ours w/o $\mathcal{F}_{concat}$	37.74	0.03	8.084	7.042	2.61	0.94
Ours w/o $\mathcal{F}_{fusion}$	37.58	0.02	8.131	7.043	2.88	0.95

Ours w/o $\mathcal{F}_{portrait}$. $\mathcal{F}_{portrait}$ extracts texture and dynamic features from input video frames. Without this module, the model lacks temporal information, potentially causing slight flickering or jitter in local areas such as the mouth and eyes.

Ours w/o $\mathcal{F}_{appearance}$. $\mathcal{F}_{appearance}$ aggregates information across frames to obtain a unified appearance representation. Removing this module weakens the consistency of appearance features across frames, which may introduce subtle variations in texture and identity-related details.

Ours w/o $\mathcal{F}_{decoupling}$. $\mathcal{F}_{decoupling}$ separates appearance representations into individual-specific and group-level features. Without this module, group-level features tend to dominate the appearance representation, resulting in reduced preservation of fine-grained identity details.

Ours w/o $\mathcal{F}_{alignment}$. $\mathcal{F}_{alignment}$ enables individual features to guide the modulation of group features, ensuring that personalized characteristics are maintained in key facial regions. Removing this module weakens the influence of individual features, producing more generic facial expressions and less distinctive details in regions such as contours and expressions.

Ours w/o $\mathcal{F}_{scale}$. $\mathcal{F}_{scale}$ captures dynamic variations across different scales, from coarse-grained to fine-grained levels. Without this module, the motion representation capability is weakened, leading to degraded temporal continuity and slight inter-frame jitter in local regions, such as the mouth.

Ours w/o $\mathcal{F}_{motion}$. $\mathcal{F}_{motion}$ integrates structural information with multi-scale motion features to construct a unified motion representation. Removing it makes it difficult to distinguish motion patterns at different scales, resulting in over-smoothed motion representations.

Ours w/o $\mathcal{F}_{concat}$. $\mathcal{F}_{concat}$ facilitates the interaction between structure and motion. Without this module, such interaction is weakened, leading to inaccurate coordination between facial motion and structure, as well as local artifacts.

Ours w/o $\mathcal{F}_{fusion}$. $\mathcal{F}_{fusion}$ progressively integrates geometric and motion features across multiple scales. Removing this module limits the effectiveness of multi-scale feature aggregation, resulting in less coherent motion patterns and local discontinuities, such as jitter around facial boundaries.

Overall, both quantitative metrics and visual comparisons demonstrate that all ablation variants exhibit noticeable performance degradation across different evaluation dimensions. In contrast, the full model (Ours) generates results that are more consistent with the real videos and achieves superior performance, particularly in preserving personalized details and maintaining motion coordination.

4.4 Discussion

Our method can reconstruct personalized portrait characteristics to a certain extent, but it still depends on the input portrait. For instance, visible attributes (such as teeth) can be enhanced when present, whereas missing details rely primarily on prior information from the dataset, which limits the application scenarios. To address this issue, we introduce a portrait attribute estimation strategy that infers latent personalized features from high-level semantics such as age and skin condition. These attributes are then used to adaptively modulate the generation parameters, thereby improving the plausibility of missing details (see Supplementary Material B3: Parameter Settings). As shown in Figure 6, even when such attributes are absent in the input, the generated results better align with the real data distribution.

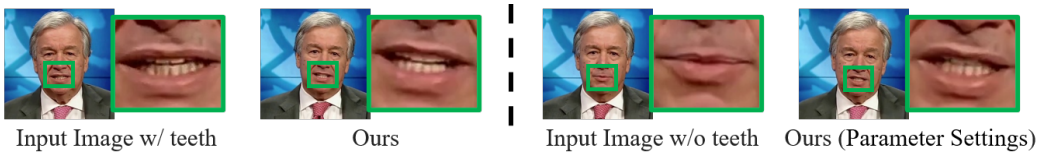

Figure 6: Our method can infer missing personalized details.

5 Conclusion

This paper proposes a talking portrait generation method based on global representation learning, aiming to enhance the personalized expression of portraits. We first propose AG2I (alignment from group to individual), which reduces the over-averaging of details through group-to-individual alignment. We then introduce FG&M (fusion of geometry and motion), which enhances motion coordination by fusing portrait geometry and motion information. Experimental results demonstrate that the proposed method achieves significant improvements in personalized expression capability compared to state-of-the-art methods.

AI USE STATEMENT

We used generative AI tools solely to improve the clarity and readability of the English writing. All research ideas, methodology, experiments, results, and interpretations were developed by the authors. All AI-assisted text was reviewed by the authors, who take full responsibility for the final content.

ETHICS STATEMENT

This work does not involve human subjects, sensitive personal data, or other foreseeable ethical concerns. All data used in this study are publicly available, and the experiments were conducted in accordance with applicable research integrity and data usage guidelines.

REFERENCES

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision*, pp. 520–535, 2018.

Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7832–7841, 2019.

Shu-Yu Chen, Wanchao Su, Lin Gao, Shihong Xia, and Hongbo Fu. Deepfacedrawing: deep generation of face images from sketches. *ACM Trans. Graph.*, 39(4), August 2020. ISSN 0730-0301. doi: 10.1145/3386569.3392386. URL https://doi.org/10.1145/3386569.3392386.

Shu-Yu Chen, Feng-Lin Liu, Yu-Kun Lai, Paul L. Rosin, Chunpeng Li, Hongbo Fu, and Lin Gao. Deepfaceediting: deep face generation and editing with disentangled geometry and appearance control. *ACM Trans. Graph.*, 40(4), July 2021. ISSN 0730-0301. doi: 10.1145/3450626.3459760. URL https://doi.org/10.1145/3450626.3459760.

Yi Chen, Sen Liang, Zixiang Zhou, Ziyao Huang, Yifeng Ma, Junshu Tang, Qin Lin, Yuan Zhou, and Qinglin Lu. Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters. *arXiv preprint arXiv:2505.20156*, 2025.

Kyusun Cho, Joungbin Lee, Heeji Yoon, Yeobin Hong, Jaehoon Ko, Sangjun Ahn, and Seungryong Kim. Gaussiantalker: Real-time talking head synthesis with 3d gaussian splatting. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10985–10994, 2024.

Chanhyuk Choi, Taesoo Kim, Donggyu Lee, Siyeol Jung, and Taehwan Kim. Cross-modal emotion transfer for emotion editing in talking face video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1759–1770, 2026.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pp. 251–263, 2017.

Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21086–21095, 2025.

Biao Dong and Lei Zhang. Talking head generation via viewpoint and lighting simulation based on global representation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10258–10267, 2025.

Chenpeng Du, Qi Chen, Tianyu He, Xu Tan, Xie Chen, Kai Yu, Sheng Zhao, and Jiang Bian. Dae-talker: High fidelity speech-driven talking face generation with diffusion autoencoder. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4281–4289, 2023.

Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5784–5794, 2021.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Steven Hogue, Chenxu Zhang, Hamza Daruger, Yapeng Tian, and Xiaohu Guo. Diffted: One-shot audio-driven ted talk video generation with diffusion-based co-speech gestures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1922–1931, 2024.

Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5901–5910, 2020.

Seyeon Kim, Siyoon Jin, Jihye Park, Kihong Kim, Jiyoung Kim, Jisu Nam, and Seungryong Kim. Moditalker: Motion-disentangled diffusion model for high-fidelity talking head generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4302–4310, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Dongze Li, Kang Zhao, Wei Wang, Bo Peng, Yingya Zhang, Jing Dong, and Tieniu Tan. Ae-nerf: Audio enhanced neural radiance field for few shot talking head synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3037–3045, 2024.

Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Jun Zhou, and Lin Gu. Instag: Learning personalized 3d talking head from few-second video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10690–10700, 2025.

Tao Liu, Feilong Chen, Shuai Fan, Chenpeng Du, Qi Chen, Xie Chen, and Kai Yu. Anitalker: Animate vivid and diverse talking faces through identity-decoupled facial motion encoding. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6696–6705, 2024.

Zhenjie Liu, Jianzhang Lu, Renjie Lu, Cong Liang, and Shangfei Wang. Consisttalk: Intensity controllable temporally consistent talking head generation with diffusion noise search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7449–7457, 2026.

Yifeng Ma, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yingya Zhang, and Zhidong Deng. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv preprint arXiv:2312.09767*, 2(3), 2023.

Rang Meng, Yan Wang, Weipeng Wu, Ruobing Zheng, Yuming Li, and Chenguang Ma. Echomimicv3: 1.3 b parameters are all you need for unified multi-modal and multi-task human animation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8008–8015, 2026.

Fatemeh Nazarieh, Zhenhua Feng, Diptesh Kanojia, Josef Kittler, and Muhammad Awais. Syncdreamer: Controllable and expressive avatar generation beyond the talking head. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25121–25130, 2026.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 484–492, 2020.

Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1982–1991, 2023.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yongjin Liu. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics*, 43(4):1–9, 2024.

Zhiyao Sun, Ziqiao Peng, Yifeng Ma, Yi Chen, Zhengguang Zhou, Zixiang Zhou, Guozhen Zhang, Youliang Zhang, Yuan Zhou, Qinglin Lu, et al. Streamavatar: Streaming diffusion models for real-time interactive human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10887–10897, 2026.

Kim Sung-Bin, Lee Chae-Yeon, Gihun Son, Oh Hyun-Bin, Janghoon Ju, Suekyeong Nam, and Tae-Hyun Oh. Multitalk: Enhancing 3d talking head generation across languages with multilingual video dataset. *arXiv preprint arXiv:2406.14272*, 2024.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pp. 244–260. Springer, 2024.

Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-express: Conditional dropout for progressive training of portrait video generation. *arXiv preprint arXiv:2406.02511*, 2024.

Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17979–17989, 2023a.

Haotian Wang, Yuzhe Weng, Jun Du, Haoran Xu, Xiaoyan Wu, Shan He, Bing Yin, Cong Liu, Jianqing Gao, and Qingfeng Liu. Read: Real-time and efficient asynchronous diffusion for audio-driven talking head generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9766–9774, 2026a.

MengChao Wang, Wang Qiang, Fan Jiang, and Mu Xu. Fantasytalking2: Timestep-layer adaptive preference optimization for audio-driven portrait animation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9966–9974, 2026b.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36:7594–7611, 2023b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024.

Haolan Xu, Keli Cheng, Lei Wang, Ning Bi, and Xiaoming Liu. Emotag: Emotion-aware talking head synthesis on gaussian splatting with few-shot personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10921–10931, 2026.

Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024a.

Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *Advances in Neural Information Processing Systems*, 37:660–684, 2024b.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Zhenhui Ye, Jinzheng He, Ziyue Jiang, Rongjie Huang, Jiawei Huang, Jinglin Liu, Yi Ren, Xiang Yin, Zejun Ma, and Zhou Zhao. Geneface++: Generalized and stable real-time audio-driven 3d talking face generation. *arXiv preprint arXiv:2305.00787*, 2023.

Hongyun Yu, Zhan Qu, Qihang Yu, Jianchuan Chen, Zhonghua Jiang, Zhiwen Chen, Shengyu Zhang, Jimin Xu, Fei Wu, Chengfei Lv, et al. Gaussiantalker: Speaker-specific talking head synthesis via 3d gaussian splatting. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3548–3557, 2024.

Shenghai Yuan, Jinfa Huang, Yujun Shi, Yongqi Xu, Ruijie Zhu, Bin Lin, Xinhua Cheng, Li Yuan, and Jiebo Luo. Magictime: Time-lapse video generation models as metamorphic simulators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Youliang Zhang, Zhaoyang Li, Duomin Wang, Jiahe Zhang, Deyu Zhou, Zixin Yin, Xili Dai, Gang Yu, and Xiu Li. Speakervid-5m: A large-scale high-quality dataset for audio-visual dyadic interactive human generation. *arXiv preprint arXiv:2507.09862*, 2025.

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*, pp. 3661–3670, 2021.

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022.

Guoxu Zhou, Andrzej Cichocki, Yu Zhang, and Danilo P Mandic. Group component analysis for multiblock data: Common and individual feature extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 27(11):2426–2439, 2015.

Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4176–4186, 2021.

Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. Vlogger: Make your dream a vlog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8806–8817, 2024.

A Related Work

A.1 Diffusion-based Video Generation

Diffusion models have made significant progress in video generation. For example, Make-A-Video Singer et al. (2022) extends text-to-image generation to text-to-video generation by learning visual and semantic representations from image-text pairs, enabling high-quality video generation without requiring large-scale paired text-video data. MagicVideo Zhou et al. (2022) combines a low-dimensional latent space with an efficient 3D U-Net and inter-frame adapters to generate high-resolution videos while reducing computational costs. Stable Video Diffusion Blattmann et al. (2023) achieves multi-view consistency through carefully constructed datasets and a staged training strategy, while providing strong motion representation capabilities.

Some works further introduce additional conditions to improve generation controllability. VideoCrafter Chen et al. (2023) and MagicTime Yuan et al. (2025) use text, image, or temporal sequence constraints to guide video generation. AnimateDiff Guo et al. (2023) extends image models to video generation through plug-and-play motion modules and can be efficiently fine-tuned for different motion patterns. VideoComposer Wang et al. (2023b) employs motion vectors and spatiotemporal conditional encoding to improve cross-frame consistency. HunyuanVideo Kong et al. (2024) provides a large-scale open-source foundation model and achieves strong performance in visual quality, motion representation, and text alignment through optimized training strategies and architectural designs.

For long-term video generation, CogVideoX Yang et al. (2025) employs a diffusion transformer and a 3D variational autoencoder to achieve spatiotemporal compression and multi-resolution training, enabling the generation of semantically coherent long videos. Movie Gen Polyak et al. (2024) builds upon large-scale pretrained models to generate high-resolution videos while also supporting personalized generation and fine-grained text-based editing.

Overall, existing methods have made significant progress in generation quality, motion representation, multi-view consistency, and long-term video modeling, while also showing strong potential for controllability and personalization. However, temporal consistency and motion coherence remain unstable in complex scenes. In long-term generation, semantic consistency and fine-grained details may degrade, while personalized control across different identities or multiple characters remains limited, providing clear directions for future research.

A.2 Audio-driven Portrait Animation

In recent years, diffusion-based portrait video generation methods have made significant progress. These methods typically use audio as the driving signal to generate dynamic talking videos from static portrait images. Early works such as Vlogger Zhuang et al. (2024) and Hallo Xu et al. (2024a) employ end-to-end diffusion models to generate high-fidelity and temporally coherent talking videos. Vlogger uses a staged generation strategy to improve spatiotemporal consistency in long videos, while Hallo introduces a hierarchical audio-driven visual synthesis module to enhance the diversity of facial expressions and poses.

For improving generation consistency, some methods focus on temporal consistency across frames. EMO Tian et al. (2024) uses a direct audio-to-video mapping to maintain smooth transitions between frames and preserve identity consistency. AniPortrait establishes a mapping between 3D facial structures and 2D expressions and combines it with a motion module to achieve stable and natural portrait animation.

Meanwhile, improving emotional expressiveness and fine-grained facial dynamics has also become an important research direction. V-Express Wang et al. (2024) improves lip synchronization by balancing multiple control signals. VASA-1 Xu et al. (2024b) and AniTalker Liu et al. (2024) employ general motion representations and fine-grained expression modeling to generate diverse head movements and facial expressions, thereby improving the realism and synchronization of the generated videos.

In addition, some works focus on multi-character and multi-audio generation in more complex scenarios. MultiTalk Sung-Bin et al. (2024) and HunyuanVideo-Avatar Chen et al. (2025) support

multilingual and multi-character interactions, enabling highly dynamic and emotionally controllable video generation. DiffTalk Shen et al. (2023) and DreamTalk Ma et al. (2023) focus on cross-identity generation while preserving personalized characteristics and emotional expressions, together with temporal consistency and high visual quality.

Overall, these methods have made progress in generation accuracy, emotional expression, and multi-character scenario modeling. However, they still face limitations in capturing individual differences and personalized characteristics. For example, facial details, expression styles, and motion patterns can become averaged or weakened across different identities, which limits the realism and identity distinctiveness of the generated results.

B More Details of Our Method

This appendix provides the detailed mathematical formulation of the proposed alignment from group to individual (AG2I) and fusion of geometry and motion (FG&M) strategies. The derivations follow the computational flow of the main method, including feature extraction, temporal regularization, domain representation construction, feature decoupling, individual-guided alignment, multi-scale motion decomposition, cross-scale interaction, progressive fusion, and spatial consistency regularization.

B.1 Diffusion Process and Conditional Denoising

The forward diffusion process progressively perturbs a clean sample x_0 with Gaussian noise. Given the noise schedule $\{\alpha_t\}_{t=1}^{T}$, the cumulative product is defined as:

$$\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s. \tag{23}$$

The noisy sample at diffusion step t can therefore be directly sampled from x_0 as:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{24}$$

The conditional denoising network predicts the perturbation from the noisy sample and the condition variable c. The prediction error can be expressed as:

$$e_\theta = \epsilon - \epsilon_\theta(x_t, t, c). \tag{25}$$

The corresponding conditional denoising objective is:

$$\mathcal{L}(c) = \mathbb{E}_{x_0,\epsilon,t}\left[\|e_\theta\|_2^2\right]. \tag{26}$$

The condition variable combines the global appearance and motion representations:

$$c = \mathcal{F}_{condition}\left(\mathcal{K}_{appearance}, \mathcal{K}_{motion}\right). \tag{27}$$

For a weighted conditioning operation, the resulting representation can be generally written as:

$$c = \omega_A \mathcal{K}_{appearance} + \omega_M \mathcal{K}_{motion}. \tag{28}$$

where ω_A and ω_M denote the effective weights assigned to appearance and motion conditions, respectively. This formulation allows the denoising process to adaptively balance identity-preserving appearance information and motion-coordination information.

The gradient of the conditional objective with respect to the predicted noise is:

$$\frac{\partial \mathcal{L}(c)}{\partial \epsilon_\theta} = -2\left(\epsilon - \epsilon_\theta(x_t, t, c)\right). \tag{29}$$

Thus, optimization encourages $\epsilon_\theta(x_t, t, c)$ to approach the actual perturbation ϵ, while the condition c provides appearance and motion information required for personalized generation.

B.2 Temporal Portrait Representation

For an input frame sequence $\{f_t\}_{t=1}^{T}$, the portrait encoder extracts an appearance representation:

$$V_t = \mathcal{F}_{portrait}(f_t). \tag{30}$$

Let the feature dimension of V_t be d. The temporal variation between adjacent frames can be written as:

$$\Delta V_t = V_t - V_{t-1}. \tag{31}$$

The second-order temporal difference is then:

$$\Delta^2 V_t = \Delta V_t - \Delta V_{t-1} = V_t - 2V_{t-1} + V_{t-2}. \tag{32}$$

Accordingly, the temporal gradient loss in the main method can be expanded as:

$$\mathcal{L}_{portrait} = \sum_t \|V_t - 2V_{t-1} + V_{t-2}\|_2^2. \tag{33}$$

This formulation penalizes abrupt changes in the first-order temporal variation. In particular, when consecutive feature differences are consistent, $\Delta V_t \approx \Delta V_{t-1}$ and the corresponding penalty approaches zero. Therefore, the loss constrains temporal discontinuities while preserving the underlying appearance variation.

B.3 Global Appearance Domain Representation

The appearance network maps the frame-wise representation into individual-specific and group-level components:

$$\mathcal{F}_{appearance}(V_t) = (\mathcal{I}_k, \mathcal{G}_r). \tag{34}$$

For a temporal window containing L frames, the self-attention operation can be expressed by first constructing queries, keys, and values:

$$Q_V = VW_Q, \qquad K_V = VW_K, \qquad Z_V = VW_V. \tag{35}$$

where $V \in \mathbb{R}^{L\times d}$ denotes the stacked feature sequence in the window. The attention weights are obtained as:

$$A_V = \mathrm{softmax}\left(\frac{Q_V K_V^\top}{\sqrt{d}}\right). \tag{36}$$

The window-level feature can consequently be written as:

$$\widetilde{V} = A_V Z_V. \tag{37}$$

The individual and group branches then extract complementary components from the aggregated representation:

$$\mathcal{I}_k = \Psi_I(\widetilde{V}), \qquad \mathcal{G}_r = \Psi_G(\widetilde{V}). \tag{38}$$

where Ψ_I and Ψ_G denote the corresponding feature projections.

The global appearance representation is jointly formed by these two components:

$$\mathcal{A}_{global} = \mathcal{T}_A\left(\mathcal{I}_k, \mathcal{G}_r\right). \tag{39}$$

For a feature matrix $X \in \mathbb{R}^{N\times d}$, its covariance matrix is calculated after centering the feature dimensions. Specifically, let:

$$\mu_X = \frac{1}{N}\sum_{i=1}^{N} X_i. \tag{40}$$

and define the centered feature matrix:

$$\widetilde{X}_i = X_i - \mu_X. \tag{41}$$

The covariance matrix is then:

$$\mathrm{Cov}(X) = \frac{1}{N-1}\sum_{i=1}^{N} \widetilde{X}_i \widetilde{X}_i^{\top}. \tag{42}$$

Thus, the appearance distribution alignment objective can be written explicitly as:

$$\mathcal{L}_{appearance} = \left\| \mathrm{Cov}(\mathcal{A}_{global}) - \frac{1}{T}\sum_{t} \mathrm{Cov}(V_t) \right\|_F. \tag{43}$$

The Frobenius norm measures the overall discrepancy between the two covariance structures:

$$\|A - B\|_F^2 = \sum_i \sum_j (A_{ij} - B_{ij})^2. \tag{44}$$

Therefore, minimizing $\mathcal{L}_{appearance}$ encourages the global appearance representation to preserve the statistical structure of the frame-wise portrait features.

B.4 Individual–Group Feature Decoupling

The individual and group representations extracted from multi-identity training data may share correlated information. The decoupling network independently transforms the two branches:

$$\mathcal{I}'_k = \mathrm{MLP}_I(\mathcal{I}_k), \qquad \mathcal{G}'_r = \mathrm{MLP}_G(\mathcal{G}_r). \tag{45}$$

For two feature sets X and Y, their cross-covariance can be written as:

$$\mathrm{Cov}(X, Y) = \frac{1}{N-1}\sum_{i=1}^{N} (X_i - \mu_X)(Y_i - \mu_Y)^{\top}. \tag{46}$$

The first term of the decoupling objective therefore minimizes the cross-correlation between individual and group representations:

$$\mathcal{L}_{cross} = \left\| \mathrm{Cov}\left(\mathcal{I}'_k, \mathcal{G}'_r\right) \right\|_F^2. \tag{47}$$

The reconstruction-related term measures the consistency between the combined representation and the global appearance representation. Let:

$$Z_{IG} = \mathcal{I}'_k + \mathcal{G}'_r. \tag{48}$$

where Z_{IG} denotes the combined representation formed by the image-related representation $\mathcal{I}'_k$ and the Gaussian-related representation $\mathcal{G}'_r$.

The cosine similarity between Z_{IG} and $\mathcal{A}_{global}$ is:

$$\cos\left(Z_{IG}, \mathcal{A}_{global}\right) = \frac{Z_{IG}^{\top}\mathcal{A}_{global}}{\|Z_{IG}\|_2\|\mathcal{A}_{global}\|_2}. \tag{49}$$

Accordingly, the complete decoupling objective is:

$$\mathcal{L}_{decoupling} = \|\mathrm{Cov}\left(\mathcal{I}'_k, \mathcal{G}'_r\right)\|_F^2 + 1 - \frac{(\mathcal{I}'_k + \mathcal{G}'_r)^{\top}\mathcal{A}_{global}}{\|\mathcal{I}'_k + \mathcal{G}'_r\|_2\|\mathcal{A}_{global}\|_2}. \tag{50}$$

The first term suppresses shared statistical components, while the second term prevents excessive removal of information by maintaining consistency with the global appearance representation.

B.5 Individual-Guided Group-to-Individual Alignment

The aligned feature is obtained by modulating the group representation with individual-specific information. For each channel, the modulation functions generate an adaptive scale and bias:

$$\gamma_k = \gamma(\mathcal{I}'_k), \qquad \beta_k = \beta(\mathcal{I}'_k). \tag{51}$$

The channel-wise modulation can then be expanded as:

$$\sigma_n = \mathcal{G}'_r + \gamma_k \odot \mathcal{G}'_r + \beta_k. \tag{52}$$

Equivalently, the modulation can be expressed as:

$$\sigma_n = (1 + \gamma_k) \odot \mathcal{G}'_r + \beta_k. \tag{53}$$

Thus, γ_k controls the degree of individual-specific feature amplification or suppression, whereas β_k introduces an individual-dependent feature offset.

The global appearance alignment representation is obtained by aggregating the aligned features:

$$\mathcal{K}_{appearance} = \mathcal{T}\left(\{\sigma_n\}_{n=1}^{N}\right). \tag{54}$$

For an averaging aggregation, for example, this operation can be written as:

$$\mathcal{T}\left(\{\sigma_n\}_{n=1}^{N}\right) = \frac{1}{N}\sum_{n=1}^{N}\sigma_n. \tag{55}$$

More generally, $\mathcal{T}$ can perform learnable weighted aggregation:

$$\mathcal{K}_{appearance} = \sum_{n=1}^{N}\eta_n\sigma_n, \qquad \sum_{n=1}^{N}\eta_n = 1. \tag{56}$$

To measure the consistency between the individual representation and the resulting aligned representation, a shared projection $\phi(\cdot)$ maps the features into a common latent space:

$$z_I = \phi(\mathcal{I}'_k), \qquad z_A = \phi(\mathcal{K}_{appearance}), \qquad z_G = \phi(\mathcal{G}'_r). \tag{57}$$

The alignment objective can therefore be expressed as:

$$\mathcal{L}_{alignment} = \|z_I - z_A\|_2^2 - \|z_G - z_A\|_2^2. \tag{58}$$

The first term pulls the aligned representation toward the individual feature, whereas the negative second term prevents the aligned representation from drifting toward the group feature.

B.6 Multi-Scale Geometry and Motion Representation

The FG&M module first extracts joint geometric and motion information from each frame:

$$\mathcal{U}_t^q = \mathcal{F}_{scale}(f_t), \qquad q = 1, \ldots, Q. \tag{59}$$

For adjacent scales, the transformed representation is:

$$\widetilde{\mathcal{U}}_t^q = \mathcal{D}(\mathcal{U}_t^q). \tag{60}$$

The cross-scale difference is consequently defined as:

$$\Delta_q = \widetilde{\mathcal{U}}_t^q - \widetilde{\mathcal{U}}_t^{q-1}, \qquad q = 2, \ldots, Q. \tag{61}$$

The corresponding consistency loss becomes:

$$\mathcal{L}_{scale} = \sum_{q=2}^{Q} \|\Delta_q\|_2^2. \tag{62}$$

Minimizing this term encourages neighboring scales to exhibit smooth transitions after transformation, thereby reducing abrupt discrepancies between adjacent motion granularities.

B.7 Motion Domain Decomposition

The attention-weighted motion network produces the weights associated with structural and multi-scale motion information:

$$(\omega_{\mathcal{P}}, \omega_{\mathcal{M}}) = \mathcal{F}_{motion}(\mathcal{U}_t^q). \tag{63}$$

The weights can be normalized using a softmax operation:

$$\widetilde{\omega}_{\mathcal{P}} = \frac{\exp(\omega_{\mathcal{P}})}{\exp(\omega_{\mathcal{P}}) + \exp(\omega_{\mathcal{M}})}, \qquad \widetilde{\omega}_{\mathcal{M}} = \frac{\exp(\omega_{\mathcal{M}})}{\exp(\omega_{\mathcal{P}}) + \exp(\omega_{\mathcal{M}})}. \tag{64}$$

The structural and motion components can consequently be represented as:

$$\mathcal{P}_t = \Psi_{\mathcal{P}}\left(\widetilde{\omega}_{\mathcal{P}} \odot \mathcal{U}_t\right). \tag{65}$$

$$\mathcal{M}_t^q = \Psi_{\mathcal{M}}^q\left(\widetilde{\omega}_{\mathcal{M}} \odot \mathcal{U}_t^q\right). \tag{66}$$

The global motion representation jointly contains structural information and motion representations at multiple scales:

$$\mathcal{M}_{global} = \mathcal{T}_M \left(\mathcal{P}_t, \{\mathcal{M}_t^q\}_{q=1}^Q \right). \tag{67}$$

For $Q = 4$, the motion domain is organized into the four granularities:

$$\mathcal{M}_{global} = \left\{ \mathcal{P}_t, \mathcal{M}_t^{(c)}, \mathcal{M}_t^{(mc)}, \mathcal{M}_t^{(mf)}, \mathcal{M}_t^{(f)} \right\}. \tag{68}$$

B.8 Motion Decorrelation

For each motion scale, let the centered feature matrix be:

$$\widetilde{\mathcal{M}}_t^q = \mathcal{M}_t^q - \mu_{\mathcal{M}^q}. \tag{69}$$

Its covariance matrix is:

$$C_q = \mathrm{Cov}\left(\mathcal{M}_t^q\right) = \frac{1}{N-1} \left(\widetilde{\mathcal{M}}_t^q \right)^\top \widetilde{\mathcal{M}}_t^q. \tag{70}$$

The diagonal elements of C_q represent the variance of individual feature dimensions, while the off-diagonal elements represent pairwise correlations.

The motion decorrelation objective is:

$$\mathcal{L}_{motion} = \sum_{q=1}^{Q} \|C_q - \mathbf{I}\|_F^2. \tag{71}$$

Expanding the Frobenius norm gives:

$$\mathcal{L}_{motion} = \sum_{q=1}^{Q} \left[\sum_i (C_{q,ii} - 1)^2 + \sum_{i \neq j} C_{q,ij}^2 \right]. \tag{72}$$

Therefore, the objective simultaneously constrains the diagonal covariance terms toward unit variance and suppresses off-diagonal correlations, reducing redundant information among feature dimensions at each motion scale.

B.9 Scale-Aware Feature Reorganization

The global motion representation is reorganized into scale-aware representations using the concatenation and cross-scale interaction network:

$$\mathcal{H}_t^q = \mathcal{F}_{concat}\left(\mathcal{M}_{global}\right), \qquad q \in \{c, mc, mf, f\}. \tag{73}$$

The Transformer-based interaction can be represented by:

$$Q_q = \mathcal{H}_t^q W_Q, \qquad K_q = \mathcal{H}_t^q W_K, \qquad V_q = \mathcal{H}_t^q W_V. \tag{74}$$

The attention matrix is then:

$$A_q = \mathrm{softmax}\left(\frac{Q_q K_q^\top}{\sqrt{d}} \right). \tag{75}$$

The scale-aware representation after cross-scale interaction is:

$$\widetilde{\mathcal{H}}_t^q = A_q V_q. \tag{76}$$

A residual formulation can further be written as:

$$\mathcal{H}_t^q \leftarrow \mathcal{H}_t^q + \widetilde{\mathcal{H}}_t^q. \tag{77}$$

which preserves the original scale-specific information while introducing dependencies across motion granularities.

B.10 Entropy-Regularized Scale Allocation

Let w_q denote the normalized weight assigned to the q-th scale:

$$w_q = \frac{\exp(a_q)}{\sum_{r=1}^{Q} \exp(a_r)}, \qquad \sum_{q=1}^{Q} w_q = 1. \tag{78}$$

The entropy of the scale distribution is:

$$\mathcal{H}(w) = -\sum_{q=1}^{Q} w_q \log w_q. \tag{79}$$

The entropy-regularized objective used in the method can therefore be written as:

$$\mathcal{L}_{concat} = -\sum_{q=1}^{Q} w_q \log w_q. \tag{80}$$

Because the entropy reaches its maximum when the scale weights are distributed more evenly, optimizing this term discourages the representation from collapsing onto a single motion scale.

For the normalized weights, the entropy gradient with respect to a weight is:

$$\frac{\partial \mathcal{H}(w)}{\partial w_q} = -(\log w_q + 1). \tag{81}$$

This formulation provides a direct regularization signal for maintaining the contribution of multiple motion granularities.

B.11 Progressive Geometry–Motion Fusion

For each scale, the fusion network processes the corresponding structure-aware representation:

$$\varphi_q = \mathcal{F}_{fusion}\left(\mathcal{H}_t^q\right), \qquad q \in \{c, mc, mf, f\}. \tag{82}$$

A progressive fusion operation can be represented recursively as:

$$Z^{(1)} = \mathcal{F}_{fusion}^{(1)}\left(\mathcal{H}_t^{(c)}\right). \tag{83}$$

$$Z^{(q)} = \mathcal{F}_{fusion}^{(q)}\left(\mathcal{H}_t^q, Z^{(q-1)}\right), \qquad q > 1. \tag{84}$$

The resulting scale-specific outputs are denoted as φ_q. Their concatenation produces the global fused motion representation:

$$\mathcal{K}_{motion} = Concat\left(\varphi_c, \varphi_{mc}, \varphi_{mf}, \varphi_f\right). \tag{85}$$

This formulation allows information from coarse motion patterns to be progressively propagated toward finer motion structures while retaining scale-specific characteristics.

B.12 Spatial Motion Consistency

To encourage coordinated motion among spatially neighboring regions, an energy representation $E^q(i)$ is defined for spatial position i at scale q. For a two-dimensional spatial feature field, the spatial gradient is:

$$\nabla E^q(i) = \left[\frac{\partial E^q(i)}{\partial x}, \frac{\partial E^q(i)}{\partial y}\right]^\top. \tag{86}$$

For two neighboring positions i and j, the gradient discrepancy is:

$$\Delta_{ij}^q = \nabla E^q(i) - \nabla E^q(j). \tag{87}$$

The corresponding spatial consistency objective is:

$$\mathcal{L}_{fusion} = \sum_i \sum_{j \in \mathcal{N}(i)} S_{ij} \left\|\Delta_{ij}^q\right\|_2^2. \tag{88}$$

The structural similarity weight can be normalized as:

$$S_{ij} = \frac{\exp(s_{ij})}{\sum_{l \in \mathcal{N}(i)} \exp(s_{il})}. \tag{89}$$

where s_{ij} denotes the similarity between the structural features at positions i and j.

Expanding the squared difference gives:

$$\left\|\Delta_{ij}^q\right\|_2^2 = \left\|\nabla E^q(i)\right\|_2^2 + \left\|\nabla E^q(j)\right\|_2^2 - 2\left(\nabla E^q(i)\right)^\top \nabla E^q(j). \tag{90}$$

Thus, minimizing $\mathcal{L}_{fusion}$ encourages spatially similar regions to exhibit consistent motion gradients, thereby reducing abrupt motion differences across neighboring portrait structures.

B.13 Joint Condition Construction

The two proposed modules produce complementary global representations:

$$\mathcal{K}_{appearance} = \mathcal{T}\left(\{\sigma_n\}_{n=1}^N\right). \tag{91}$$

$$\mathcal{K}_{motion} = Concat\left(\{\varphi_q\}\right). \tag{92}$$

These representations are jointly transformed into the diffusion condition:

$$c = \mathcal{F}_{condition}\left(\mathcal{K}_{appearance}, \mathcal{K}_{motion}\right). \tag{93}$$

Let the concatenated representation be:

$$Z_c = \left[\mathcal{K}_{appearance}; \mathcal{K}_{motion}\right]. \tag{94}$$

A generic weighting function can then be expressed as:

$$c = W_c Z_c + b_c. \tag{95}$$

or, when adaptive weighting is employed:

$$c = \omega_A \odot W_A \mathcal{K}_{appearance} + \omega_M \odot W_M \mathcal{K}_{motion} + b_c. \tag{96}$$

The resulting condition is injected into the denoising network together with the noisy sample and diffusion timestep:

$$\epsilon_\theta = \epsilon_\theta \left(x_t, t, c \right). \tag{97}$$

This establishes the complete computational path from portrait appearance and multi-scale motion information to the conditional denoising process.

B.14 Overall Optimization Objective

The AG2I module contains the temporal portrait constraint, appearance distribution alignment, decoupling, and individual-guided alignment objectives. Their weighted combination can be written as:

$$\mathcal{L}_{AG2I} = \lambda_{portrait}\mathcal{L}_{portrait} + \lambda_{appearance}\mathcal{L}_{appearance} + \lambda_{decoupling}\mathcal{L}_{decoupling} + \lambda_{alignment}\mathcal{L}_{alignment}. \tag{98}$$

Similarly, the FG&M module combines cross-scale consistency, motion decorrelation, scale-aware reorganization, entropy regularization, and spatial fusion constraints:

$$\mathcal{L}_{FG\&M} = \lambda_{scale}\mathcal{L}_{scale} + \lambda_{motion}\mathcal{L}_{motion} + \lambda_{concat}\mathcal{L}_{concat} + \lambda_{fusion}\mathcal{L}_{fusion}. \tag{99}$$

The complete auxiliary objective introduced by the two strategies is:

$$\mathcal{L}_{condition} = \mathcal{L}_{AG2I} + \mathcal{L}_{FG\&M}. \tag{100}$$

Combining the conditional representation learning objectives with the diffusion reconstruction objective gives the overall training objective:

$$\mathcal{L}_{total} = \mathcal{L}(c) + \lambda_{condition}\mathcal{L}_{condition}. \tag{101}$$

Expanding the conditional denoising component yields:

$$\mathcal{L}_{total} = \mathbb{E}_{x_0,\epsilon,t}\left[\left\| \epsilon - \epsilon_\theta \left(x_t, t, \mathcal{F}_{condition}(\mathcal{K}_{appearance}, \mathcal{K}_{motion}) \right) \right\|_2^2 \right] + \lambda_{condition} \left(\mathcal{L}_{AG2I} + \mathcal{L}_{FG\&M} \right). \tag{102}$$

Through joint optimization, the AG2I objective emphasizes the preservation and alignment of individual-specific appearance characteristics, while the FG&M objective constrains motion representations across multiple granularities and spatially related regions. The resulting condition representation therefore provides the diffusion model with both personalized appearance information and coordinated motion information.

C Lightweight Teacher–Student Model

The two proposed strategies consist of multiple submodules. To further improve efficiency, we adopt a knowledge distillation strategy to distill the key submodules into a unified lightweight student model, as shown in Figure 7.

Specifically, our approach consists of eight core submodules: portrait encoding component $\mathcal{F}_{portrait}$, appearance domain construction component $\mathcal{F}_{appearance}$, feature disentanglement component $\mathcal{F}_{decoupling}$, alignment component $\mathcal{F}_{alignment}$, multi-scale joint component $\mathcal{F}_{scale}$, motion domain construction component $\mathcal{F}_{motion}$, feature reconstruction component $\mathcal{F}_{concat}$, and fusion component $\mathcal{F}_{fusion}$. These submodules are responsible for prediction and updating for feature computation, playing a crucial role in generating the final output.

During training, we adopt a feature-level knowledge distillation scheme, where each student submodule $\mathcal{F}_z^{student}$ learns to approximate its teacher counterpart by minimizing feature discrepancies.

Here, $z \in 1, \ldots, 8$ denotes the index of the eight student submodules, including $\mathcal{F}_{portrait}^{student}$, $\mathcal{F}_{appearance}^{student}$, $\mathcal{F}_{decoupling}^{student}$, $\mathcal{F}_{alignment}^{student}$, $\mathcal{F}_{scale}^{student}$, $\mathcal{F}_{motion}^{student}$, $\mathcal{F}_{concat}^{student}$, and $\mathcal{F}_{fusion}^{student}$. This design enables a lightweight and efficient student model while preserving the performance of the teacher model.

For each submodule, we define the following distillation loss:

$$\mathcal{L}_z = \left\| \mathbf{F}_z^{student} - \mathbf{F}_z \right\|_2^2 . \tag{103}$$

where $\mathbf{F}_z$ denotes the output feature of the corresponding teacher submodule, including $\mathcal{F}_{portrait}$, $\mathcal{F}_{appearance}$, $\mathcal{F}_{decoupling}$, $\mathcal{F}_{alignment}$, $\mathcal{F}_{scale}$, $\mathcal{F}_{motion}$, $\mathcal{F}_{concat}$, and $\mathcal{F}_{fusion}$.

The total distillation loss is then formulated as a weighted sum:

$$\mathcal{L}_{total} = \sum_{z=1}^{8} \lambda_z \mathcal{L}_z = \sum_{z=1}^{8} \lambda_z \left\| \mathbf{F}_z^{student} - \mathbf{F}_z \right\|_2^2 . \tag{104}$$

where λ_z represents an adjustable hyperparameter used to balance the relative importance of different submodules.

Furthermore, we conducted quantitative comparisons. As shown in Table 3, we compare the proposed lightweight model with the full model and other methods. Experimental results show that the designed lightweight model can effectively improve generation efficiency while maintaining visual quality.

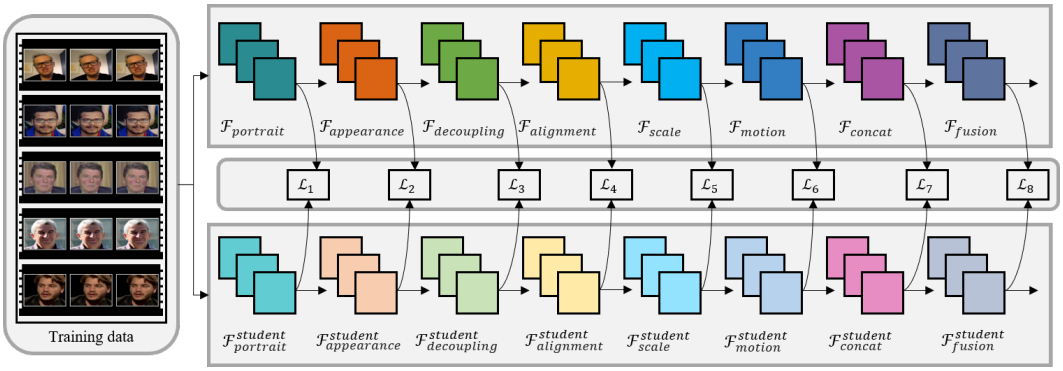

Figure 7: A lightweight model is proposed based on this paper.

D More Details on Experiments

D.1 User Study

We designed a user study to evaluate the subjective performance of the generated videos. One hundred participants were invited to view multiple sets of talking videos, including outputs from our method and several baselines. Participants rated the videos across four dimensions, including image quality, video fluency, identity preservation, and personalization, using a scale of 1 to 5, where higher scores indicate better performance.

Table 3: Quantitative results.

Methods	PSNR ↑	LPIPS ↓	SyncNet ↑	LSE-C ↑	LMD ↓	CSIM ↑	FPS ↑
Ground Truth	N/A	N/A	8.743	7.937	N/A	1	N/A
Ours	**37.65**	**0.01**	**8.465**	**7.615**	**1.04**	**0.96**	**10.16**
Ours (lightweight)	**37.63**	**0.01**	**8.464**	**7.613**	**1.05**	**0.96**	**25.14**
StreamAvatar	32.57	0.07	7.944	7.172	1.74	0.91	5.154
EchoMimicV3	33.71	0.04	8.115	7.049	1.75	0.89	0.297
Hallo3	34.27	0.05	7.031	7.265	1.63	0.89	0.065
AniPortrait	33.18	0.09	7.149	7.164	1.34	0.84	0.758
MoDiTalker	35.29	0.03	7.442	7.067	1.35	0.92	2.047
V-Express	31.45	0.06	7.763	7.265	1.63	0.85	0.746

Image Quality: This criterion evaluates the overall visual quality of individual generated frames, including image clarity, facial structure, and local texture details. It also considers the presence of visual defects such as blurring, artifacts, distortions, or unnatural boundaries, as well as whether the overall illumination and appearance remain natural and consistent.

Video Fluency: This criterion evaluates the temporal continuity and stability of the generated videos. It focuses on whether transitions between adjacent frames are smooth and whether the overall motion remains continuous and natural. Participants also consider the presence of flickering, jitter, sudden changes, or other temporal inconsistencies, particularly during long-sequence generation.

Identity Protection: This criterion evaluates how well the generated results preserve the identity characteristics of the target subject throughout the video. It considers the consistency of facial structure, key facial features, and overall appearance across different frames, while also examining whether identity drift, facial distortion, or identity confusion occurs during dynamic generation.

Personalized Display: This criterion evaluates the ability of the generated results to preserve and express individual-specific characteristics. It focuses on personalized facial expression styles, distinctive facial features, fine-grained local details, and other unique characteristics of the target subject. It also considers whether these personalized traits remain recognizable during dynamic changes rather than being averaged or weakened across different identities.

The experimental results are presented in Figure 8. Compared to baseline methods, our approach achieves superior performance across all evaluation dimensions. For image quality, our method achieves the highest scores, demonstrating clear advantages in clarity and detail retention while effectively reducing visual defects such as blur and artifacts. For video fluency, our method excels in temporal continuity and motion naturalness, producing smooth transitions between adjacent frames and significantly reducing jitter and discontinuities. For identity preservation, our method better maintains original identity characteristics, ensuring high consistency in facial structure and key details and effectively avoiding identity drift or confusion. Finally, for personalization, our method better captures individual differences, exhibiting greater distinctiveness in expression styles and fine-grained features, and avoiding feature dilution often observed in averaged results.

Overall, our method achieves consistent improvements in both personalization and realism for talking video generation across all evaluation dimensions.

D.2 Model Performance

We analyze the convergence speed and training stability of the model by examining loss curves during training. Figure 9 illustrates the convergence trends of the proposed method and several baselines under identical training settings. To ensure a fair comparison, all methods are trained on a unified hardware environment with eight NVIDIA A100 GPUs, using a batch size of 1, a learning rate of 10^{-5}, 20,000 iterations, and an input resolution of 720×720.

Overall, existing diffusion-based methods such as Hallo3 and EchoMimicV3 exhibit relatively stable but slow convergence behavior, with loss values tending to plateau in the later stages of training, indicating limited optimization efficiency. In contrast, methods including MoDiTalker, AniPortrait, and V-Express achieve faster initial convergence by incorporating intermediate representations or structural constraints, but often suffer from oscillations or stagnation in later optimization stages,

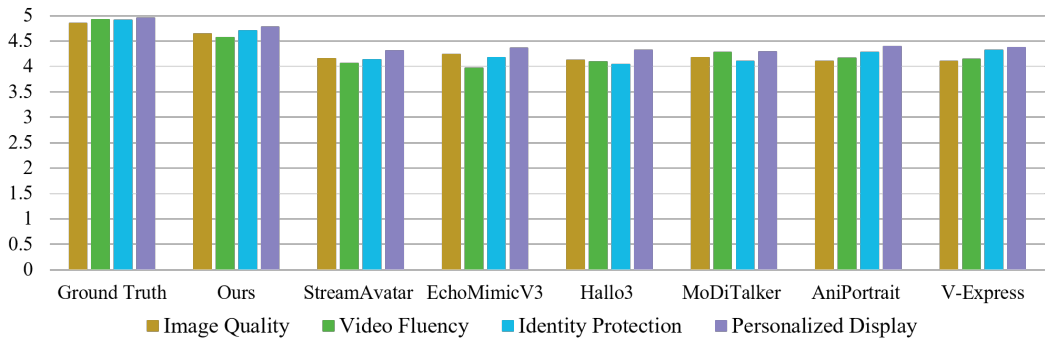

Figure 8: User study.

reflecting instability in long-term training dynamics. StreamAvatar improves convergence speed through a distillation strategy, yet still exhibits limited stability during later training stages.

In contrast, our method demonstrates consistently faster and more stable convergence throughout training. Even the lightweight variant, which also adopts distillation, achieves rapid loss reduction in early stages and maintains a steady downward trend without noticeable oscillation in later iterations. These results indicate that our global representation modeling improves optimization stability and convergence efficiency for diffusion-based training, and remains effective under both full-scale and lightweight configurations.

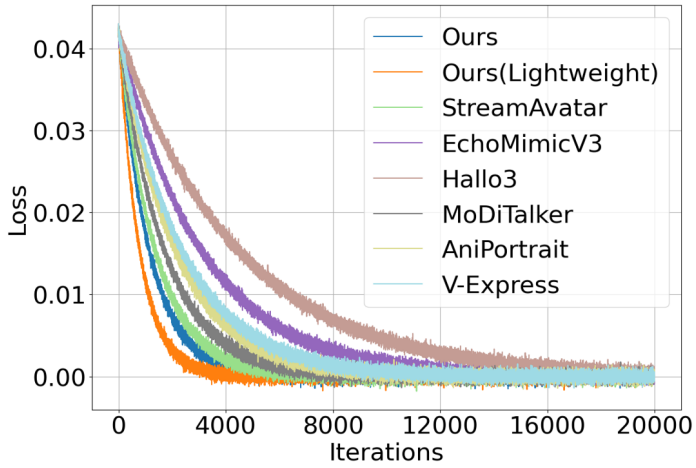

Figure 9: Model performance.

D.3 Parameter Settings

Our method can reconstruct personalized portrait characteristics to a certain extent, but it still relies on the input portrait. For example, visible attributes (e.g., teeth) can be enhanced when present but are difficult to recover when missing.

To address this issue, we introduce a portrait attribute estimation strategy, which employs a prediction network $\mathcal{F}_{prediction}$ to infer latent personalized features from high-level semantic information of the input portrait (e.g., age and skin condition), and generates a control parameter ρ as follows:

$$\rho = \mathcal{F}_{prediction}(Image) \tag{105}$$

where $\rho \in [0, 1]$ controls the balance between data-driven features and prior information, improving the plausibility of fine details. $\mathcal{F}_{prediction}$ is implemented as a lightweight multi-layer MLP-based regression network, which maps semantic attributes into a compact representation for estimating ρ, ensuring stable and efficient prediction.

As shown in Figure 10, ρ controls the strength of prior incorporation: larger values increase reliance on prior information. The best performance is achieved when $\rho = 0.5$. Smaller values lead to blurry teeth details, while larger values may introduce unnatural distortions.

Overall, by properly setting ρ, the proposed strategy improves generation quality even when teeth are not explicitly present in the input portrait, demonstrating its effectiveness in compensating for missing fine-grained details.

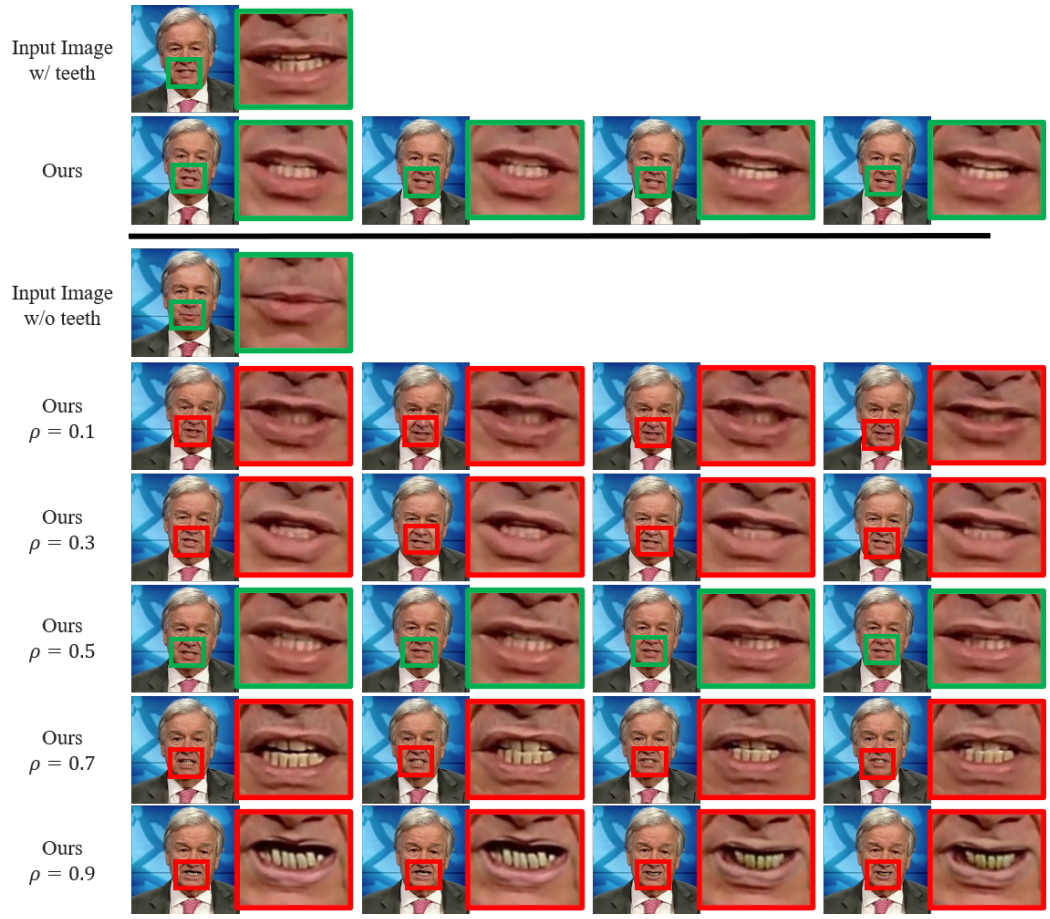

Figure 10: Parameter settings for ρ. Our prediction network can infer plausible details even when personalized details are missing. When $\rho = 0.5$, the generated results achieve the highest consistency with the input portrait.

D.4 Limits

Our method achieves personalized talking portrait video generation through global representation learning. However, it still has certain limitations. Specifically, the model relies on saliency-driven alignment between individual-specific and population-level features, as well as the integration of geometric information and motion patterns. This makes the generation process highly dependent on temporal modeling capability. Although our method can generate videos of approximately 45 seconds, performance degradation may occur in longer sequences, such as gradually blurred textures and audio-visual misalignment, as shown in Figure 11.

From the perspective of spatial modeling, these issues partly stem from the model's limited ability to capture facial spatial structures. Existing methods mainly rely on implicit temporal modeling to maintain consistency, lacking explicit constraints on geometric structures and spatial relationships. To address this, we consider incorporating spatial structural priors to improve stability and realism in long-duration video generation while preserving temporal coherence.



Figure 11: Some limitations of this method. When the generation duration exceeds 45 seconds, instability in spatial structure may occur.