WHAT ENABLES MOTION COORDINATION IN TALKING PORTRAIT GENERATION? FROM CROSS-DOMAIN MOTION REPRESENTATION TO ALIGNMENT AND FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

3DGS-based talking portrait generation has achieved significant progress. However, motion coordination still presents challenges, particularly in the heterogeneous modality differences between audio and portraits, as well as in interactions among motion regions. This primarily stems from the fact that 3DGS is fundamentally based on multi-view appearance fitting, where portrait motion is coupled with appearance only implicitly or indirectly, leading to weakened and limited motion representations. To address these issues, we propose a cross-domain motion representation framework for talking portrait generation, aiming to move beyond the traditional paradigm that relies on appearance variations. Specifically, we construct motion representations in the audio-visual domain and the density domain by leveraging motion variation characteristics across modalities and portrait regions, thereby enhancing motion coordination. The framework consists of two key strategies: from audio-visual domain representation to alignment (AR-A) and from density domain representation to fusion (DR-F). For AR-A, we construct the audio-visual representation based on the correlation between audio and different portrait structures, and leverage strongly correlated structures to guide the alignment of weakly correlated ones, thereby enhancing cross-modal synchronization. For DR-F, we construct the density representation based on motion variations in continuous sequences, and map the motion sequences aligned in AR-A into density space for fusion, thereby enabling cross-region motion interaction. Experimental results demonstrate that our method achieves superior motion coordination and outperforms existing state-of-the-art approaches.

1 INTRODUCTION

Talking portrait generation refers to synthesizing portrait animations based on audio signals, providing key technical support for applications such as human-computer interaction, virtual avatars, and digital human reconstruction Chen et al. (2026); Li et al. (2026); Wang et al. (2026b); Dong & Zhang (2025); Liu et al. (2026b). The primary goal of this task is to achieve audiovisual synchronization and natural motion generation. However, existing methods still face challenges in motion coordination, often leading to inconsistent or unstable motion across different components.

In recent years, 3D Gaussian Splatting has developed rapidly in human-centric video generation Fei et al. (2024); Bao et al. (2025); Hu et al. (2024). These methods aim to integrate multimodal inputs, such as text, images, and audio, to synthesize high-fidelity human videos Hu et al. (2024); Jiang et al. (2024). Existing works have made notable progress in human modeling Kocabas et al. (2024); Liu et al. (2025); Li et al. (2024b), rendering, and animation generation. These methods focus on capturing global human shape, pose variations, and motion trajectories, so as to maintain temporal consistency in generated videos. Within this area, talking portrait generation has recently received increasing attention Zhou et al. (2020); Ma et al. (2023). Compared with full-body modeling, this task places more emphasis on audio-lip synchronization and the coherence of both local and global motions, such as lip movements, eye motions, and facial expressions Wang et al. (2026a); Chen et al. (2026); Peng et al. (2026). However, most existing methods generate animations mainly by modeling appearance changes or performing temporal interpolation over long sequences, which limits motion expressiveness and leads to poor motion coordination.

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.

We observe that in talking portrait generation, the correlations between different portrait structures and audio vary Wang et al. (2022); Jiang et al. (2025). For example, the mouth region typically exhibits stronger correlation, while weakly correlated regions tend to be underestimated or accumulate errors during alignment with strongly correlated regions. In addition, in continuous sequences, motion properties such as amplitude and frequency exhibit non-uniform distributions Chu et al. (2024); Liu et al. (2024a), which can lead to inconsistencies between local and global motion during fusion, resulting in structural distortions or loss of details. Based on these observations, we propose a cross-domain motion representation framework for talking portrait generation, consisting of two main strategies: from audio-visual domain representation to alignment (AR-A) and from density domain representation to fusion (DR-F). For AR-A, we construct the audio-visual representation based on the correlations between audio and different portrait structures, and use strongly correlated structures to align weakly correlated ones, establishing correspondences across heterogeneous modalities. For DR-F, we construct the density representation based on motion variations within a continuous sequence and map the aligned sequences from AR-A into this density space for fusion, maintaining the coherence and integrity of different motion regions.

This paper proposes a talking portrait generation method based on cross-domain motion representation. Specifically, we propose the AR-A (from audio-visual domain representation to alignment) strategy to enhance cross-modal synchronization, and the DR-F (from density domain representation to fusion) strategy to improve cross-region motion interaction. Experimental results demonstrate that the proposed method achieves significant improvements in motion coordination.

2 OVERVIEW

The overall framework of the proposed method is illustrated in Figure 1, which consists of two core strategies: from audio-visual domain representation to alignment (AR-A) and from density domain representation to fusion (DR-F).

In AR-A, we first project the audio and portrait structures into an audio-visual domain. Due to differences in correlation between the two (e.g., the mouth exhibits the strongest correlation), we use $MLP_{audio-visual}$ to characterize the strength of the correlations. Subsequently, based on the corresponding feature sequence $\mathcal{Z}_{audio-visual}$, we utilize highly correlated structures to guide the alignment of weakly correlated ones, thereby enhancing cross-modal synchronization.

In DR-F, we construct a density domain based on motion variations (e.g., amplitude and frequency) across consecutive sequences. We then compute its hierarchical distribution, denoted as $\mathcal{Z}_{density}$, to characterize the motion properties of different regions. Subsequently, we map the aligned feature sequences from AR-A into this density space and perform density-guided fusion according to the hierarchical density distribution, thereby facilitating cross-regional motion interactions.

Finally, based on 3DGS, we integrate the audio-visual alignment representation $\mathcal{R}_{audio-visual}$ and the density fusion representation $\mathcal{R}_{density}$ obtained from AR-A and DR-F, in order to enhance motion coordination in talking videos.

2.1 3DGS AND UPDATES

In 3DGS-based human portrait reconstruction, each Gaussian primitive is modeled as a 3D Gaussian function, as follows:

$$G(x) = \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right). \tag{1}$$

where x denotes a spatial point, μ is the Gaussian center, and Σ is the covariance matrix. Furthermore, each Gaussian primitive is parameterized by the following attributes:

$$G = \{c, s, r, C, \alpha\}. \tag{2}$$

where geometric attributes include the center position c, scaling vector s, and rotation parameter r, and appearance attributes include surface color C and opacity α.

To enhance motion coordination, we incrementally update the Gaussian parameters to obtain the updated parameters G', as shown below:

$$G' = G \oplus \Delta G. \tag{3}$$

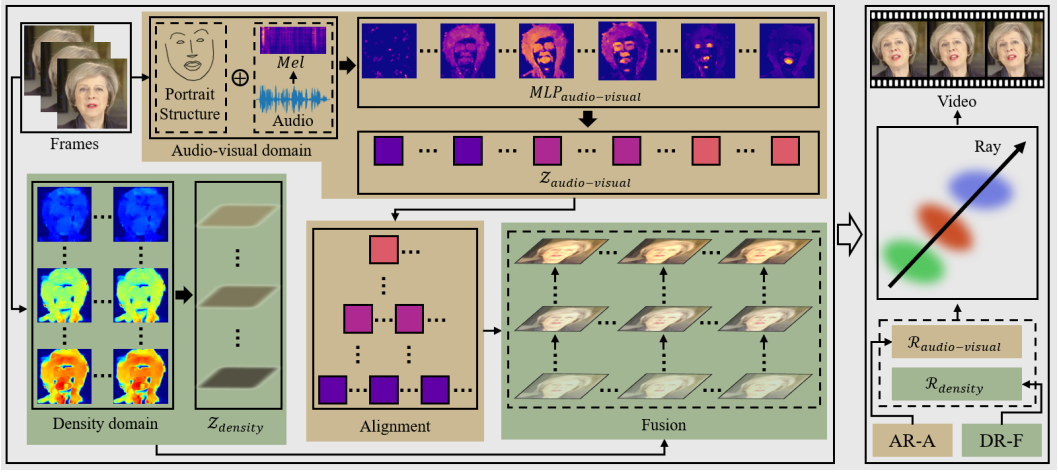


Figure 1: The framework of our method. The method consists of two modules: from audio-visual domain representation to alignment (AR-A) and from density domain representation to fusion (DR-F). AR-A leverages strongly correlated structures to guide the alignment of weakly correlated structures. DR-F characterizes motion attributes through the density domain to achieve fusion.

where $\oplus$ denotes parameter-wise additive or multiplicative composition. For the increment ΔG, we define a conditional mapping function as follows:

$$\Delta G = \mathcal{F}_\theta(G, \mathcal{R}_{audio-visual}, \mathcal{R}_{density}). \tag{4}$$

where $\mathcal{F}_\theta$ is implemented using an MLP and performs differentiated processing across parameter.

3 Method

3.1 From Audio-visual Domain Representation to Alignment

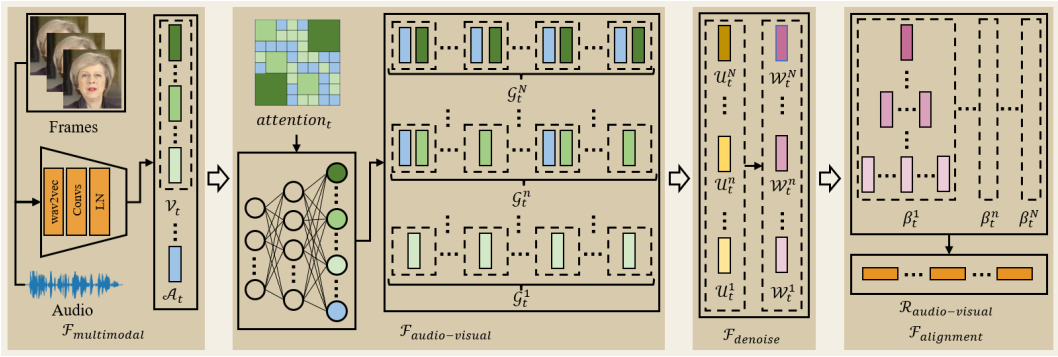


Figure 2: The process of from audio-visual domain representation to alignment (AR-A).

The AR-A process is shown in Figure 2. First, given a video frame sequence $\{f_t\}_{t=1}^T$ and the corresponding audio signal a, we use a multimodal feature extractor $\mathcal{F}_{multimodal}$ to extract visual features $\mathcal{V}_t$ and audio features $\mathcal{A}_t$:

$$\mathcal{F}_{multimodal} : (f_t, a) \rightarrow (\mathcal{V}_t, \mathcal{A}_t). \tag{5}$$

where $\mathcal{F}_{multimodal}$ consists of a multi-layer convolutional network and a *wav2vec*-based audio encoder Baevski et al. (2020); Schneider et al. (2019).

To ensure cross-modal alignment between visual and audio features, we introduce a contrastive loss:

$$\mathcal{L}_{multimodal} = -\sum_{i} \log \frac{\exp\left(\cos(\mathcal{V}_t^i, \mathcal{A}_t^i)\right)}{\sum_j \exp\left(\cos(\mathcal{V}_t^i, \mathcal{A}_t^j)\right)}. \tag{6}$$

where i and j denote frame indices.

Next, we construct an audio-visual representation of the audio and visual features, denoted as $\{\mathcal{G}_t^n\}_{n=1}^N$. Considering that different portrait structures exhibit varying degrees of coupling with audio, we compute feature weights $attention_t$ from $\mathcal{V}_t$ and $\mathcal{A}_t$ based on a dynamic attention mechanism, and further map them to $\{\mathcal{G}_t^n\}_{n=1}^N$ as follows:

$$\mathcal{F}_{audio\text{-}visual} : (\mathcal{V}_t, \mathcal{A}_t) \to \mathcal{G}_t^n, \quad 1 \le n \le N. \tag{7}$$

where $\mathcal{F}_{audio\text{-}visual}$ is built upon a dynamic attention mechanism and implemented using a Transformer architecture to capture the cross-modal interactions between audio and visual features.

We introduce a correlation loss to reflect the sensitivity of different portrait structures to audio:

$$\mathcal{L}_{audio\text{-}visual} = \sum_{n=1}^{N} \left\| \mathrm{corr}(\mathcal{G}_t^n, \mathcal{A}_t) - \gamma_n^t \right\|_2^2. \tag{8}$$

where $\gamma_n^t \in [0, 1]$ denotes the coupling strength, with parameter settings provided in Appendix C.3.

Subsequently, considering that motion data in consecutive frames often contains noise, we introduce a denoising network $\mathcal{F}_{denoise}$ to process the features and obtain $\{\mathcal{W}_t^n\}_{n=1}^N$:

$$\mathcal{F}_{denoise} : \mathcal{U}_t^n \to \mathcal{W}_t^n. \tag{9}$$

where $\mathcal{F}_{denoise}$ consists of convolutional and residual layers to suppress noise while preserving temporal continuity. For ease of visualization in the figures, we use $\{\mathcal{U}_t^n\}_{n=1}^N$ to denote $\{\mathcal{G}_t^n\}_{n=1}^N$.

To remove noise while retaining dynamic information, we design the following loss function:

$$\mathcal{L}_{denoise} = \sum_{t=1}^{T-1} \sum_{n=1}^{N} \left\| \mathcal{W}_{t+1}^n - \mathcal{W}_t^n \right\|_2^2 + \sum_{t=1}^{T} \sum_{n=1}^{N} \left\| \mathcal{W}_t^n - \mathcal{U}_t^n \right\|_2^2. \tag{10}$$

where the first term enforces temporal smoothness to suppress noise, while the second term preserves the original motion information and prevents over-smoothing.

Finally, given the denoised hierarchical representations $\{\mathcal{W}_t^n\}_{n=1}^N$, we introduce an alignment network $\mathcal{F}_{alignment}$ to generate the cross-modal features $\{\beta_t^n\}_{n=1}^N$, formulated as:

$$\mathcal{F}_{alignment} : \mathcal{W}_t^n \to \beta_t^n. \tag{11}$$

where β_t^n is used to obtain the audio-visual alignment representation $\mathcal{R}_{audio-visual}$ via the aggregation function $\mathcal{T}(\cdot)$, as shown below:

$$\mathcal{R}_{audio-visual} = \mathcal{T}\left(\{\beta_t^n\}_{n=1}^N\right). \tag{12}$$

where $\mathcal{T}(\cdot)$ consists of a linear layer and a normalization operation.

To enforce consistency between visual and audio features across different levels, we design a weighted alignment loss:

$$\mathcal{L}_{alignment} = \sum_{t=1}^{T} \sum_{n=1}^{N} \left| \hat{\mathcal{V}}_t^n - \hat{\mathcal{A}}_t^n \right|_2^2. \tag{13}$$

where $\hat{\mathcal{V}}_t^n$ and $\hat{\mathcal{A}}_t^n$ denote the normalized visual features and audio features, respectively.

3.2 From Density Domain Representation to Fusion

The DR-F process is illustrated in Figure 3. First, given the training data $\{f_t\}_{t=1}^T$, we introduce a window function $\mathcal{F}_{window}$ with a temporal window radius k to extract motion features $\mathcal{H}_t$ from consecutive frames, as follows:

$$\mathcal{F}_{window} : (f_{t-k:t+k}) \to \mathcal{H}_t. \tag{14}$$

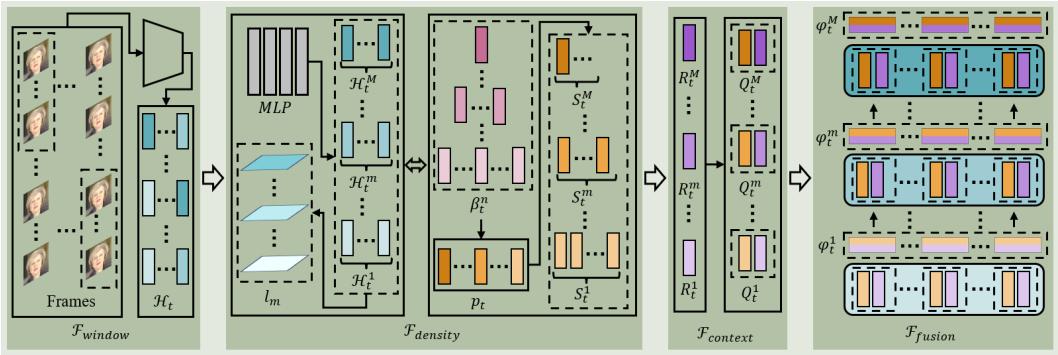


Figure 3: The process of from density domain representation to fusion (DR-F).

where $\mathcal{F}_{window}$ is implemented using a multi-layer convolutional network designed to capture motion characteristics such as amplitude and frequency.

We introduce a self-supervised loss function to capture the amplitude and frequency characteristics of local motion, expressed as follows:

$$\mathcal{L}_{window} = -\operatorname{Var}\left(\sum_{t-k}^{t+k} |\mathcal{H}_t|\right). \tag{15}$$

where $\operatorname{Var}(\cdot)$ denotes the variance over the temporal window centered at t with size $2k+1$.

Based on the motion features $\mathcal{H}_t$, we design a mapping function $\mathcal{F}_{density}$ to decompose $\mathcal{H}_t$ into M hierarchical features $\{\mathcal{H}_t^m\}_{m=1}^M$ and further map them to density representations l_m, as follows:

$$\mathcal{F}_{density}\colon \mathcal{H}_t \xrightarrow{MLP} \{\mathcal{H}_t^m\} \xrightarrow{mapping} l_m. \tag{16}$$

where $\mathcal{F}_{density}$ is implemented using a lightweight multilayer perceptron and convolutional layers.

We denote the alignment sequence $\{\beta_t^n\}_{n=1}^N$ as p_t and express it based on l_m as S_t^m, as follows:

$$S_t^m = f(l_m; Q_t). \tag{17}$$

where $f(\cdot)$ represents a hierarchical conditional mapping that projects the alignment sequence into the corresponding density space.

To accurately capture the motion characteristics at each level, we introduce a hierarchical loss:

$$\mathcal{L}_{density} = \sum_{t=1}^{T}\sum_{m=1}^{M-1}\left\|\frac{S_t^{m+1}-S_t^m}{l_{m+1}-l_m+\delta_{(t,m)}} - \mathbf{1}\right\|_2^2.$$

where the weight $\delta_{(t,m)} \in [0,1]$ is used to balance the magnitude of variation when mapping aligned feature sequences to density representations, with parameter settings provided in Appendix C.3.

Subsequently, we capture contextual information R_t^m from the continuous motion features $\mathcal{H}_t$ using a set of convolutional layers. We then combine R_t^m with S_t^m to generate a context-enhanced representation Q_t, as follows:

$$\mathcal{F}_{context} : (S_t^m, R_t^m) \to Q_t^m. \tag{18}$$

where $\mathcal{F}_{context}$ consists of a 2-3 layer MLP to mitigate conflicts between hierarchical features.

Here, to enforce consistency between the enhanced representations Q_t^m and the original hierarchical features S_t^m, we introduce an adaptive context consistency loss:

$$\mathcal{L}_{context} = \sum_{t=1}^{T}\sum_{m=1}^{M}\|Q_t^m - S_t^m\|_2^2. \tag{19}$$

Finally, in the fusion stage, we apply the fusion mechanism $\mathcal{F}_{fusion}$ to Q_t^m, generating the cross-region features φ_t^m, as shown below:

$$\mathcal{F}_{fusion} : (Q_t^m) \to \varphi_t^m, \quad m = 1, \ldots, M. \tag{20}$$

where $\mathcal{F}_{fusion}$ consists of convolutional layers, and the set $\{\varphi_t^m\}_{m=1}^M$ is aggregated via a fully connected layer to obtain the density fusion representation $\mathcal{R}_{density}$.

We introduce the loss function $\mathcal{L}_{fusion}$ to ensure consistency of features at different levels during the fusion process, as follows:

$$\mathcal{L}_{fusion} = \sum_{t=1}^{T} \sum_{m=1}^{M-1} \left\| \left(\varphi_t^{m+1} - \varphi_t^m\right) - \left(Q_t^{m+1} - Q_t^m\right) \right\|_2^2 . \tag{21}$$

4 EXPERIMENT

To comprehensively validate the effectiveness of the proposed method, we validate the proposed method through qualitative and quantitative comparisons, ablation study, and related discussions. Additional details, including user study (Section C.1), model performance (Section C.2), parameter settings (Section C.3), and limitations (Section C.4), are provided in the appendix.

Implementation details. Our training data are sourced from publicly accessible lectures and interview videos on YouTube. We preprocess the videos by cropping, extracting frames at 25 FPS, and resizing them to 512×512. For a fair comparison, all compared methods are evaluated under the same experimental settings. Experiments are carried out on an NVIDIA RTX 5090 GPU.

Comparison methods. We compared our method with other 3DGS-based approaches: EmoDiffTalk Liu et al. (2026a), EmoTaG Xu et al. (2026), InsTaG Li et al. (2025), PointTalk Xie et al. (2025), GaussianTalker Cho et al. (2024), TalkingGaussian Li et al. (2024a).

4.1 QUALITATIVE COMPARISON

We compare our method with the aforementioned approaches. As shown in Figure 4, (a) synthesizing novel pose sequences beyond those observed in the original dataset, and (b) reconstructing fine-grained wrinkle variations across multiple facial regions. We also recommend viewing the supplementary video (*supplementary.mp4*), which provides a dynamic demonstration.

EmoTaG models motion in the parameter space and incorporates emotion-aware mechanisms, improving emotional expressiveness and geometric stability. However, as a few-shot method, it suffers from limited generalization and a certain degradation in visual sharpness. EmoDiffTalk introduces emotion modeling into a Gaussian diffusion framework and combines it with text-driven AU control to enable fine-grained and controllable multimodal expression editing, but it remains limited in modeling extreme expressions. PointTalk employs a dynamic differential encoder to capture lip motion in detail, yet it lacks effective modeling of global structural motion, which tends to be oversmoothed. InsTaG, based on lightweight 3DGS, adapts to new identities with limited data, but compromises visual quality and temporal coherence. TalkingGaussian models facial and oral motion using deformable Gaussian primitives. However, its densification process may introduce noise, affecting structural consistency. GaussianTalker achieves high-quality lip synchronization via a shared implicit feature space and spatiotemporal attention, but under extreme poses, its structural modeling remains coarse due to limited training constraints.

In contrast, our method enhances synchronization between audio and visual modalities via AR-A module, thereby enabling more diverse pose variations without requiring additional motion priors. Furthermore, with the DR-F module, we can effectively capture interactions among different motion regions, thereby reconstructing richer personalized details, such as wrinkles caused by eyebrow raising. Overall, our method demonstrates superior performance in motion coordination, particularly regarding fine-grained details and motion coherence.

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 smoothness (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).

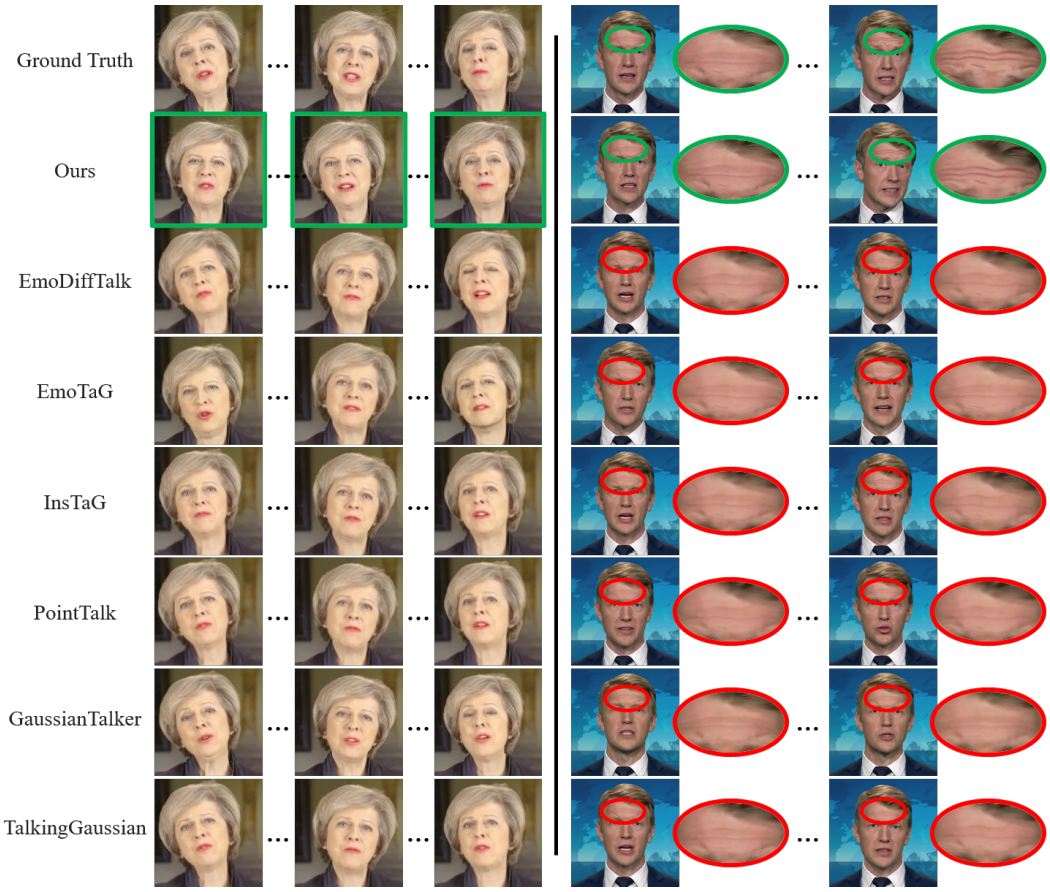


Figure 4: Qualitative results. Green boxes indicate that our method can generate diverse pose sequences while faithfully recovering portrait details.

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	**33.47**	**0.019**	**7.624**	**7.615**	**1.13**	**0.96**	**127.2**
EmoDiffTalk	25.48	0.028	5.832	7.172	3.14	0.91	86.4
EmoTaG	30.14	0.029	6.154	7.049	2.76	0.95	75.6
InsTaG	28.46	0.035	5.341	7.265	2.39	0.93	80.2
PointTalk	30.11	0.041	6.143	7.164	2.44	0.94	77.3
GaussianTalker	30.29	0.023	6.352	7.067	2.39	0.92	103.6
TalkingGaussian	28.37	0.036	6.148	7.108	2.65	0.95	92.4

Specifically, in terms of image quality, EmoTaG and InsTaG exhibit a certain degree of degradation due to their few-shot settings. TalkingGaussian introduces noise during the structural modeling process, which affects structural consistency to some extent. In terms of lip synchronization and motion smoothness, PointTalk achieves good synchronization performance. However, due to the lack of effective modeling of global structural motion, it tends to produce over-smoothed results. GaussianTalker performs well under normal head poses, but its performance degrades significantly under extreme poses. Meanwhile, EmoDiffTalk shows advantages in emotion-related tasks, but still suffers from limited stability when handling extreme expressions.

In contrast, the proposed method achieves more balanced performance across multiple metrics. On the one hand, it produces clearer and more detailed visual results in terms of image quality. On the other hand, in motion performance, it effectively captures both fine-grained and global motion

variations, leading to improved motion coherence and structural stability. In addition, our method maintains competitive inference efficiency. Overall, the quantitative results are consistent with the qualitative analysis, demonstrating that our method outperforms other approaches.

4.3 Ablation Study

To evaluate the contributions of AR-A and DR-F, we conduct ablation studies by removing key components. All experiments use the same training settings and dataset for fair comparison. Results are shown in Figure 5 and Table 2.

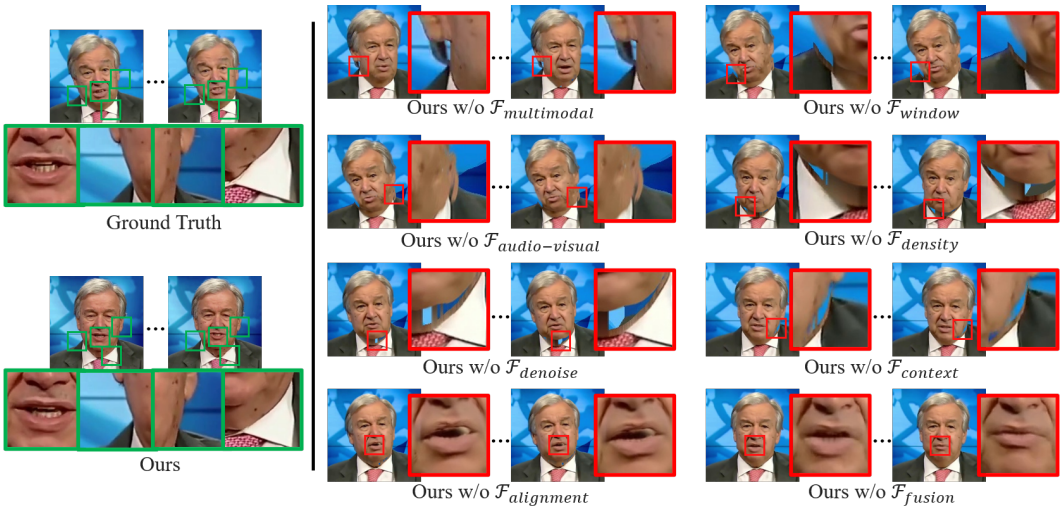


Figure 5: Qualitative results of the ablation study.

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	**33.64**	**0.019**	**7.625**	**7.562**	**1.14**	**0.96**
Ours w/o $\mathcal{F}_{multimodal}$	33.58	0.020	8.026	8.484	1.16	0.95
Ours w/o $\mathcal{F}_{audio\text{-}visual}$	33.61	0.021	8.085	8.526	1.16	0.95
Ours w/o $\mathcal{F}_{denoise}$	33.59	0.021	8.428	8.022	1.15	0.94
Ours w/o $\mathcal{F}_{alignment}$	33.57	0.020	8.507	8.094	1.16	0.94
Ours w/o $\mathcal{F}_{window}$	28.74	0.041	8.128	7.621	1.17	0.95
Ours w/o $\mathcal{F}_{density}$	29.89	0.042	8.083	7.623	1.15	0.94
Ours w/o $\mathcal{F}_{context}$	30.84	0.041	8.084	7.622	1.16	0.94
Ours w/o $\mathcal{F}_{fusion}$	29.28	0.040	8.131	7.623	1.15	0.95

Ours w/o $\mathcal{F}_{multimodal}$. $\mathcal{F}_{multimodal}$ extracts motion and audio features to build cross-modal representations. Without this module, the correspondence between audio and visual signals is significantly weakened, leading to degraded lip synchronization and motion consistency.

Ours w/o $\mathcal{F}_{audio\text{-}visual}$. $\mathcal{F}_{audio\text{-}visual}$ models audio–visual coupling via dynamic attention. Removing this module will reduce the ability to capture interactions, which in turn will affect motion performance and structural consistency.

Ours w/o $\mathcal{F}_{denoise}$. $\mathcal{F}_{denoise}$ suppresses noise in motion features while maintaining temporal continuity. Without it, the generated results become less stable over time, often showing frame jitter and structural flickering, which degrade motion smoothness and visual quality.

Ours w/o $\mathcal{F}_{alignment}$. $\mathcal{F}_{alignment}$ aligns the denoised hierarchical features with audio to ensure cross-modal consistency. Removing this module weakens the alignment between audio and visual signals, resulting in poorer lip synchronization and noticeable delays or offsets in motion responses.

Ours w/o $\mathcal{F}_{window}$. $\mathcal{F}_{window}$ extracts the amplitude and frequency of local windows from consecutive frames to form motion features. Without this module, the model struggles to capture local dynamics, leading to inaccurate motion rhythm, insufficient fine-grained motion, and overall overly smooth motion lacking dynamic variation.

Ours w/o $\mathcal{F}_{density}$. $\mathcal{F}_{density}$ maps motion features to density representations to describe motion structures at different granularities. Removing it limits the model's ability to represent multi-scale motion, resulting in weak distinction between fine-grained and global motion, and leading to monotonous motion with unclear structure.

Ours w/o $\mathcal{F}_{context}$. $\mathcal{F}_{context}$ enhances hierarchical features with contextual information to reduce conflicts across levels. Without this module, the coordination between features decreases, leading to inconsistent local motion, artifacts in detailed regions, and slight structural misalignment.

Ours w/o $\mathcal{F}_{fusion}$. $\mathcal{F}_{fusion}$ integrates motion features across levels and produces density-motion fusion parameters to enable cross-level coordination. Without it, features from different levels cannot be effectively combined, resulting in discontinuous motion responses and a disconnect between local and global motion, which reduces overall motion naturalness.

Overall, the experimental results demonstrate that all ablation variants exhibit noticeable performance degradation across different evaluation dimensions, validating the effectiveness of each module in improving the generation quality. In contrast, the full model (Ours) generates results that are closer to the real videos, with clear advantages in motion coordination and stability.

4.4 Discussion

As shown in Figure 6 (a), we evaluate the robustness of our method under challenging scenarios, including occlusions (e.g., hand gestures and glasses) and extreme poses. Unlike existing methods that primarily rely on appearance modeling, our approach leverages cross-domain motion perception, resulting in improved structural preservation and reduced distortion during dynamic generation.

In addition, as shown in Figure 6 (b), when trained with only about 5 seconds of video data, existing methods often suffer from artifacts and distortions, whereas our method reduces the reliance on identity-specific motion data and alleviates the need for long-term observations, thereby still generating visually coherent results.

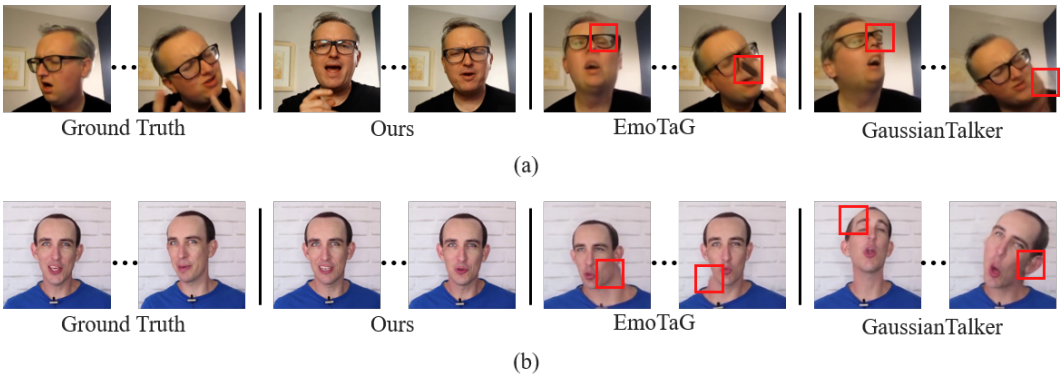


Figure 6: Applications in extreme scenarios.

5 Conclusion

This paper proposes a talking portrait generation framework based on cross-domain motion representation. First, we introduce AR-A (from audio-visual domain representation to alignment) to improve cross-modal synchronization between audio and portrait structures. Subsequently, we propose DR-F (from density domain representation to fusion) to enhance cross-region motion interaction within continuous sequences. Experiments demonstrate that the proposed method outperforms state-of-the-art methods in terms of motion coordination.

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

Shivangi Aneja, Artem Sevastopolsky, Tobias Kirschstein, Justus Thies, Angela Dai, and Matthias Nießner. Gaussianspeech: Audio-driven personalized 3d gaussian avatars. In *2025 IEEE/CVF International Conference on Computer Vision*, pp. 13065–13075. IEEE, 2025.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

Yanqi Bao, Tianyu Ding, Jing Huo, Yaoli Liu, Yuxin Li, Wenbin Li, Yang Gao, and Jiebo Luo. 3d gaussian splatting: Survey, technologies, challenges, and opportunities. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(7):6832–6852, 2025.

Jianchuan Chen, Jingchuan Hu, Gaige Wang, Zhonghua Jiang, Tiansong Zhou, Zhiwen Chen, and Chengfei Lv. Taoavatar: Real-time lifelike full-body talking avatars for augmented reality via 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10723–10734, 2025.

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.

Shunian Chen, Hejin Huang, Yexin Liu, Zihan Ye, Pengcheng Chen, Chenghao Zhu, Michael Guan, Rongsheng Wang, Junying Chen, Jianye Hou, et al. Talkvid: A large-scale diversified dataset for audio-driven talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3492–3500, 2026.

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.

Zhaojie Chu, Kailing Guo, Xiaofen Xing, Yilin Lan, Bolun Cai, and Xiangmin Xu. Corrtalk: Correlation between hierarchical speech and facial activity variances for 3d animation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(9):8953–8965, 2024.

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.

Helisa Dhamo, Yinyu Nie, Arthur Moreau, Jifei Song, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Headgas: Real-time animatable head avatars via 3d gaussian splatting. In *European Conference on Computer Vision*, pp. 459–476. Springer, 2024.

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.

Ben Fei, Jingyi Xu, Rui Zhang, Qingyuan Zhou, Weidong Yang, and Ying He. 3d gaussian splatting as new era: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

Shengjie Gong, Haojie Li, Jiapeng Tang, Dongming Hu, Shuangping Huang, Hao Chen, Tianshui Chen, and Zhuoman Liu. Monocular and generalizable gaussian talking head animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5523–5534, 2025.

Qianyun He, Xinya Ji, Yicheng Gong, Yuanxun Lu, Zhengyu Diao, Linjia Huang, Yao Yao, Siyu Zhu, Zhan Ma, Songcen Xu, et al. Emotalk3d: High-fidelity free-view synthesis of emotional 3d talking head. In *European Conference on Computer Vision*, pp. 55–72. Springer, 2024.

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.

Shoukang Hu, Tao Hu, and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20418–20431, 2024.

Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, and Qing Wang. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4568–4577, 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.

Jianwen Jiang, Chao Liang, Jiaqi Yang, Gaojie Lin, Tianyun Zhong, and Yanbo Zheng. Loopy: Taming audio-driven portrait avatar with long-term motion dependency. In *International Conference on Learning Representations*, volume 2025, pp. 15245–15263, 2025.

Yuheng Jiang, Zhehao Shen, Penghao Wang, Zhuo Su, Yu Hong, Yingliang Zhang, Jingyi Yu, and Lan Xu. Hifi4g: High-fidelity human performance rendering via compact gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19734–19745, 2024.

Tobias Kirschstein, Javier Romero, Artem Sevastopolsky, Matthias Nießner, and Shunsuke Saito. Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12089–12100, 2025.

Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 505–515, 2024.

Hebeizi Li, Benyuan Sun, Yi Yang, Zihao Liang, Zihao Yin, Xiao Sha, and Chenliang Wang. Unitalking: A unified audio-video framework for talking portrait generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4647–4656, 2026.

Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Talkinggaussian: Structure-persistent 3d talking head synthesis via gaussian splatting. In *European Conference on Computer Vision*, pp. 127–145. Springer, 2024a.

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.

Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19711–19722, 2024b.

Chang Liu, Tianjiao Jing, Chengcheng Ma, Xuanqi Zhou, Zhengxuan Lian, Qin Jin, Hongliang Yuan, and Shi-Sheng Huang. Emodifftalk: Emotion-aware diffusion for editable 3d gaussian talking head. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18063–18073, 2026a.

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, 2024a.

Yang Liu, Xiang Huang, Minghan Qin, Qinwei Lin, and Haoqian Wang. Animatable 3d gaussian: Fast and high-quality reconstruction of multiple human avatars. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1120–1129, 2024b.

Yifan Liu, Shengjun Zhang, Chensheng Dai, Yang Chen, Hao Liu, Chen Li, and Yueqi Duan. Learning efficient and generalizable human representation with human gaussian model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11797–11806, 2025.

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, 2026b.

Yifeng Ma, Suzhen Wang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, Zhidong Deng, and Xin Yu. Styletalk: One-shot talking head generation with controllable speaking styles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1896–1904, 2023.

Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1165–1175, 2024.

Pramish Paudel, Anubhav Khanal, Danda Pani Paudel, Jyoti Tandukar, and Ajad Chhatkuli. ihuman: Instant animatable digital humans from monocular videos. In *European Conference on Computer Vision*, pp. 304–323. Springer, 2024.

Ziqiao Peng, Yi Chen, Yifeng Ma, Guozhen Zhang, Zhiyao Sun, Zixiang Zhou, Youliang Zhang, Zhengguang Zhou, Zhaoxin Fan, Hongyan Liu, et al. Actavatar: Temporally-aware precise action control for talking avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39983–39993, 2026.

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.

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*, pp. 20299–20309, 2024a.

Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5020–5030. IEEE, 2024b.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1606–1616. IEEE, 2024.

Shuai Shen, Wanhua Li, Yunpeng Zhang, Weipeng Hu, Jiwen Lu, and Yap-Peng Tan. Hp-gaussian: Head prior guided gaussian splatting for personalized talking head synthesis from few-second video. *IEEE Transactions on Image Processing*, 2026.

Baiqin Wang, Xiangyu Zhu, Fan Shen, Hao Xu, and Zhen Lei. Pc-talk: precise facial animation control for audio-driven talking face generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25153–25162, 2026a.

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, 2026b.

Jie Wang, Jiu-Cheng Xie, Xianyan Li, Feng Xu, Chi-Man Pun, and Hao Gao. Gaussianhead: High-fidelity head avatars with learnable gaussian derivation. *IEEE Transactions on Visualization and Computer Graphics*, 2025.

Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. One-shot talking face generation from single-speaker audio-visual correlation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2531–2539, 2022.

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.

Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alexander G Schwing, and Shenlong Wang. Gomavatar: Efficient animatable human modeling from monocular video using gaussians-on-mesh. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2059–2069. IEEE, 2024.

Yifan Xie, Tao Feng, Xin Zhang, Xiangyang Luo, Zixuan Guo, Weijiang Yu, Heng Chang, Fei Ma, and Fei Richard Yu. Pointtalk: Audio-driven dynamic lip point cloud for 3d gaussian-based talking head synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8753–8761, 2025.

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.

Yuhang Yang, Fengqi Liu, Yixing Lu, Qin Zhao, Pingyu Wu, Wei Zhai, Ran Yi, Yang Cao, Lizhuang Ma, Zheng-Jun Zha, et al. Sigman: Scaling 3d human gaussian generation with millions of assets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5122–5133, 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.

Ye Yuan, Xueting Li, Yangyi Huang, Shalini De Mello, Koki Nagano, Jan Kautz, and Umar Iqbal. Gavatar: Animatable 3d gaussian avatars with implicit mesh learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 896–905. IEEE, 2024.

Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. Avatarrex: Real-time expressive full-body avatars. *ACM Transactions on Graphics*, 42(4):1–19, 2023.

Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics*, 39(6): 1–15, 2020.

Zhenglin Zhou, Fan Ma, Hehe Fan, Zongxin Yang, and Yi Yang. Headstudio: Text to animatable head avatars with 3d gaussian splatting. In *European Conference on Computer Vision*, pp. 145–163. Springer, 2024.

A Related Work

3D Gaussian Splatting (3DGS) has recently become an effective representation for digital human generation due to its ability to model high-fidelity appearance while supporting efficient rendering and controllable animation. Existing 3DGS-based methods have been extensively studied for both full-body digital humans and audio-driven talking portraits, with particular emphasis on motion modeling, cross-modal interaction, and fine-grained dynamic details. We therefore review these two research directions in the following subsections.

A.1 3DGS-based Full-body Digital Human

In recent years, research on full-body digital humans based on 3D Gaussian Splatting (3DGS) has developed rapidly Zheng et al. (2023); Huang et al. (2024); Chen et al. (2025); Wen et al. (2024). For single-person animatable avatars, 3DGS-Avatar Qian et al. (2024b) combines 3DGS with non-rigid deformation to reconstruct animatable clothed humans from monocular videos, while deformation regularization improves its generalization to highly articulated poses. GaussianAvatar Qian et al. (2024a) introduces animatable 3D Gaussians with dynamic appearance modeling and jointly optimizes motion and appearance to improve reconstruction from monocular videos. Animatable Gaussians Li et al. (2024b) employs a learned parametric template and 2D Gaussian maps to model pose-dependent appearance, enabling detailed and dynamic garment reconstruction with better generalization to novel poses.

Second, for efficient training and real-time rendering, GauHuman Hu et al. (2024), HiFi4G Jiang et al. (2024), and ASH Pang et al. (2024) explore different strategies to improve the efficiency of 3DGS-based human modeling. GauHuman uses canonical-space Gaussians with LBS and efficient Gaussian optimization, achieving 1–2 minute training. HiFi4G combines 3D Gaussians with non-rigid tracking and coarse-to-fine motion priors to improve optimization efficiency and reduce storage overhead. ASH attaches Gaussians to a deformable character model and learns their parameters in 2D texture space, enabling efficient real-time rendering of animatable humans. These methods demonstrate the potential of 3DGS for efficient and high-fidelity full-body digital human modeling.

Third, for cross-modal or text-driven generation, GAvatar Yuan et al. (2024) uses pose-driven primitives and implicit field prediction to generate high-fidelity, animatable digital humans from textual descriptions. In addition, methods such as Sigman Yang et al. (2025), HFGaussian, iHuman Paudel et al. (2024), and Animatable 3D Gaussian Liu et al. (2024b) further explore scalable and generalizable 3DGS-based human digitization. Sigman introduces a latent-space generation paradigm with a UV-structured VAE and DiT, enabling high-quality 3D human Gaussian generation from multi-view images. HFGaussian integrates Gaussian splatting with pose regression and feature splatting to jointly reconstruct human appearance and structural features in real time. iHuman binds 3D Gaussians to a parametric body template and uses normal supervision to improve geometry and animation quality from monocular videos. Animatable 3D Gaussian models skinned Gaussians with pose-dependent shape and appearance, supporting efficient reconstruction of complex motions and multi-human scenes.

Overall, these works demonstrate the potential of 3DGS for full-body digital human generation, animation control, and efficient rendering, while also highlighting challenges in motion coordination, cross-modal fusion, and fine-grained detail capture.

A.2 Audio-driven Talking Portrait Generation

In recent years, 3DGS-based digital head avatars have also made significant progress Kirschstein et al. (2025); Shen et al. (2026); Gong et al. (2025). First, for animatable and expressive heads, methods such as EmoTalk3D He et al. (2024), GaussianSpeech Aneja et al. (2025), and GaussianHead Wang et al. (2025) have explored different strategies for modeling expressive and fine-grained facial dynamics. EmoTalk3D predicts 3D facial geometry from speech and synthesizes appearance with 4D Gaussians, enabling emotion-controllable and multi-view talking head generation. GaussianSpeech employs an audio-conditioned Transformer with 3DGS to generate temporally coherent lip and facial movements, while using wrinkle- and perceptual-based losses to enhance fine-grained facial details. GaussianHead combines a motion deformation field with a tri-plane representation

to capture intricate head dynamics and detailed textures, achieving high-fidelity reconstruction and animation with a compact model.

Second, for real-time and efficient head animation, GaussianTalker Cho et al. (2024) and HeadGaS Dhamo et al. (2024) explore explicit 3DGS representations for controllable and efficient facial animation. GaussianTalker binds Gaussians to a 3D facial model and introduces a speaker-specific motion translator to generate accurate lip movements from audio, achieving 130 FPS rendering. HeadGaS combines 3DGS with learnable latent features and parametric head-model parameters to generate expression-dependent appearance, achieving state-of-the-art real-time inference with over 10× faster rendering than previous methods. For few-shot and multimodal synthesis, InsTaG Li et al. (2025), MGGTalk Gong et al. (2025), and PointTalk Xie et al. (2025) explore personalized and generalizable digital head generation from limited training data or monocular inputs. These methods further improve geometry and appearance reconstruction through depth information or symmetry constraints.

In addition, several works further explore different extensions of 3DGS-based digital head generation. HeadStudio Zhou et al. (2024) generates realistic and animatable head avatars from text prompts by combining 3D Gaussians with an animatable head prior, while supporting real-time novel-view rendering and speech- or video-driven animation. SplattingAvatar Shao et al. (2024) integrates Gaussian Splatting with explicit mesh geometry, using the mesh to model low-frequency motion and deformation while Gaussians capture high-frequency geometry and appearance details. TalkingGaussian Li et al. (2024a) models facial motion through smooth deformation of persistent Gaussian primitives and further separates the face and inner-mouth regions to improve detailed lip-motion reconstruction. Avat3r Kirschstein et al. (2025) leverages large reconstruction models and pretrained 3D human priors to reconstruct high-quality animatable head avatars from only a few input images, substantially reducing inference requirements. These methods further improve the quality, controllability, efficiency, and applicability of 3DGS-based digital head generation.

Overall, these works demonstrate the potential of 3DGS for head animation, audio-driven generation, and high-fidelity rendering. However, challenges remain in few-shot adaptation, cross-modal control, and fine-grained facial detail representation.

B More Details of Our Method

This section provides the detailed mathematical formulation of the proposed audio-visual alignment and density fusion strategies. Following the computation flow of the framework, we further derive the multimodal feature extraction, cross-modal correlation modeling, hierarchical denoising and alignment in AR-A, as well as the motion-window construction, density decomposition, contextual enhancement, and cross-region fusion in DR-F. The resulting representations are subsequently used to condition the Gaussian state updates.

B.1 Audio-Visual Feature Extraction

Given a frame sequence $\{f_t\}_{t=1}^{T}$ and the corresponding audio signal a, the multimodal feature extractor produces visual and audio representations. For each frame, the visual feature can be written as:

$$\mathcal{V}_t = \mathcal{E}_V(f_t). \tag{22}$$

where $\mathcal{E}_V(\cdot)$ denotes the visual feature extractor. Similarly, the audio representation is obtained as:

$$\mathcal{A}_t = \mathcal{E}_A(a_t). \tag{23}$$

where $\mathcal{E}_A(\cdot)$ denotes the audio encoder and a_t represents the audio segment associated with frame t.

The multimodal extractor can therefore be viewed as:

$$\mathcal{F}_{multimodal} : (f_t, a) \to (\mathcal{V}_t, \mathcal{A}_t). \tag{24}$$

Before calculating cross-modal similarity, the two modalities can be projected into a common feature space. Let the projected representations be:

$$\widetilde{\mathcal{V}}_t = W_V \mathcal{V}_t + b_V, \qquad \widetilde{\mathcal{A}}_t = W_A \mathcal{A}_t + b_A. \tag{25}$$

Their normalized representations are:

$$\hat{\mathcal{V}}_t = \frac{\widetilde{\mathcal{V}}_t}{|\widetilde{\mathcal{V}}_t|_2}, \qquad \hat{\mathcal{A}}_t = \frac{\widetilde{\mathcal{A}}_t}{|\widetilde{\mathcal{A}}_t|_2}. \tag{26}$$

The cosine similarity between the visual representation at frame i and the audio representation associated with frame j is consequently:

$$s_{ij} = \cos\left(\widetilde{\mathcal{V}}_i, \widetilde{\mathcal{A}}_j\right) = \hat{\mathcal{V}}_i^\top \hat{\mathcal{A}}_j. \tag{27}$$

For the positive audio-visual pair (i, i), the probability assigned by the contrastive objective can be expressed as:

$$p_{i\to i} = \frac{\exp(s_{ii})}{\sum_j \exp(s_{ij})}. \tag{28}$$

The multimodal contrastive objective therefore encourages the matched audio-visual pair to have a larger similarity than mismatched pairs:

$$\mathcal{L}_{multimodal} = -\sum_i \log p_{i\to i}. \tag{29}$$

For a given visual feature, the gradient of the contrastive loss with respect to the similarity score s_{ij} is:

$$\frac{\partial \mathcal{L}_{multimodal}}{\partial s_{ij}} = p_{ij}\mathbb{I}(i = j). \tag{30}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Thus, the optimization increases the similarity of the matched audio-visual pair while decreasing the relative similarity of mismatched pairs.

B.2 Audio-Visual Domain Representation

Different portrait structures exhibit different degrees of correlation with the audio signal. The cross-modal interaction is modeled using a dynamic attention mechanism. Let the visual and audio features be projected into query, key, and value representations:

$$Q_V = \mathcal{V}_t W_Q^V, \qquad K_A = \mathcal{A}_t W_K^A, \qquad V_A = \mathcal{A}_t W_V^A. \tag{31}$$

Specifically, the visual queries attend to the audio keys to obtain cross-modal feature weights:

$$attention_t = \mathrm{softmax}\left(\frac{Q_V K_A^\top}{\sqrt{d_A}}\right). \tag{32}$$

where d_A denotes the dimension of the audio key representation. The resulting audio-conditioned visual representation is then obtained by applying these attention weights to the projected audio values:

$$\mathrm{Attn}_{V\leftarrow A} = attention_t V_A. \tag{33}$$

Similarly, visual information can be used to condition the audio representation. The corresponding projections are:

$$Q_A = \mathcal{A}_t W_Q^A, \qquad K_V = \mathcal{V}_t W_K^V, \qquad V_V = \mathcal{V}_t W_V^V. \tag{34}$$

The reverse cross-modal attention is then formulated as:

$$\mathrm{Attn}_{A\leftarrow V} = \mathrm{softmax}\left(\frac{Q_A K_V^\top}{\sqrt{d_V}}\right) V_V. \tag{35}$$

where d_V denotes the dimension of the visual key representation.

The cross-modal interaction can consequently be represented abstractly as:

$$\mathcal{G}_t^n = \Phi_n\left(\mathcal{V}_t, \mathcal{A}_t, \mathrm{Attn}_{V\leftarrow A}, \mathrm{Attn}_{A\leftarrow V}\right). \tag{36}$$

where $\Phi_n(\cdot)$ denotes the pattern-specific projection for the n-th portrait structure.

The resulting audio-visual mapping is therefore:

$$\mathcal{F}_{audio\text{-}visual} : (\mathcal{V}_t, \mathcal{A}_t) \to \mathcal{G}_t^n, \qquad 1 \le n \le N. \tag{37}$$

To characterize the coupling strength of each portrait structure with audio, define the correlation coefficient:

$$\rho_n^t = \mathrm{corr}\left(\mathcal{G}_t^n, \mathcal{A}_t\right). \tag{38}$$

The correlation objective is then:

$$\mathcal{L}_{audio\text{-}visual} = \sum_{n=1}^{N} \left\| \rho_n^t - \gamma_n^t \right\|_2^2. \tag{39}$$

For each portrait structure, the derivative of the corresponding loss term with respect to the predicted correlation is:

$$\frac{\partial \left(\rho_n^t - \gamma_n^t\right)^2}{\partial \rho_n^t} = 2\left(\rho_n^t - \gamma_n^t\right). \tag{40}$$

Hence, the optimization drives the estimated correlation toward the predefined coupling strength γ_n^t. Larger values of γ_n^t impose stronger audio-visual coupling, whereas smaller values allow the corresponding portrait structure to retain a weaker dependence on the audio signal.

B.3 Hierarchical Denoising

The audio-visual representations obtained from consecutive frames may contain local temporal noise. For each hierarchical feature, the denoising network can be abstractly written as:

$$\mathcal{W}_t^n = \mathcal{F}_{denoise}\left(\mathcal{U}_t^n\right). \tag{41}$$

where $\mathcal{U}_t^n$ denotes the input hierarchical representation and $\mathcal{W}_t^n$ denotes the denoised representation.

The denoising objective consists of a temporal smoothness term and a feature-preservation term. The temporal smoothness component is:

$$\mathcal{L}_{smooth} = \sum_{t=1}^{T-1} \sum_{n=1}^{N} \left| \mathcal{W}_{t+1}^n - \mathcal{W}_t^n \right|_2^2. \tag{42}$$

For a single temporal transition, this term can be expanded as:

$$\left| \mathcal{W}_{t+1}^n - \mathcal{W}_t^n \right|_2^2 = |\mathcal{W}_{t+1}^n|_2^2 + |\mathcal{W}_t^n|_2^2 - 2(\mathcal{W}_{t+1}^n)^\top \mathcal{W}_t^n. \tag{43}$$

Thus, minimizing $\mathcal{L}_{smooth}$ encourages neighboring temporal features to remain close and suppresses high-frequency variations that are not consistent across consecutive frames.

The feature-preservation component is:

$$\mathcal{L}_{preserve} = \sum_{t=1}^{T} \sum_{n=1}^{N} |\mathcal{W}_t^n - \mathcal{U}_t^n|_2^2. \tag{44}$$

The complete denoising objective can therefore be interpreted as:

$$\mathcal{L}_{denoise} = \mathcal{L}_{smooth} + \mathcal{L}_{preserve}. \tag{45}$$

The first component suppresses abrupt temporal fluctuations, while the second prevents the denoising network from moving excessively far from the original dynamic representation. Consequently, the optimization provides a balance between temporal stability and preservation of motion-related information.

B.4 AUDIO-VISUAL ALIGNMENT

The denoised hierarchical representations are transformed into alignment features through:

$$\beta_t^n = \mathcal{F}_{alignment}\left(\mathcal{W}_t^n\right). \tag{46}$$

The complete alignment representation is obtained through aggregation:

$$\mathcal{R}_{audio-visual} = \mathcal{T}\left(\{\beta_t^n\}_{n=1}^N\right). \tag{47}$$

Let the normalized visual and audio features at the n-th hierarchical level be:

$$\hat{\mathcal{V}}_t^n = \frac{\mathcal{V}_t^n}{|\mathcal{V}_t^n|_2}, \qquad \hat{\mathcal{A}}_t^n = \frac{\mathcal{A}_t^n}{|\mathcal{A}_t^n|_2}. \tag{48}$$

Their squared Euclidean distance can be written as:

$$\left|\hat{\mathcal{V}}_t^n - \hat{\mathcal{A}}_t^n\right|_2^2 = (\hat{\mathcal{V}}_t^n)^\top \hat{\mathcal{A}}_t^n. \tag{49}$$

Therefore, the alignment loss:

$$\mathcal{L}_{alignment} = \sum_{t=1}^{T}\sum_{n=1}^{N}\left|\hat{\mathcal{V}}_t^n - \hat{\mathcal{A}}_t^n\right|_2^2. \tag{50}$$

is equivalent to maximizing the cosine similarity between the normalized visual and audio representations.

The corresponding gradient with respect to the visual feature is:

$$\frac{\partial\left|\hat{\mathcal{V}}_t^n - \hat{\mathcal{A}}_t^n\right|_2^2}{\partial\hat{\mathcal{V}}_t^n} = \left(\hat{\mathcal{V}}_t^n - \hat{\mathcal{A}}_t^n\right). \tag{51}$$

Thus, the alignment optimization progressively reduces the distance between the two normalized modalities at each hierarchical level.

B.5 MOTION WINDOW CONSTRUCTION

The density representation is constructed from local temporal motion rather than from a single frame. Given a temporal radius k, the input sequence centered at frame t is:

$$\mathcal{F}_{window} : (f_{t-k:t+k}) \to \mathcal{H}_t. \tag{52}$$

where the temporal window contains $2k+1$ consecutive frames.

For a local feature sequence, the temporal motion difference can be represented as:

$$\Delta\mathcal{H}_{t+\tau} = \mathcal{H}_{t+\tau+1}\mathcal{H}_{t+\tau}, \qquad \tau = -k, \ldots, k-1. \tag{53}$$

The corresponding local motion magnitude is:

$$m_{t+\tau} = |\Delta\mathcal{H}_{t+\tau}|_2. \tag{54}$$

These temporal variations provide the local motion statistics from which amplitude and frequency characteristics can be inferred.

The window-based self-supervised objective can be expressed as:

$$\mathcal{L}_{window} = -\operatorname{Var}\left(\sum_{t-k}^{t+k}|\mathcal{H}_t|\right). \tag{55}$$

where the variance measures the dispersion of the aggregated motion responses within the temporal window. This objective encourages the learned motion representation to preserve distinguishable temporal variation patterns rather than collapsing the responses of different temporal positions into an indistinguishable representation.

B.6 Density Hierarchical Decomposition

The motion representation $\mathcal{H}_t$ is decomposed into M hierarchical features. The mapping can be written as:

$$\mathcal{H}_t^m = \phi_m(\mathcal{H}_t), \qquad m = 1, \ldots, M. \tag{56}$$

Each hierarchical feature is subsequently mapped to a density coordinate:

$$l_m = \psi_m(\mathcal{H}_t^m). \tag{57}$$

The complete density mapping is therefore:

$$\mathcal{F}_{density} : \mathcal{H}_t \xrightarrow{MLP} \mathcal{H}_t^m \xrightarrow{mapping} l_m. \tag{58}$$

The alignment sequence is denoted by p_t and conditioned on the density coordinate l_m. The corresponding density-space representation is:

$$S_t^m = f(l_m; Q_t). \tag{59}$$

For two neighboring density levels, the finite difference of the mapped representation is:

$$\Delta_l S_t^m = \frac{S_t^{m+1} - S_t^m}{l_{m+1} - l_m + \delta_{(t,m)}}. \tag{60}$$

The hierarchical density objective constrains this local variation:

$$\mathcal{L}_{density} = \sum_{t=1}^{T} \sum_{m=1}^{M-1} |\Delta_l S_t^m - \mathbf{1}|_2^2 . \tag{61}$$

Expanding one term gives:

$$|\Delta_l S_t^m - \mathbf{1}|_2^2 = |\Delta_l S_t^m|_2^2 + |\mathbf{1}|_2^2 \ - 2(\Delta_l S_t^m)^\top \mathbf{1}.$$

Therefore, the optimization constrains the local rate of variation of the mapped alignment representation with respect to the density coordinate. The term $\delta_{(t,m)}$ provides an adaptive offset that prevents unstable scaling when the difference between adjacent density coordinates becomes small.

B.7 Contextual Enhancement

The continuous motion features are further processed to extract contextual information. For each density level, the contextual representation can be written as:

$$R_t^m = \mathcal{C}_m(\mathcal{H}_{t-k:t+k}). \tag{62}$$

where $\mathcal{C}_m(\cdot)$ denotes the convolutional contextual feature extractor.

The density-space representation and contextual representation are then jointly transformed:

$$Q_t^m = \mathcal{F}_{context}\left(S_t^m, R_t^m\right). \tag{63}$$

A general residual formulation of this contextual enhancement can be expressed as:

$$Q_t^m = S_t^m + \Psi_m\left(S_t^m, R_t^m\right). \tag{64}$$

where $\Psi_m(\cdot)$ denotes the learned context-dependent residual transformation.

The context consistency objective is:

$$\mathcal{L}_{context} = \sum_{t=1}^{T} \sum_{m=1}^{M} |Q_t^m - S_t^m|_2^2 . \tag{65}$$

For one hierarchical level, the corresponding gradient is:

$$\frac{\partial \mathcal{L}_{context}}{\partial Q_t^m} = \left(Q_t^m - S_t^m\right). \tag{66}$$

Thus, the contextual enhancement is encouraged to remain close to the original density representation while incorporating additional temporal context.

B.8 CROSS-REGION FUSION

The context-enhanced representations are mapped into cross-region features:

$$\varphi_t^m = \mathcal{F}_{fusion}\left(Q_t^m\right), \qquad m = 1, \ldots, M. \tag{67}$$

The hierarchical features can be aggregated into a joint density representation through:

$$\mathcal{R}_{density} = \mathcal{T}_{density}\left(\{\varphi_t^m\}_{m=1}^{M}\right). \tag{68}$$

where $\mathcal{T}_{density}(\cdot)$ denotes the fully connected aggregation operation.

To preserve the relative variation between neighboring hierarchical levels, define:

$$\Delta\varphi_t^m = \varphi_t^{m+1}\varphi_t^m. \tag{69}$$

and:

$$\Delta Q_t^m = Q_t^{m+1}Q_t^m. \tag{70}$$

The fusion objective can then be written as:

$$\mathcal{L}_{fusion} = \sum_{t=1}^{T}\sum_{m=1}^{M-1} |\Delta\varphi_t^m - \Delta Q_t^m|_2^2. \tag{71}$$

Expanding one term gives:

$$|\Delta\varphi_t^m - \Delta Q_t^m|_2^2 = |\Delta\varphi_t^m|_2^2 + |\Delta Q_t^m|_2^2 - 2(\Delta\varphi_t^m)^\top \Delta Q_t^m. \tag{72}$$

Consequently, minimizing the fusion objective encourages the fusion mapping to preserve the relative changes between neighboring density levels. Rather than forcing the absolute representations φ_t^m and Q_t^m to be identical, the loss preserves their hierarchical variation structure.

B.9 JOINT REPRESENTATION AND GAUSSIAN CONDITIONING

The two factorized representations obtained from AR-A and DR-F are combined to provide complementary conditioning information for the Gaussian state transition. Their joint representation can be written as:

$$Z_t = \Phi_G\left([\mathcal{R}_{audio-visual}; \mathcal{R}_{density}]\right). \tag{73}$$

where $\Phi_G(\cdot)$ projects the two representations into the feature space required by the Gaussian update network.

For the Gaussian state:

$$G = c, s, r, C, \alpha. \tag{74}$$

the state variation is obtained through:

$$\Delta G = \mathcal{F}_\theta\left(G, \mathcal{R}_{audio-visual}, \mathcal{R}_{density}\right). \tag{75}$$

The updated Gaussian state is then:

$$G' = G \oplus \Delta G. \tag{76}$$

where $\oplus$ denotes the parameter-wise additive or multiplicative composition specified for the corresponding Gaussian attributes.

For additive parameters, the update can be expressed as:

$$c' = c + \Delta c, \tag{77}$$

$$s' = s + \Delta s, \tag{78}$$

$$r' = r + \Delta r, \tag{79}$$

$$C' = C + \Delta C, \tag{80}$$

$$\alpha' = \alpha + \Delta\alpha. \tag{81}$$

where the increments are generated conditionally on the current Gaussian state and the two factorized representations.

More generally, the parameter-specific update can be written as:

$$\Delta g_i = \mathcal{F}_{\theta_i}\left(g_i, \mathcal{R}_{audio-visual}, \mathcal{R}_{density}\right). \tag{82}$$

where g_i denotes an individual Gaussian attribute and $\mathcal{F}_{\theta_i}$ denotes the corresponding parameter-specific mapping.

This formulation allows the audio-visual representation to provide information about the correspondence between audio signals and portrait structures, while the density representation provides information about the motion magnitude, temporal variation, and cross-region motion characteristics. Their joint conditioning consequently enables the Gaussian parameters to evolve according to both cross-modal synchronization and regional motion coordination.

B.10 Overall Optimization

The individual objectives introduced by AR-A and DR-F can be jointly optimized. The AR-A objective is:

$$\mathcal{L}_{AR-A} = \lambda_{multimodal}\mathcal{L}_{multimodal} + \lambda_{audio\text{-}visual}\mathcal{L}_{audio\text{-}visual} + \lambda_{denoise}\mathcal{L}_{denoise} + \lambda_{alignment}\mathcal{L}_{alignment}. \tag{83}$$

where the coefficients balance the multimodal contrastive, audio-visual correlation, denoising, and alignment objectives.

The DR-F objective is:

$$\mathcal{L}_{DR-F} = \lambda_{window}\mathcal{L}_{window} + \lambda_{density}\mathcal{L}_{density} + \lambda_{context}\mathcal{L}_{context} + \lambda_{fusion}\mathcal{L}_{fusion}. \tag{84}$$

The complete factorized objective can therefore be written as:

$$\mathcal{L}_{factor} = \mathcal{L}_{AR-A} + \mathcal{L}_{DR-F}. \tag{85}$$

When the reconstruction objective of the 3D Gaussian representation is included, the complete training objective becomes:

$$\mathcal{L} = \mathcal{L}_{render} + \mathcal{L}_{factor}. \tag{86}$$

where $\mathcal{L}_{render}$ supervises the visual fidelity of the reconstructed or rendered portrait frames.

Through joint optimization, AR-A learns to establish hierarchical audio-visual correspondences and suppress temporal noise while preserving cross-modal motion information. DR-F further organizes local motion into a hierarchical density space and preserves contextual and cross-region relationships during fusion. The resulting $\mathcal{R}_{audio-visual}$ and $\mathcal{R}_{density}$ jointly condition the Gaussian state transition, enabling the reconstructed portrait to maintain coherent motion coordination across temporal, cross-modal, and regional dimensions.

C More Details on Experiments

C.1 User Study

We conduct a user study to evaluate the subjective quality of the generated videos. A total of 100 participants are invited to watch multiple sets of talking-head videos. They rate each video on four aspects: image quality, video smoothness, motion naturalness, and identity preservation (scores range from 1 to 5, with higher scores indicating better performance).

Image Quality: This criterion evaluates the visual realism and detail fidelity of individual generated frames, including image sharpness, the reconstruction of facial and local texture details, and the presence of issues such as blurring, artifacts, or structural distortions. It also considers whether the overall illumination and appearance are natural and consistent, thereby assessing the quality of the generated results at the static visual level.

Video Smoothness: This criterion focuses on temporal consistency and stability in the generated videos, including whether transitions between adjacent frames are smooth, whether flickering, jittering, or abrupt changes occur, and whether the overall motion remains continuous. It therefore measures the temporal stability of the method in long-sequence generation.

Motion Plausibility: This criterion evaluates the realism and physical plausibility of the generated motion, including the correspondence between facial expressions, lip movements, and audio, whether head and overall pose movements follow natural human motion patterns, and whether local and global motions remain coordinated and consistent. It thereby assesses the credibility of the generated motion.

Identity Preservation: This criterion measures whether the generated results effectively preserve the identity characteristics of the target subject, including the stability of facial structure, facial features, and overall identity consistency across different frames. It also considers whether identity drift or identity confusion occurs, thereby evaluating the model's ability to preserve identity information.

We summarize the results in Figure 7. In terms of image quality, EmoTaG and InsTaG suffer from blurred details. TalkingGaussian may introduce noise during densification, which degrades structural clarity. GaussianTalker shows coarse structures under extreme poses. For video smoothness, PointTalk tends to produce over-smoothed or unnatural dynamics due to the lack of global modeling. InsTaG exhibits temporal discontinuities. Noise accumulation in TalkingGaussian also affects sequence stability. For motion naturalness, EmoDiffTalk is insufficient under extreme expressions. PointTalk has limited capability in modeling global motion and may lead to incoherent dynamics. GaussianTalker remains unstable under complex poses. Regarding identity preservation, InsTaG may sacrifice identity details in few-shot adaptation. TalkingGaussian and GaussianTalker may exhibit structural shifts under large motions or pose variations, which affects identity consistency.

In contrast, our method produces clearer details and more stable temporal consistency. By leveraging AR-A to enhance cross-modal alignment and DR-F to model structural motion relationships, our method generates more natural and coordinated dynamics while maintaining stable identity features under diverse motion variations. Overall, the user study shows that our method achieves higher subjective scores across all metrics.

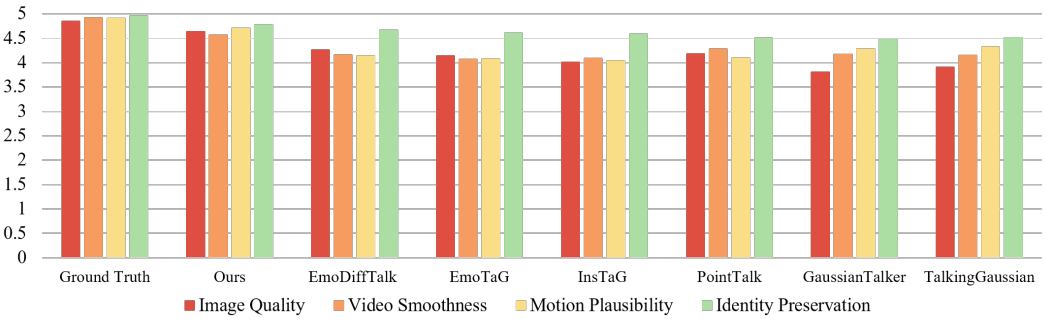


Figure 7: User study.

C.2 Model Performance

We evaluate the convergence speed and stability of the models by analyzing the evolution of the loss curves during training. Figure 8 shows the loss decreasing trends of our method and several baselines under the same training setup. To ensure fairness, all models are trained under identical hardware conditions with the following key settings: an initial learning rate of 1×10^{-3}, a batch size of 16 frames (adjusted according to GPU memory), and a total of 100 epochs. Most models begin to converge at around 80 epochs.

Overall, GaussianTalker achieves the best convergence performance, with a more sufficient loss reduction and an earlier transition to a stable stage in the later phase, indicating a relatively well-optimized process with a clear convergence boundary. TalkingGaussian follows, showing a generally smooth decreasing trend but still exhibiting noticeable fluctuations in the middle and later stages,

suggesting some instability during optimization. EmoDiffTalk and InsTaG perform at a moderate level, maintaining a balance between convergence speed and stability, but with limited convergence depth and a slower decline in the later stage. In contrast, EmoTaG and PointTalk perform relatively poorly, with limited loss reduction and more pronounced fluctuations, indicating that their optimization processes are more susceptible to disturbances and exhibit weaker overall convergence. The performance of these two methods is also quite similar.

In comparison, our method shows a faster decrease in the early training stage and maintains a more stable convergence trend in the middle and later stages. Benefiting from cross-domain motion-aware modeling, the model can better handle interactions across heterogeneous modalities and achieve more consistent motion representation. Meanwhile, the loss curve exhibits smaller fluctuations, indicating a more stable training process and overall favorable convergence behavior.

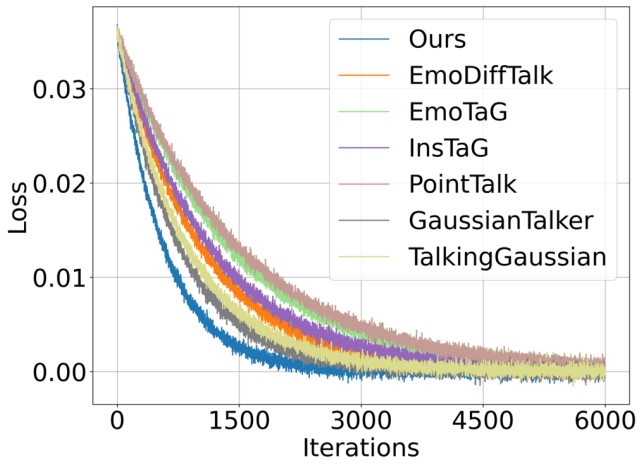


Figure 8: Model performance.

C.3 Parameter Settings

We conduct parameter setting experiments focusing on the coupling strength parameter γ_n^t and the weighting parameter $\delta_{(t,m)}$ to validate the correlation between audio and portrait structures as well as the hierarchical variations among different motion components, thereby improving model performance.

Coupling strength parameter γ_n^t: In the proposed correlation loss $\mathcal{L}_{audio-visual}$, the coupling strength parameter $\gamma_n^t \in [0, 1]$ is used to characterize the sensitivity of different portrait structures to the audio signal. As illustrated in Figure 9, when γ_n^t takes a larger value, the model tends to strengthen the correlation between audio and the corresponding visual structures, making these regions more sensitive to audio variations and potentially leading to unnatural motion patterns. In contrast, when γ_n^t is small, the model's response to the audio signal is weakened, and the motion of the related structures relies more on their own visual priors rather than audio-driven cues, which may result in temporal inconsistency in the portrait structures.

When $\gamma_n^t = 0.6$, the model achieves the best visual quality, indicating a favorable balance between audio-driven motion and visual priors. Therefore, a proper setting of γ_n^t is important for balancing audio-driven information and visual priors, thereby improving the robustness and expressive capability of the model.

Weight parameter $\delta_{(t,m)}$: In the hierarchical density loss $\mathcal{L}_{density}$, a weighting parameter $\delta(t, m) \in [0, 1]$ is introduced to regulate the magnitude of variation between different density levels. As illustrated in Figure 10, when $\delta_{(t,m)}$ takes a smaller value, the loss function becomes more sensitive to subtle inter-level variations, which may lead to training instability or even gradient explosion, potentially manifesting as blurry artifacts in local regions. In contrast, when $\delta_{(t,m)}$ is larger, it improves training stability but weakens the distinction between different levels, making it difficult for

the model to capture fine-grained hierarchical motion structures. This may result in overly smooth motion with insufficient detail variations.

In practice, when $\delta_{(t,m)} = 0.7$, the model achieves a favorable balance between training stability and expressive capability, while maintaining reasonable hierarchical differences across different time steps and levels, thereby achieving the best visual quality.



Figure 9: Comparison of parameters of γ_n^t. When $\gamma_n^t = 0.6$, the visual performance is optimal, as shown in the green box.

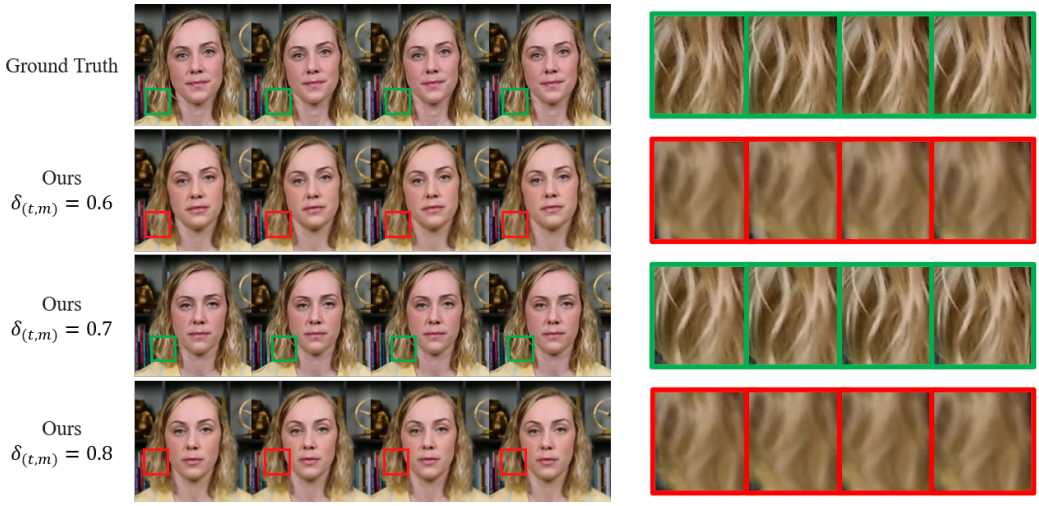


Figure 10: Comparison of parameters of $\delta_{(t,m)}$. When $\delta_{(t,m)} = 0.7$, the visual performance is optimal, as shown in the green box.

C.4 LIMITS

Our method proposes cross-domain motion perception to achieve the generation of coordinated and natural talking videos. Compared with existing methods that rely on limited observations in training data, our approach can generate more realistic motions with a certain degree of generalization capability.

However, it should be noted that, apart from lip movements, there is no strict deterministic mapping between audio and other motion patterns, which may lead to physically implausible motions in some

cases. As shown in Figure 11, for long-duration video generation (e.g., over 70 minutes), structural distortions may occur.

In future work, we plan to incorporate spatial physical constraints and stronger logical modeling to further improve the stability and robustness of multimodal generation.



Figure 11: Limitations of this method. Structural distortions in long-duration talking portrait generation.