WHY NOT RETHINK TALKING PORTRAIT GENERATION THROUGH DYNAMIC EVOLUTION? MOTION INTERACTION REPRESENTATION MAKES IT POSSIBLE

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, 3DGS-based methods have made significant progress in portrait reconstruction and generation. However, existing methods still struggle to capture the interactions among multiple motion factors over time, often resulting in incoherent fine-grained dynamics and insufficient coordination across different motion components. The fundamental reason is that 3DGS adopts an explicit parameterized representation based on discrete Gaussian primitives, which primarily describes spatial states but lacks an effective mechanism for characterizing the interactions among motion factors and their state evolution. We argue that the key to dynamic portrait generation lies in learning an interaction-oriented state evolution mechanism that enables the model to predict future changes based on the current dynamic state, rather than relying on local observations for frame-by-frame fitting. To this end, we propose a talking portrait generation framework based on motion interaction representation, which establishes relationships among dynamic factors from two aspects: static-dynamic factor fusion (S-DF) and rigid-flexible factor alignment (R-FA). S-DF establishes the static-dynamic interaction representation to enhance the expression of fine-grained dynamic evolution, while R-FA develops the rigid-flexible interaction representation to enable coordinated dynamic changes across different regions. Extensive experiments on talking portrait generation demonstrate that our method effectively captures the interactions among motion factors, resulting in portrait videos with higher fidelity in fine-grained details and more natural motion coordination.

1 INTRODUCTION

Talking portrait generation aims to synthesize videos with accurate lip synchronization, natural facial expressions, and visual consistency from an input portrait and driving signals Dong & Zhang (2025); Chu et al. (2026); Wang et al. (2026); Sun et al. (2026). It has broad applications in virtual humans, film production, and human-computer interaction. However, existing methods typically rely on local observations or explicit motion constraints for frame-by-frame generation, making it difficult to capture the relationships among different motion factors as they evolve over time.

In recent years, 3D Gaussian Splatting has made significant progress in human-centric 3D reconstruction and dynamic modeling Bao et al. (2025); Hu et al. (2024); Zheng et al. (2024), owing to its explicit 3D Gaussian representation and efficient differentiable rendering capabilities Kerbl et al. (2023); Yu et al. (2024); Fei et al. (2024). Building upon this representation, recent studies have further extended 3DGS to full-body human animation by incorporating pose-driven control, deformation modeling, and dynamic Gaussian optimization Liu et al. (2024); Hu et al. (2024); Wen et al. (2024), aiming to maintain generalization while achieving consistency between appearance and motion. However, talking portrait generation involves a series of complex and fine-grained facial dynamics Li et al. (2024a); Peng et al. (2025); Shen et al. (2026), including variations in lip opening and closing, intermittently appearing wrinkles, hair motion, and local texture changes. These variations are not independent of each other but are jointly influenced by different motion factors, spatial regions, and temporal stages. Existing 3D Gaussian-based methods mainly focus on optimizing Gaussian parameters or improving rendering efficiency, but lack effective mechanisms to model the interactions and coordinated evolution among different dynamic factors.

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 further observe that portrait animation can be essentially regarded as a state evolution process jointly driven by multiple factors. On the one hand, static structure provides spatial references for dynamic changes Jang et al. (2023); Xie et al. (2024). However, static structures remain relatively stable while dynamic details continuously change across space and time, making it difficult to preserve fine-grained variations such as tooth exposure, mouth shape, and facial expression. On the other hand, different regions exhibit distinct motion characteristics Weise et al. (2011); Athar et al. (2022). For example, head pose mainly involves rigid motion, whereas the eyes, hair strands, and skin textures exhibit more complex non-rigid variations. Based on these observations, we argue that the key to dynamic portrait generation lies in learning a state evolution mechanism that captures interactions among motion factors, enabling the model to infer future changes from the current dynamic state. To this end, we propose a motion interaction representation framework for talking portrait generation. Specifically, we first propose static-dynamic factor fusion (S-DF). It establishes the relationship between static structure and dynamic changes in different temporal stages and spatial regions by fusing them, forming a static-dynamic interactive representation, thereby enhancing the modeling of fine-grained time-varying dynamics. Simultaneously, we propose rigid-flexible factor alignment (R-FA). It aligns different rigid-flexible motion regions based on the motion properties, establishing a rigid-flexible interactive representation, thereby enabling the coordinated evolution of global motion trends and local dynamic changes.

The main contribution of this paper is to propose a talking portrait generation framework based on motion interaction representation, which characterizes the relationships among multiple motion factors. Specifically, we propose the static-dynamic factor fusion (S-DF) strategy to capture fine-grained time-varying details. Furthermore, we propose the rigid-flexible factor alignment (R-FA) strategy to coordinate motion across multiple regions. Extensive experimental results demonstrate that the proposed method effectively improves the reconstruction of fine-grained dynamic details and motion coordination in talking portrait generation.

2 OVERVIEW

The overall framework of the proposed method is illustrated in Figure 1, which consists of two core strategies: static-dynamic factor fusion (S-DF) and rigid-flexible factor alignment (R-FA).

In the S-DF module, we first decompose the input video into static structural information and dynamic variations, which differ substantially in spatial extent, motion magnitude, and temporal persistence, ranging from low-motion regions such as the background to highly dynamic regions such as the mouth. Based on these observations, we progressively fuse dynamic information from different spatiotemporal dimensions with static structural information, enhancing the preservation of fine-grained time-varying details. In the R-FA module, we further analyze the differences in regional motion properties within the dynamic variations. In general, head-pose changes can be regarded as a typical form of rigid motion, whereas local variations such as hair movements exhibit flexible motion characteristics. To capture the relationships between different motion properties, we use rigid regions as structural motion references to guide the dynamic alignment of flexible regions, thereby learning coordinated motion patterns across different regions.

Finally, S-DF and R-FA respectively produce the static-dynamic interactive representation $\mathcal{Z}_{S-DF}$ and the rigid-flexible interactive representation $\mathcal{Z}_{R-FA}$. Together, these representations characterize the relationships among different motion factors and drive the continuous evolution of Gaussian states, enabling high-fidelity, natural, and coherent talking portrait generation.

2.1 3DGS AND UPDATES

For the 3D Gaussian Splatting representation, each Gaussian point is parameterized as:

$$G = \{c, s, r, C, \alpha\}. \tag{1}$$

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

To characterize the dynamic evolution of the Gaussian representation, we formulate it as an interaction-driven state update process. Specifically, given the Gaussian parameters G_t at time t,

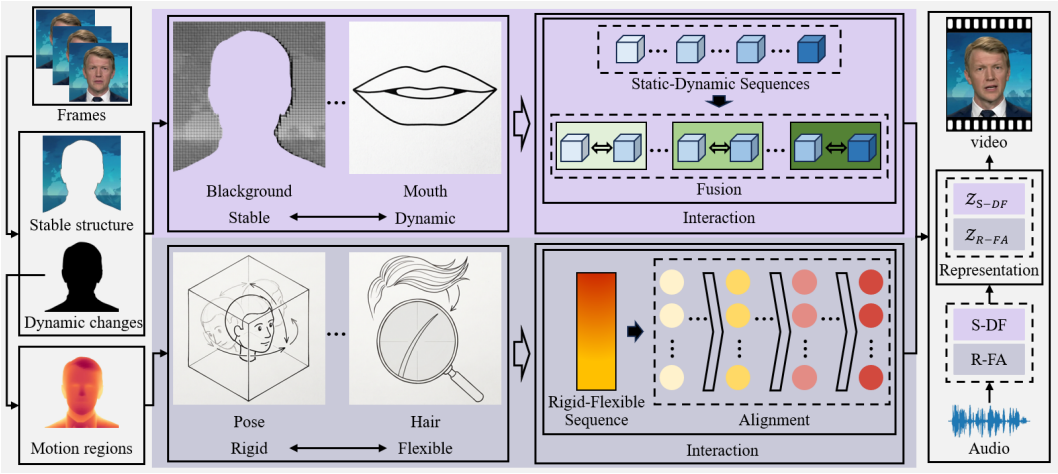

Figure 1: Overview of our framework, consisting of static-dynamic factor fusion (S-DF) and rigid-flexible factor alignment (R-FA). S-DF establishes the static-dynamic interaction through fusion, while R-FA establishes the rigid-flexible interaction through alignment.

the representation at the next time step can be written as:

$$G_{t+1} = f\left(G_t, \mathcal{Z}_{S-DF}, \mathcal{Z}_{R-FA}\right). \tag{2}$$

where $\mathcal{Z}_{S-DF}$ and $\mathcal{Z}_{R-FA}$ denote the static-dynamic interactive representation and rigid-flexible interactive representation, respectively. The mapping function f is implemented using a parameter-modulated multilayer perceptron.

3 METHOD

3.1 STATIC-DYNAMIC FACTOR FUSION

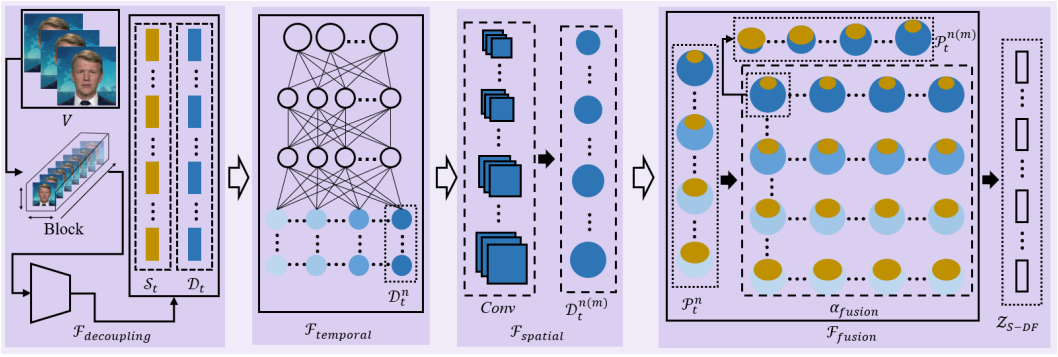

Figure 2: The process of static-dynamic factor fusion (S-DF).

The S-DF process is shown in Figure 2. First, given a continuous video frame sequence $V = \{I_t\}_{t=1}^{T}$, we define the video as a set of spatiotemporal blocks and design a static-dynamic decoupling module $\mathcal{F}_{decoupling}$ to extract the static structure S_t and dynamic change D_t, as follows:

$$\mathcal{F}_{decoupling} : I_t \rightarrow (S_t, D_t). \tag{3}$$

where $\mathcal{F}_{decoupling}$ employs a shared spatial feature extractor followed by a dual-branch architecture.

To prevent dynamic information from being interfered with by identity-related structural information, we further introduce a loss function based on static-dynamic decoupling constraints:

$$\mathcal{L}_{decoupling} = \left(\frac{S_t^{\top} D_t}{|S_t|_2 \, |D_t|_2}\right)^2. \tag{4}$$

Then, we design a temporal dynamic decomposition module $\mathcal{F}_{temporal}$, which uses a Transformer to capture the temporal relationships between continuous dynamic features and decompose them into multiple dynamic subspaces with different temporal variation patterns:

$$\mathcal{F}_{temporal} : D_t \rightarrow \{D_t^n\}_{n=1}^N. \tag{5}$$

where D_t^n denotes the n-th dynamic variation pattern. For dynamic patterns exhibiting recurrent temporal variations, we introduce a l-step temporal consistency constraint:

$$\mathcal{L}_{temporal} = \sum_n \left\| D_t^n - D_{t+l}^n \right\|_2^2. \tag{6}$$

where l denotes the temporal interval and is used to enhance the temporal correlation.

Next, to further capture local dynamic variations at different spatial scales, we design a spatial granularity decomposition module $\mathcal{F}_{spatial}$, which decomposes each dynamic pattern into multiple granularity levels according to the magnitude of motion variation, as follows:

$$\mathcal{F}_{spatial} : D_t^n \rightarrow \{D_t^{n(m)}\}_{m=1}^M. \tag{7}$$

where m denotes the dynamic granularity level. To maintain dynamic continuity between adjacent granularity levels while preserving their distinctiveness, we introduce the spatial granularity loss:

$$\mathcal{L}_{spatial} = \sum_{n=1}^{N} \left[\sum_{m=1}^{M-1} \left(1 - \text{sim}(D_t^{n(m)}, D_t^{n(m+1)}) \right) + \sum_{m=1}^{M-2} \text{sim}(D_t^{n(m)}, D_t^{n(m+2)}) \right]. \tag{8}$$

Finally, during the fusion stage, we design a static-dynamic fusion module $\mathcal{F}_{fusion}$ to establish associations between static structures and dynamic variations. Specifically, we first compute the interactions between static and dynamic features at different temporal and spatial scales:

$$(S_t, D_t^n) \rightarrow \mathcal{P}_t^n, \qquad (S_t, D_t^{n(m)}) \rightarrow \mathcal{P}_t^{n(m)}. \tag{9}$$

We then integrate these interaction features using $\mathcal{F}_{fusion}$ to obtain the fusion sequence α_{fusion}, which forms the static-dynamic interactive representation $\mathcal{Z}_{S\text{-}DF}$, as follows:

$$\mathcal{F}_{fusion} : (\mathcal{P}_t^n, \mathcal{P}_t^{n(m)}) \rightarrow \alpha_{fusion}. \tag{10}$$

We introduce a fusion loss that promotes effective association between static and dynamic features while suppressing irrelevant dynamic changes from interfering with the static structure:

$$\mathcal{L}_{fusion} = \sum_n \sum_m \sigma_{n(m)} \left(1 - \frac{\mathcal{P}_t^n \cdot \mathcal{P}_t^{n(m)}}{\|\mathcal{P}_t^n\|_2 \|\mathcal{P}_t^{n(m)}\|_2} \right). \tag{11}$$

where $\sigma_{n,m}$ denotes the adaptive weight assigned to different spatiotemporal scales. It is obtained from the relative similarity between the corresponding interaction representations:

$$\sigma_{n(m)} = \exp\left(\mathcal{P}_t^n \cdot \mathcal{P}_t^{n(m)} / \tau \right) \Big/ \sum_m \exp\left(\mathcal{P}_t^n \cdot \mathcal{P}_t^{n(m)} / \tau \right). \tag{12}$$

where τ is a temperature parameter that controls the smoothness of the similarity distribution, with parameter settings provided in Appendix C.3.

3.2 Rigid-Flexible Factor Alignment

The R-FA process is illustrated in Figure 3. First, for the dynamic variations in the video frame I_t, we use the region dynamic feature extraction module $\mathcal{F}_{region}$, based on ResNet-18 He et al. (2016) and 3D convolution, to extract motion features from different regions:

$$\mathcal{F}_{region} : (I_t) \rightarrow \mathcal{R}_t^x. \tag{13}$$

where $x \in [1, X]$ denotes the index of the spatial region, and $\mathcal{R}_t^x$ represents the dynamic feature of the x-th region at time t. We then introduce a regional attention mechanism to establish dynamic relationships $\mathcal{A}_t$ among different regions.

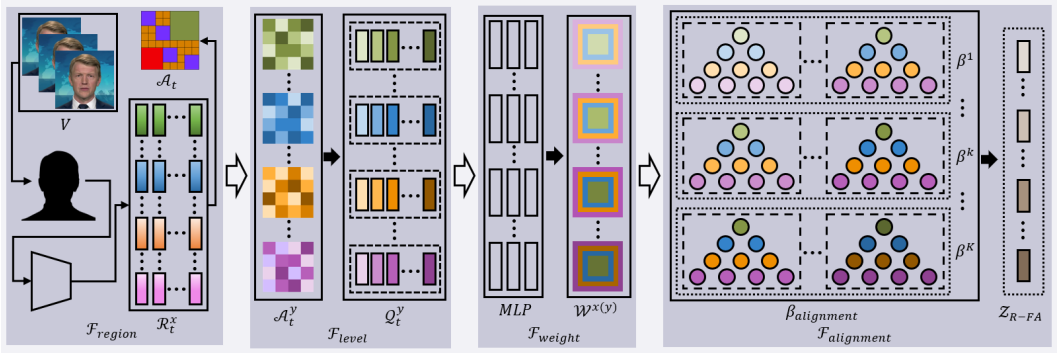

Figure 3: The process of rigid-flexible factor alignment (R-FA).

To constrain the temporal evolution of motion variations across different regions, we introduce a regional dynamic evolution loss:

$$\mathcal{L}_{region} = \sum_{x=1}^{X} \left\| (\mathcal{R}_{t+1}^x - \mathcal{R}_t^x) - (\mathcal{R}_{t+\Delta+1}^x - \mathcal{R}_{t+\Delta}^x) \right\|_2^2. \tag{14}$$

Subsequently, to further distinguish the motion properties of different regions, we design a rigid-flexible hierarchical decomposition module $\mathcal{F}_{level}$, which maps the dynamic attention representation into different motion levels, as follows:

$$\mathcal{F}_{level} : \mathcal{A}_t \rightarrow \{\mathcal{A}_t^y\}_{y=1}^Y. \tag{15}$$

where y denotes the hierarchical index ranging from rigid to flexible motion. The representations at different levels are further transformed into rigid-flexible feature sequences:

$$\{\mathcal{A}_t^y\}_{y=1}^Y \rightarrow \mathcal{Q}_t^y. \tag{16}$$

We introduce a loss function to ensure sufficient differences between different motion levels:

$$\mathcal{L}_{level} = \sum_{y} \max\left(0, d_y - d_{y+1} + h\right). \tag{17}$$

where $d_y = \left\| Q_t^{y+1} - Q_t^y \right\|_2$ denotes the change distance, and h is a margin parameter.

Next, to enable the adaptive combination of motion information from different regions, we design a weight prediction module $\mathcal{F}_{weight}$, which is calculated as follows:

$$\mathcal{F}_{weight} : (\mathcal{R}_t^x, \mathcal{A}_t^y) \rightarrow \mathcal{W}^{x(y)}. \tag{18}$$

where $\mathcal{F}_{weight}$ consists of multiple region-aware MLPs, and $\mathcal{W}^{x(y)}$ denotes the contribution weight of the x-th region at the y-th motion level during dynamic generation.

To encourage the predicted weights to reflect the dynamic importance of different regions, we further introduce a dynamic saliency constraint loss:

$$\mathcal{L}_{weight} = \sum_{y=1}^{Y} \sum_{x=1}^{X} \left\| \mathcal{W}^{x(y)} - \frac{h^x}{\sum_{x=1}^{X} h^x} \right\|_2^2. \tag{19}$$

where $h^x = \|\mathcal{R}_{t+1}^x - \mathcal{R}_t^x\|_2$ denotes the motion magnitude of the x-th region between adjacent temporal stages and serves as a measure of its dynamic saliency.

Finally, in the alignment module $\mathcal{F}_{alignment}$, we use the predicted weights $\mathcal{W}^{x(y)}$ to guide the alignment, obtaining the motion features β^k for different alignment levels k:

$$\mathcal{F}_{alignment} : \left(\mathcal{W}^{x(y)}, \mathcal{Q}_t^{(y)}\right) \rightarrow \beta^k. \tag{20}$$

Concatenating β^k produces the alignment feature sequence $\beta_{alignment}$, which is encoded to derive the rigid-flexible interaction representation $\mathcal{Z}_{R\text{-}FA}$. To preserve spatial structure and hierarchical semantic information during feature alignment, we introduce a spatially weighted alignment loss:

$$\mathcal{L}_{alignment} = \sum_{k=1}^{K} \lambda_k \left(1 - \frac{\beta^{k\top} \hat{\beta}^k}{\|\beta^k\|_2 \|\hat{\beta}^k\|_2} \right). \tag{21}$$

where λ_k denotes the weight coefficient for the k-th hierarchical level, which balances the contributions of multi-scale features during the alignment process. The parameter settings for λ_k are provided in Appendix C.3.

4 EXPERIMENT

To comprehensively evaluate the effectiveness of the proposed method, we conduct qualitative and quantitative comparisons, ablation studies, and related discussions. Additional details, including the 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. The training videos are collected from publicly available YouTube lectures and interviews. The videos are cropped, sampled at 25 FPS, and resized to 512×512, with all frames containing the speaker. All experiments are conducted on an NVIDIA RTX 5090 GPU.

Comparison methods. We compare our method with the following 3DGS-based approaches: GaussianTalker Cho et al. (2024), InsTaG Li et al. (2025), MGGTalk Gong et al. (2025), TaoAvatar Chen et al. (2025), EmoTaG Xu et al. (2026), and EmoDiffTalk Liu et al. (2026).

4.1 QUALITATIVE COMPARISON

As shown in Figure 4, we compare our method with the aforementioned representative approaches. Ground Truth denotes the real video corresponding to the input audio. We select two representative types of dynamic details: continuously varying tooth exposure and intermittently emerging wrinkles.

GaussianTalker performs frame-wise Gaussian attribute prediction based on 3DGS and spatial-audio attention, providing effective pose control but tending to exhibit local inconsistencies under complex motions and extreme viewpoints. InsTaG learns motion priors through pretraining and rapidly adapts to different identities. However, its ability to capture personalized details remains limited when the target identity differs substantially from the pretraining distribution. MGGTalk enhances geometric and appearance reconstruction using depth estimation and facial symmetry priors, but may still produce local distortions under extreme illumination or facial asymmetry. TaoAvatar enhances high-frequency detail modeling through personalized parameters, non-rigid deformation modeling, and distillation. However, its ability to model dynamics remains limited for complex poses beyond the training distribution. EmoTaG predicts motion in the parameter space and incorporates emotion and pose information to enhance facial expression generation, but remains limited in representing fine-grained variations such as extreme expressions. EmoDiffTalk employs emotion-aware Gaussian diffusion to enhance fine-grained facial animation and emotional editing, but incurs relatively high computational costs and remains limited in editing extreme expressions.

In contrast, our method provides a more effective representation of fine-grained dynamics. Specifically, S-DF establishes interactions between static structures and dynamic factors across different temporal stages, allowing local details to adapt to the current dynamic state while preserving identity-related structures. R-FA further uses rigid motion as a structural reference to align flexible regions, thereby maintaining consistency between local dynamics and overall facial motion. Together, these designs enable more accurate recovery of dynamic details with varying magnitudes and dynamic evolution patterns, resulting in more natural and coordinated talking portrait videos.

4.2 QUANTITATIVE COMPARISON

Table 1 reports the quantitative comparison between our method and existing approaches. We consider five aspects of talking portrait generation, including visual fidelity measured by PSNR,



Figure 4: Qualitative results. Green boxes highlight regions where Ours preserves details close to the Ground Truth, while red boxes indicate corresponding regions in the comparison methods with varying degrees of distortion.

Table 1: Quantitative results.

Methods	PSNR ↑	LPIPS ↓	SyncNet ↑	LSE-C ↑	LMD ↓	CSIM ↑	FPS ↑
Ground Truth	N/A	N/A	8.649	7.846	N/A	1	N/A
Ours	**36.48**	**0.028**	**7.942**	**7.514**	**1.08**	**0.97**	**163.7**
GaussianTalker	30.76	0.051	7.156	7.105	4.61	0.92	103.6
InsTaG	31.82	0.074	7.364	7.138	3.27	0.93	80.3
MGGTalk	33.15	0.068	7.269	7.346	2.63	0.91	41.8
TaoAvatar	30.26	0.092	7.441	7.267	3.06	0.92	121.5
EmoTaG	32.04	0.084	7.524	7.203	2.97	0.94	75.6
EmoDiffTalk	31.53	0.063	7.418	7.294	2.65	0.96	86.4

SSIM, and LPIPS Wang et al. (2004), audio–lip synchronization evaluated using SyncNet and LSE-C Chung & Zisserman (2017); Prajwal et al. (2020), motion smoothness assessed by AUE and LMD Chen et al. (2018), identity consistency measured by FID and CSIM Heusel et al. (2017); Huang et al. (2020), and inference efficiency reported in terms of FPS Ye et al. (2023a); Chen et al. (2019).

Our method demonstrates strong performance across the evaluation metrics. Compared with existing approaches, the proposed framework better preserves identity characteristics and visual realism, while more accurately recovering fine-grained variations during dynamic generation and improving audio–visual correspondence. In particular, the improved performance in image quality, lip synchronization, and motion smoothness indicates that the proposed state evolution mechanism with motion factor interactions can improve local detail representation and motion coordination. Meanwhile, the model maintains high inference efficiency, suggesting that the S-DF and R-FA modules introduce only limited computational overhead. These results demonstrate the effectiveness of modeling

interactions between different motion factors for talking portrait generation, and further show that interaction-driven Gaussian state updates provide an effective approach for dynamic generation.

4.3 Ablation Study

To evaluate the contributions of S-DF and R-FA, we conduct ablation studies by removing key components. All experiments use the same training settings and dataset for fair comparison. The results are shown in Figure 5 and Table 2.

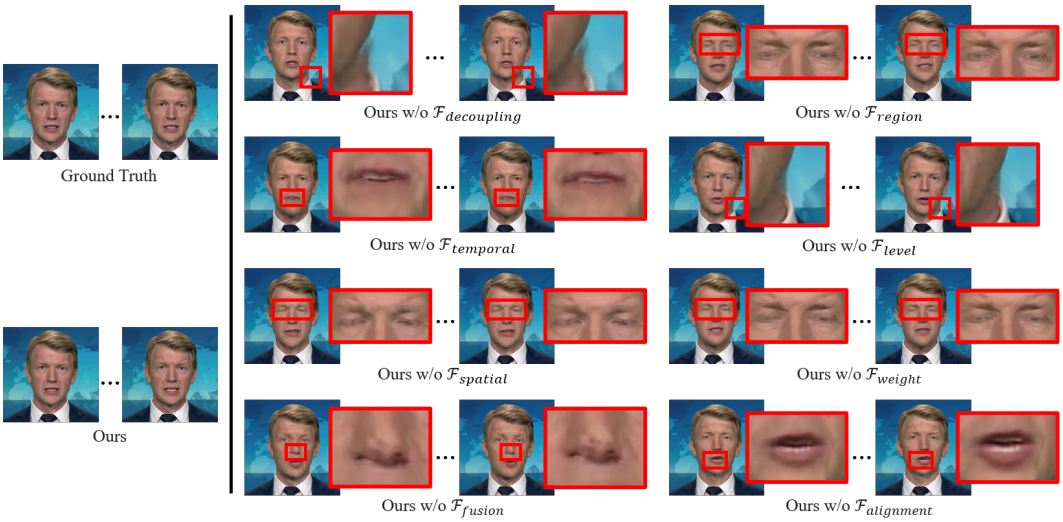

Figure 5: Qualitative results of the ablation study.

Table 2: Quantitative ablation study.

Methods	PSNR ↑	LPIPS ↓	SyncNet ↑	LSE-C ↑	LMD ↓	CSIM ↑
Ground Truth	N/A	N/A	8.649	7.846	N/A	1
Ours	**36.48**	**0.028**	**7.942**	**7.514**	**1.08**	**0.97**
Ours w/o $\mathcal{F}_{decoupling}$	32.47	0.046	7.226	7.304	2.79	0.97
Ours w/o $\mathcal{F}_{temporal}$	32.34	0.042	7.235	7.217	2.82	0.97
Ours w/o $\mathcal{F}_{spatial}$	32.64	0.047	7.281	7.161	2.93	0.97
Ours w/o $\mathcal{F}_{fusion}$	32.38	0.043	7.273	7.215	2.87	0.96
Ours w/o $\mathcal{F}_{region}$	32.75	0.044	7.304	7.283	2.77	0.96
Ours w/o $\mathcal{F}_{level}$	32.61	0.049	7.287	7.126	2.85	0.96
Ours w/o $\mathcal{F}_{weight}$	32.52	0.051	7.304	7.151	2.61	0.96
Ours w/o $\mathcal{F}_{alignment}$	32.67	0.045	7.391	7.311	2.74	0.97

Ours w/o $\mathcal{F}_{decoupling}$. $\mathcal{F}_{decoupling}$ separates structural information from dynamic variations in the input video frames. Without this module, static structures and dynamic patterns are difficult to distinguish, affecting both identity preservation and dynamic detail representation and leading to identity drift, structural instability, and weaker local dynamic details.

Ours w/o $\mathcal{F}_{temporal}$. $\mathcal{F}_{temporal}$ decomposes dynamic features into subspaces with different temporal variation patterns based on temporal dependencies, thereby distinguishing periodic motions from aperiodic details. Without this module, dynamic variations at different temporal scales are difficult to separate, affecting the representation of lip movements and fine-grained dynamics.

Ours w/o $\mathcal{F}_{spatial}$. $\mathcal{F}_{spatial}$ further refines dynamic features at different spatial granularities to jointly represent global motion trends and local details. Without this module, the model struggles to balance global motion and local details, leading to mixing of information across different scales and weaker dynamic representations.

Ours w/o $\mathcal{F}_{fusion}$. $\mathcal{F}_{fusion}$ establishes interactions between static structures and dynamic variations, allowing structural information to participate in the dynamic generation process. Without this module, the association between static and dynamic information is weakened, potentially resulting in local structural deviations and unnatural motion patterns.

Ours w/o $\mathcal{F}_{region}$. $\mathcal{F}_{region}$ extracts dynamic features from different spatial regions and models their motion relationships. Without this module, the model has difficulty distinguishing regional motion properties, causing local dynamic variations in regions such as the mouth and eyes to be weakened by global motion information and affecting coordinated motion across regions.

Ours w/o $\mathcal{F}_{level}$. $\mathcal{F}_{level}$ maps regional dynamic features to different motion levels to establish a hierarchical relationship from global rigid motion to local flexible variations. Without this module, the explicit hierarchy among different motion properties is lost. This weakens the representation of fine-grained non-rigid variations and reduces the coherence of overall motion.

Ours w/o $\mathcal{F}_{weight}$. $\mathcal{F}_{weight}$ adaptively adjusts the contributions of different regions and motion levels, allowing the model to emphasize regions with stronger dynamic variations. Without this module, different regions and motion levels lack differentiated contribution weights, affecting the stability of fine-grained detail representation.

Ours w/o $\mathcal{F}_{alignment}$. $\mathcal{F}_{alignment}$ integrates dynamic features from different regions and motion levels and constructs a unified rigid-flexible interaction representation. Without this module, an effective coordination mechanism across regions and motion levels is lost, making local dynamics more prone to deviating from the overall motion and thereby reducing temporal coherence.

4.4 Discussion

Our method has the potential to further expand its interactive capabilities. On the one hand, additional motion factors can be incorporated. As shown in Figure 6 (a), our method can further introduce hand gestures while preserving the original portrait dynamics, allowing different motion factors to remain coordinated without disrupting existing movements. This supports richer forms of interaction for digital humans. On the other hand, explicit user control signals can be introduced to regulate head motion and facial expressions. As shown in Figure 6 (b), independently adjusting different control conditions enables fine-grained control over factors such as head-motion amplitude and expression intensity. Overall, our method provides a potential direction toward more interactive, controllable, and state-evolving digital human systems.

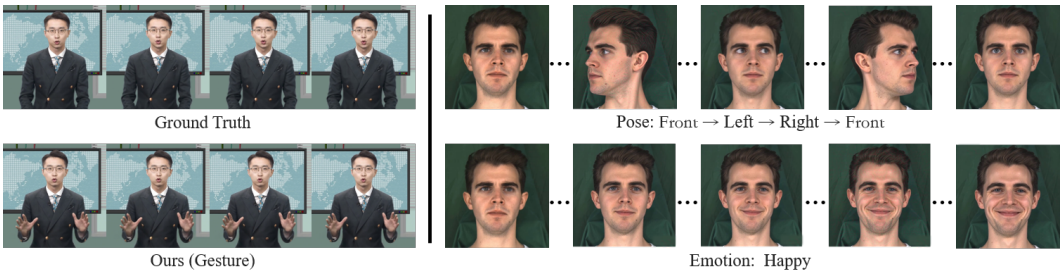

Figure 6: Applications in more interactive scenarios.

5 Conclusion

This paper proposes a talking portrait generation framework based on motion interaction representation. Specifically, S-DF (static-dynamic factor fusion) establishes the mapping between static structures and dynamic variations in different spatiotemporal dimensions, enhancing the expression of time-varying details. The R-FA (rigid-flexible factor alignment) aligns different motion regions according to their motion properties, enabling the coordinated evolution of global motion trends and local dynamic variations. Extensive experiments demonstrate that the proposed method can achieve effective interaction between motion factors, outperforming other 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

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.

ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20332–20341. IEEE, 2022.

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.

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.

Xuangeng Chu, Ruicong Liu, Yifei Huang, Yun Liu, Yichen Peng, and Bo Zheng. Unils: End-to-end audio-driven avatars for unified listening and speaking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25142–25152, 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, Yao Yao, Hao Zhu, Hanlin Shang, Kaihui Cheng, Hang Zhou, Siyu Zhu, and Jingdong Wang. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. In *International Conference on Learning Representations*, volume 2025, pp. 91659–91671, 2025a.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21086–21095. IEEE, 2025b.

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.

Yuan Gan, Zongxin Yang, Xihang Yue, Lingyun Sun, and Yi Yang. Efficient emotional adaptation for audio-driven talking-head generation. In *2023 IEEE/CVF International Conference on Computer Vision*, pp. 22577–22588. IEEE, 2023.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5523–5534. IEEE, 2025.

Jiazhi Guan, Zhanwang Zhang, Hang Zhou, Tianshu Hu, Kaisiyuan Wang, Dongliang He, Haocheng Feng, Jingtuo Liu, Errui Ding, Ziwei Liu, et al. Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1505–1515. IEEE, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

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.

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.

Youngjoon Jang, Kyeongha Rho, Jongbin Woo, Hyeongkeun Lee, Jihwan Park, Youshin Lim, Byeong-Yeol Kim, and Joon Son Chung. That's what i said: Fully-controllable talking face generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3827–3836, 2023.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions On Graphics*, 42(4):139–1, 2023.

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.

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, 2026.

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, 2024.

Ziqiao Peng, Wentao Hu, Junyuan Ma, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Hui Tian, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk++: High-fidelity and efficient synchronized talking heads synthesis using gaussian splatting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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.

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.

Xuli Shen, Hua Cai, Dingding Yu, Weilin Shen, Qing Xu, and Xiangyang Xue. Emohead: Emotional talking head via manipulating semantic expression parameters. In *2025 IEEE International Conference on Multimedia and Expo*, pp. 1–6. IEEE, 2025.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

Shuai Tan, Bin Ji, and Ye Pan. Emmn: Emotional motion memory network for audio-driven emotional talking face generation. In *2023 IEEE/CVF International Conference on Computer Vision*, pp. 22089–22099. IEEE, 2023.

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, 2026.

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

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.

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.

Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. Realtime performance-based facial animation. *ACM Transactions on Graphics*, 30(4):1–10, 2011.

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.

You Xie, Hongyi Xu, Guoxian Song, Chao Wang, Yichun Shi, and Linjie Luo. X-portrait: Expressive portrait animation with hierarchical motion attention. In *ACM SIGGRAPH 2024 conference papers*, pp. 1–11, 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*, 2024.

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*, 2023a.

Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, Jinzheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *arXiv preprint arXiv:2301.13430*, 2023b.

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19447–19456. IEEE, 2024.

Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19680–19690. IEEE, 2024.

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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4174–4184. IEEE, 2021.

A Related Work

A.1 Facial Animation

In the field of facial animation, early representative methods such as Wav2Lip Prajwal et al. (2020) and StyleSync Guan et al. (2023) primarily focus on accurate alignment between audio and lip movements. Wav2Lip introduces a discriminator to learn a strong audio-visual synchronization constraint, achieving accurate lip synchronization across different identities. StyleSync, on the other hand, is built on a style-based generative framework. It employs mask-guided spatial encoding and modulated convolutions to improve generation quality and generalization while preserving identity consistency.

To enhance motion expressiveness, a line of work uses videos as driving signals, including FOMM Siarohin et al. (2019), PC-AVS Zhou et al. (2021), and PD-FGC Wang et al. (2023). FOMM decouples appearance and motion using keypoints and local affine transformations, enabling strong generalization. PC-AVS introduces an implicit pose code for controllable head pose generation. PD-FGC further disentangles lip motion, eye movement, head pose, and emotion, allowing fine-grained control over multiple motion factors.

For emotion modeling, methods such as EAMN Tan et al. (2023), EmoHead Shen et al. (2025), and EAT Gan et al. (2023) incorporate explicit emotional conditions to improve expressiveness and realism. EAMN models the relationship between mouth motion and facial expressions using a memory network, alleviating regional inconsistencies. EmoHead leverages semantic expression parameters to enhance emotion controllability. EAT introduces parameter-efficient adaptation modules to incorporate emotion control into pretrained models, improving generalization.

With the rapid development of generative models, diffusion models and neural radiance fields have emerged as important paradigms. Diffusion-based approaches, such as the Hallo Xu et al. (2024); Cui et al. (2025a;b) series, employ hierarchical audio-driven synthesis and temporal alignment strategies, achieving improvements in visual quality, temporal consistency, and long-duration video generation. In contrast, NeRF-based methods such as GeneFace Ye et al. (2023b) and GeneFace++ Ye et al. (2023a) leverage 3D consistency to enhance realism, but still face limitations in training efficiency and long-term temporal stability. Overall, although existing methods have made notable progress in lip synchronization, motion control, and emotion modeling, challenges remain in achieving fine-grained dynamic consistency, multimodal coordination, and stable high-fidelity generation.

A.2 3DGS-based Talking Portrait Generation

Recent research on talking portrait generation has focused on 3D representations, cross-modal driving, and efficient modeling. With the emergence of 3D Gaussian Splatting, works such as GaussianAvatars Qian et al. (2024a), HeadGaS Dhamo et al. (2024), and SplattingAvatar Shao et al. (2024) adopt explicit Gaussian representations combined with parametric face models, enabling precise control over facial expressions, head poses, and viewpoints while maintaining high rendering efficiency.

In the audio-driven setting, methods including GaussianTalker Cho et al. (2024), PointTalk Xie et al. (2025), and GaussianSpeech Aneja et al. (2025) improve lip synchronization through cross-modal feature alignment and temporal modeling. However, they still face challenges in modeling complex expressions and maintaining training stability. For expression modeling and fine-grained detail synthesis, EmoTalk3D He et al. (2024) and GaussianHead Wang et al. (2025) enhance subtle dynamics such as micro-expressions and wrinkles through geometry-appearance disentanglement and dynamic Gaussian representations.

Furthermore, works such as InsTaG Li et al. (2025), MGGTalk Gong et al. (2025), and Avat3r Kirschstein et al. (2025) explore few-shot learning and cross-identity generalization, while TalkingGaussian improves dynamic consistency through deformation-aware modeling and disentangled representations.

Meanwhile, 3D Gaussian representations have also been extended to full-body digital human modeling. For example, 3DGS-Avatar Qian et al. (2024b) and GauHuman Hu et al. (2024) combine skeletal driving with deformation modeling to reconstruct animatable humans from monocular videos.

Animatable Gaussians Li et al. (2024b) efficiently model complex dynamic details such as clothing through template-based and 2D parameterized representations. TaoAvatar Chen et al. (2025) further extends this representation to full-body talking scenarios, enabling high-fidelity and real-time cross-device animation.

Overall, 3D Gaussian representations demonstrate strong expressive capability for digital human modeling. However, compared with full-body modeling, talking portrait generation imposes stricter requirements on fine-grained facial dynamics, such as lip movements and expression variations, as well as cross-modal alignment between audio and visual signals. These challenges make talking portrait generation a more delicate and demanding problem.

B More Details of Our Method

This section provides the detailed mathematical formulation of the proposed framework. Following the computation flow of the method, we further describe the feature decoupling, temporal and spatial decomposition, multi-scale fusion, regional motion modeling, hierarchical rigid-flexible decomposition, and Gaussian state evolution.

B.1 Static-Dynamic Decoupling

The static-dynamic decoupling begins with a shared spatial representation of each input frame. Let the output of the shared spatial feature extractor be denoted by:

$$F_t = \mathcal{E}(I_t). \tag{22}$$

where $\mathcal{E}(\cdot)$ extracts spatial features while sharing the common visual information between the static and dynamic branches. The two branches then project the shared representation into their respective feature spaces:

$$S_t = \mathcal{E}_S(F_t), \qquad D_t = \mathcal{E}_D(F_t). \tag{23}$$

Here, S_t captures the relatively persistent structural information, whereas D_t focuses on time-varying information.

To reduce the overlap between the two representations, their normalized correlation can be written as:

$$\rho_t = \frac{S_t^\top D_t}{|S_t|_2 |D_t|_2}. \tag{24}$$

The decoupling constraint is therefore:

$$\mathcal{L}_{decoupling} = \rho_t^2 \tag{25}$$

Minimizing this objective drives ρ_t toward zero and consequently encourages the two representations to become orthogonal. In particular, if D_t is decomposed into components parallel and orthogonal to S_t:

$$D_t = D_t^{\parallel} + D_t^{\perp}. \tag{26}$$

the parallel component is:

$$D_t^{\parallel} = \frac{S_t^\top D_t}{|S_t|_2^2} S_t. \tag{27}$$

Its squared magnitude is:

$$\left| D_t^{\parallel} \right|_2^2 = \frac{(S_t^\top D_t)^2}{|S_t|_2^2}. \tag{28}$$

Thus, minimizing $\mathcal{L}_{\text{decoupling}}$ suppresses the component of dynamic information that lies in the direction of the static representation, reducing the interference of identity-related structure in the dynamic branch.

B.2 Temporal Dynamic Decomposition

The dynamic representation D_t is further decomposed according to its temporal variation patterns. First, the dynamic features are projected into a common latent space:

$$E_t = W_D D_t + b_D. \tag{29}$$

For the sequence of continuous frames, the resulting features form:

$$E = [E_1, E_2, \ldots, E_T]^\top. \tag{30}$$

The temporal dependencies are then modeled using self-attention. For one attention head, the query, key, and value representations are:

$$Q = EW_Q, \qquad K = EW_K, \qquad V = EW_V. \tag{31}$$

The temporal interaction is calculated as:

$$H = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V. \tag{32}$$

where d is the feature dimension used for computing the attention scores.

The resulting representation H_t contains information aggregated from different temporal positions. Multiple dynamic variation patterns are then extracted through pattern-specific projections:

$$D_t^n = \phi_n(H_t), \qquad n = 1, \ldots, N. \tag{33}$$

Each D_t^n therefore describes a distinct temporal variation pattern while retaining the temporal context captured by the shared Transformer.

For a dynamic pattern that exhibits recurrent temporal behavior, its representation is encouraged to remain consistent over a temporal interval l. Considering all valid temporal positions, the consistency objective can be written as:

$$\mathcal{L}_{temporal} = \sum_{n=1}^{N}\sum_{t=1}^{T-l}\left|D_t^n - D_{t+l}^n\right|_2^2. \tag{34}$$

Expanding one term gives:

$$\left|D_t^n - D_{t+l}^n\right|_2^2 = |D_t^n|_2^2 + |D_{t+l}^n|_2^2 - 2(D_t^n)^\top D_{t+l}^n. \tag{35}$$

Therefore, when the feature magnitudes are controlled, minimizing the temporal consistency loss increases the correlation between the same dynamic pattern at temporally separated positions. The constraint consequently preserves recurrent temporal structures without requiring the absolute dynamic features at different frames to be identical.

B.3 Spatial Granularity Decomposition

Each temporal variation pattern is further decomposed according to its spatial granularity. For the n-th dynamic pattern, the representation at the m-th granularity level is obtained as:

$$D_t^{n(m)} = \psi_m(D_t^n), \qquad m = 1, \ldots, M. \tag{36}$$

where $\psi_m(\cdot)$ denotes the corresponding spatial decomposition and feature transformation.

The different granularity levels are expected to preserve related dynamic information while remaining distinguishable. This property is imposed by separating the spatial granularity objective into adjacent-level consistency and non-adjacent-level distinction:

$$\mathcal{L}_{spatial} = \mathcal{L}_{adj} + \mathcal{L}_{non-adj}. \tag{37}$$

where:

$$\mathcal{L}_{adj} = \sum_{n=1}^{N}\sum_{m=1}^{M-1}\left[1 - \mathrm{sim}\left(D_t^{n(m)}, D_t^{n(m+1)}\right)\right]. \tag{38}$$

and:

$$\mathcal{L}_{non-adj} = \sum_{n=1}^{N} \sum_{m=1}^{M-2} \operatorname{sim}\left(D_t^{n(m)}, D_t^{n(m+2)}\right). \tag{39}$$

The first term encourages neighboring granularity levels to share common dynamic information. In contrast, the second term suppresses excessive similarity between non-adjacent levels. Hence, the learned representations satisfy the desired hierarchical relationship:

$$\operatorname{sim}\left(D_t^{n(m)}, D_t^{n(m+1)}\right) \uparrow . \tag{40}$$

while:

$$\operatorname{sim}\left(D_t^{n(m)}, D_t^{n(m+2)}\right) \downarrow . \tag{41}$$

This construction prevents the spatial decomposition from collapsing into multiple redundant representations while maintaining continuity between neighboring spatial scales.

B.4 Static-Dynamic Interaction and Fusion

The static representation S_t is subsequently associated with the dynamic representations at different temporal and spatial scales. For the n-th temporal pattern, the corresponding interaction representation is written as:

$$\mathcal{P}_t^n = \Phi\left(S_t, D_t^n\right). \tag{42}$$

while the interaction at the m-th spatial granularity is:

$$\mathcal{P}_t^{n(m)} = \Phi\left(S_t, D_t^{n(m)}\right). \tag{43}$$

where $\Phi(\cdot,\cdot)$ denotes the static-dynamic interaction operation.

A general feature interaction can be expressed as:

$$\Phi(S, D) = W_P\left[S \odot D; S + D; S - D\right] + b_P. \tag{44}$$

where $\odot$ denotes element-wise multiplication and $[\cdot;\cdot]$ denotes feature concatenation. This formulation allows the interaction representation to retain both complementary and correlated information between static structure and dynamic variation.

The relative importance of different spatial granularities is determined from the interaction similarity. Define:

$$e_{n(m)} = \mathcal{P}_t^n \cdot \mathcal{P}_t^{n(m)}. \tag{45}$$

The adaptive weight is then obtained by a temperature-controlled softmax:

$$\sigma_{n(m)} = \frac{\exp(e_{n(m)}/\tau)}{\sum_{j=1}^{M} \exp(e_{n(j)}/\tau)}. \tag{46}$$

It follows directly that:

$$\sum_{m=1}^{M} \sigma_{n(m)} = 1, \qquad \sigma_{n(m)} \geq 0. \tag{47}$$

A smaller τ produces a more concentrated distribution and therefore places greater emphasis on the most similar interaction scale, whereas a larger τ produces a smoother weighting distribution.

Using these adaptive weights, the fusion constraint becomes:

$$\mathcal{L}_{fusion} = \sum_{n=1}^{N} \sum_{m=1}^{M} \sigma_{n(m)} \left(1 - \frac{\mathcal{P}_t^n \cdot \mathcal{P}_t^{n(m)}}{|\mathcal{P}_t^n|_2 |\mathcal{P}_t^{n(m)}|_2}\right). \tag{48}$$

Thus, the interaction features that are more strongly associated with the corresponding temporal dynamic pattern receive larger weights during optimization.

The resulting interaction features are integrated by the fusion function:

$$\alpha_{fusion} = \mathcal{F}_{fusion}\left(\mathcal{P}_t^n, \mathcal{P}_t^{n(m)}\right). \tag{49}$$

which provides the static-dynamic interactive representation:

$$\mathcal{Z}_{S\text{-}DF} = \mathcal{E}_{S\text{-}DF}\left(\alpha_{fusion}\right). \tag{50}$$

B.5 Regional Dynamic Representation

The dynamic variation is further analyzed according to its spatial regions. Let the extracted regional representations be:

$$\mathcal{R}_t = [\mathcal{R}_t^1, \mathcal{R}_t^2, \ldots, \mathcal{R}_t^X]. \tag{51}$$

The temporal and spatial feature extractor provides each region with a dynamic representation $\mathcal{R}_t^x$, allowing the subsequent modeling to distinguish regional motion properties.

The relationships between regions are obtained through regional attention. Specifically, the query, key, and value representations are:

$$Q_R = \mathcal{R}_t W_Q^R, \qquad K_R = \mathcal{R}_t W_K^R, \qquad V_R = \mathcal{R}_t W_V^R. \tag{52}$$

The regional attention representation is then:

$$\mathcal{A}_t = \mathrm{softmax}\left(\frac{Q_R K_R^\top}{\sqrt{d_R}}\right) V_R. \tag{53}$$

where d_R denotes the regional feature dimension.

The resulting representation incorporates interactions among different spatial regions rather than modeling each region independently. This allows the subsequent decomposition to capture coordinated motion patterns, including the relationship between relatively rigid regions and locally flexible regions.

B.6 Regional Dynamic Evolution

To constrain the temporal evolution of each regional motion pattern, define its first-order dynamic variation as:

$$\Delta\mathcal{R}_t^x = \mathcal{R}_{t+1}^x - \mathcal{R}_t^x. \tag{54}$$

The regional evolution constraint can then be expressed as:

$$\mathcal{L}_{region} = \sum_{x=1}^{X} \left\| \Delta\mathcal{R}_t^x - \Delta\mathcal{R}_{t+\Delta}^x \right\|_2^2. \tag{55}$$

Expanding the distance gives:

$$\mathcal{L}_{region} = \sum_{x=1}^{X} \left(\|\Delta\mathcal{R}_t^x\|_2^2 + \|\Delta\mathcal{R}_{t+\Delta}^x\|_2^2 - 2(\Delta\mathcal{R}_t^x)^\top \Delta\mathcal{R}_{t+\Delta}^x \right).$$

Consequently, the constraint encourages the temporal changes of the same region to exhibit consistent evolution over the interval Δ. Since the loss is defined on feature differences rather than absolute feature states, it focuses on the evolution of motion instead of forcing the regional features themselves to remain unchanged.

B.7 Rigid-Flexible Hierarchical Decomposition

The regional attention representation is organized into multiple motion levels:

$$\mathcal{A}_t^y = \phi_y(\mathcal{A}_t), \qquad y = 1, \ldots, Y. \tag{56}$$

where the hierarchy progressively describes different motion properties from rigid to flexible.

Each level is subsequently transformed into the corresponding motion feature:

$$\mathcal{Q}_t^y = W_Q^y \mathcal{A}_t^y + b_Q^y. \tag{57}$$

The difference between two consecutive motion levels is measured by:

$$d_y = \left| \mathcal{Q}_t^{y+1} - \mathcal{Q}_t^y \right|_2. \tag{58}$$

To maintain a meaningful hierarchy, the distances between consecutive levels are constrained by a margin:

$$\mathcal{L}_{level} = \sum_{y} \left[d_y - d_{y+1} + h \right]_+ . \tag{59}$$

where $[\cdot]_+$ denotes the hinge operator:

$$[z]_+ = \max(0, z). \tag{60}$$

For the corresponding term to become zero, the distance needs to satisfy:

$$d_{y+1} \geq d_y + h. \tag{61}$$

Therefore, successive motion levels are encouraged to maintain progressively distinguishable representations. The margin h prevents adjacent levels from becoming indistinguishable while avoiding unnecessary separation once the required difference has been established.

B.8 Dynamic Weight Prediction

The contribution of each region is determined according to its dynamic magnitude together with its motion-level representation. For region x and motion level y, the predicted contribution is written as:

$$\mathcal{W}^{x(y)} = \mathcal{F}_{weight} \left(\mathcal{R}_t^x, \mathcal{A}_t^y \right). \tag{62}$$

Since different regions may exhibit substantially different amounts of motion, the dynamic magnitude of region x is defined as:

$$h^x = \left| \mathcal{R}_{t+1}^x - \mathcal{R}_t^x \right|_2 . \tag{63}$$

Normalizing the regional motion magnitudes gives:

$$\pi^x = \frac{h^x}{\sum_{j=1}^{X} h^j}. \tag{64}$$

with:

$$\sum_{x=1}^{X} \pi^x = 1. \tag{65}$$

The dynamic saliency constraint encourages the predicted contribution weights to follow this relative motion importance:

$$\mathcal{L}_{weight} = \sum_{y=1}^{Y} \sum_{x=1}^{X} \left| \mathcal{W}^{x(y)} - \pi^x \right|_2^2 . \tag{66}$$

The gradient with respect to an individual predicted weight is:

$$\frac{\partial \mathcal{L}_{weight}}{\partial \mathcal{W}^{x(y)}} = 2 \left(\mathcal{W}^{x(y)} - \pi^x \right). \tag{67}$$

Thus, regions exhibiting larger temporal changes are encouraged to receive larger contributions during the alignment process, while regions with relatively weak motion are prevented from dominating the dynamic representation.

B.9 Rigid-Flexible Feature Alignment

The predicted regional weights are used to combine the motion representations at different hierarchical levels. For the k-th alignment level, the resulting motion feature is obtained through the alignment function:

$$\beta^k = \mathcal{F}_{alignment}^k \left(\mathcal{W}^{x(y)}, \mathcal{Q}_t^{(y)} \right). \tag{68}$$

where $\mathcal{F}_{alignment}^k(\cdot)$ denotes the alignment operation at the k-th level and incorporates the predicted regional contributions into the corresponding hierarchical motion representation.

Concatenating the features from all alignment levels gives:

$$\beta_{alignment} = [\beta^1; \beta^2; \cdots ; \beta^K]. \tag{69}$$

The rigid-flexible interaction representation is subsequently obtained through an encoding function:

$$\mathcal{Z}_{R\text{-}FA} = \mathcal{E}_{R\text{-}FA}\left(\beta_{alignment}\right). \tag{70}$$

To preserve the hierarchical spatial structure during this process, each alignment feature β^k is encouraged to remain consistent with its corresponding reference feature $\hat{\beta}^k$. The alignment objective is:

$$\mathcal{L}_{alignment} = \sum_{k=1}^{K} \lambda_k \left(1 - \frac{(\beta^k)^\top \hat{\beta}^k}{|\beta^k|_2 |\hat{\beta}^k|_2}\right). \tag{71}$$

When the two representations are normalized, the cosine distance has the equivalent form:

$$1 - (\beta^k)^\top \hat{\beta}^k = \frac{1}{2}\left|\beta^k - \hat{\beta}^k\right|_2^2. \tag{72}$$

Therefore, minimizing the alignment loss encourages the learned motion feature to approach the corresponding reference in the normalized feature space. The coefficient λ_k determines the contribution of each hierarchical alignment level.

B.10 Interaction-Driven Gaussian State Evolution

The two interaction representations provide complementary conditioning information for the evolution of the Gaussian representation. Their joint conditioning feature can be written as:

$$Z_t = \Phi_G\left([\mathcal{Z}_{S\text{-}DF}; \mathcal{Z}_{R\text{-}FA}]\right). \tag{73}$$

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

For a Gaussian state:

$$G_t = \{c_t, s_t, r_t, C_t, \alpha_t\}. \tag{74}$$

the parameter-modulated mapping produces a state variation:

$$\Delta G_t = \mathcal{M}\left(G_t, Z_t\right). \tag{75}$$

A residual formulation gives:

$$G_{t+1} = G_t + \Delta G_t. \tag{76}$$

The individual Gaussian parameters consequently evolve as:

$$c_{t+1} = c_t + \Delta c_t, \tag{77}$$
$$s_{t+1} = s_t + \Delta s_t, \tag{78}$$
$$r_{t+1} = r_t + \Delta r_t, \tag{79}$$
$$C_{t+1} = C_t + \Delta C_t, \tag{80}$$
$$\alpha_{t+1} = \alpha_t + \Delta \alpha_t. \tag{81}$$

The parameter-modulated multilayer perceptron can be expressed abstractly as:

$$H_t^{(1)} = \varphi\left(W_1 G_t + b_1\right). \tag{82}$$

followed by:

$$H_t^{(j+1)} = \varphi\left(\mathcal{M}_{j+1}\left(H_t^{(j)}, Z_t\right)\right). \tag{83}$$

where $\varphi(\cdot)$ is the nonlinear activation and $\mathcal{M}_{j+1}(\cdot,\cdot)$ represents the parameter modulation induced by the static-dynamic and rigid-flexible interaction representations.

This formulation allows the update of each Gaussian to depend jointly on its current state and the learned motion factors. The static-dynamic representation provides information about how dynamic changes should interact with persistent structure, while the rigid-flexible representation provides information about how motion should be coordinated across regions with different motion properties.

B.11 JOINT OPTIMIZATION

The constraints introduced throughout the framework can be jointly optimized as:

$$\begin{aligned}\mathcal{L}_{factor} = \lambda_{decoupling}\mathcal{L}_{decoupling} + \lambda_{temporal}\mathcal{L}_{temporal} + \lambda_{spatial}\mathcal{L}_{spatial} + \lambda_{fusion}\mathcal{L}_{fusion} \\ + \lambda_{region}\mathcal{L}_{region} + \lambda_{level}\mathcal{L}_{level} + \lambda_{weight}\mathcal{L}_{weight} + \lambda_{alignment}\mathcal{L}_{alignment}.\end{aligned} \tag{84}$$

where the coefficients control the relative contribution of the corresponding constraints.

The objective can be further separated according to the two factorization strategies:

$$\mathcal{L}_{S\text{-}DF} = \lambda_{decoupling}\mathcal{L}_{decoupling} + \lambda_{temporal}\mathcal{L}_{temporal} + \lambda_{spatial}\mathcal{L}_{spatial} + \lambda_{fusion}\mathcal{L}_{fusion}. \tag{85}$$

and:

$$\mathcal{L}_{R\text{-}FA} = \lambda_{region}\mathcal{L}_{region} + \lambda_{level}\mathcal{L}_{level} + \lambda_{weight}\mathcal{L}_{weight} + \lambda_{alignment}\mathcal{L}_{alignment}. \tag{86}$$

Hence:

$$\mathcal{L}_{factor} = \mathcal{L}_{S\text{-}DF} + \mathcal{L}_{R\text{-}FA}. \tag{87}$$

When the rendering objective of the Gaussian representation is included, the complete training objective takes the form:

$$\mathcal{L} = \mathcal{L}_{render} + \mathcal{L}_{factor}. \tag{88}$$

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

Through the joint optimization, the static-dynamic branch is encouraged to separate persistent structure from time-varying information while retaining dynamic details across temporal and spatial scales. Meanwhile, the rigid-flexible branch constrains the evolution and coordination of regional motion according to their different motion properties. The resulting interaction representations jointly condition the Gaussian state transition and consequently provide a factorized representation for coherent temporal evolution.

C MORE DETAILS ON EXPERIMENTS

C.1 USER STUDY

We conducted a user study to evaluate the subjective quality of the generated videos. A total of 100 participants were invited to watch multiple sets of talking portrait videos. Participants rated the results from four aspects: image quality, video smoothness, detail fidelity, and identity preservation, using a 1–5 scale, where higher scores indicate 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 artifacts such as blurring, distortions, or structural abnormalities. Overall illumination and appearance consistency are also considered to assess the quality of the generated results at the static visual level.

Video Smoothness: This criterion focuses on temporal consistency and stability, 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.

Detail Fidelity: This criterion primarily evaluates the reconstruction of fine-grained facial structures and textures, including the accuracy and symmetry of facial features, the clarity and completeness of key local regions, and the realism of skin textures and subtle details such as pores and fine wrinkles. It also considers whether illumination and shadows appear physically plausible, whether the boundaries between the face, background, and hair are rendered naturally without artifacts, and whether distinctive identity-related features are consistently preserved.

Identity Preservation: This criterion measures whether the generated results effectively preserve the identity characteristics of the target subject, including facial structure, facial features, and overall identity consistency across different frames. Participants also consider whether identity drift or

identity confusion occurs, thereby assessing the model's ability to preserve identity information during dynamic generation.

The subjective evaluation results in Figure 7 show that our method achieves higher scores across all four dimensions. For image quality, benefiting from the static-dynamic factor fusion (S-DF) strategy, the model more effectively captures fine-grained structural and texture information, producing clearer and more detailed facial regions while reducing blurring and artifacts. For video smoothness, the structured modeling of temporal dynamics enables more continuous transitions between adjacent frames and reduces flickering and abrupt changes. The rigid-flexible factor alignment (R-FA) strategy further coordinates motion relationships across different regions, making lip movements, facial expressions, and head motions more natural and better synchronized with the input audio. For identity preservation, the continuous constraint imposed by static structural information during dynamic generation enables the model to maintain stable identity characteristics under complex motions, effectively mitigating identity drift and distortion.

Overall, S-DF establishes interactions between static and dynamic information at the feature level to enhance fine-grained detail representation, while R-FA aligns rigid and flexible regions at the motion level to improve overall motion coordination. These results demonstrate the effectiveness of the proposed framework in improving generation quality, motion coordination, and identity preservation.

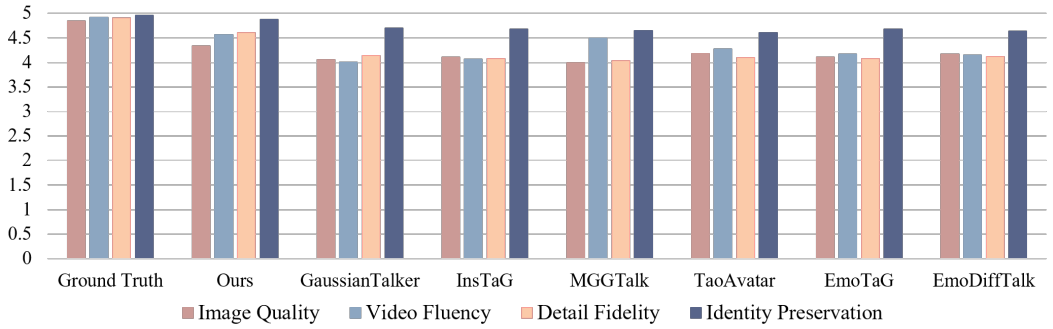

Figure 7: User study.

C.2 Model Performance

To analyze the optimization behavior of different methods, we recorded and compared the loss variations during training, as shown in Figure 8. All models were trained under the same hardware conditions to ensure a fair comparison. The initial learning rate was set to 1×10^{-3}, and the batch size was set to 16 frames, with appropriate adjustments according to the available GPU memory.

As shown by the curves in Figure 8, TaoAvatar and GaussianTalker exhibit relatively pronounced loss reductions during training and stabilize earlier, indicating that their optimization processes can reach a relatively stable state efficiently. EmoTaG, InsTaG, and EmoDiffTalk show more gradual changes in their loss curves and become increasingly stable in the later stages of training, although the subsequent reduction remains relatively limited. In contrast, MGGTalk shows a slower decrease with noticeable fluctuations, indicating a relatively less stable optimization process.

Our method achieves a more pronounced loss reduction during the early training stage and maintains a relatively stable trend throughout the subsequent iterations. This observation suggests that the motion interaction representation provides an effective organization of dynamic features, facilitating the learning of relationships between different motion factors. Meanwhile, the relatively small fluctuations in the loss curve indicate the favorable optimization stability of the proposed framework.

C.3 Parameter Settings

We conduct a parameter sensitivity study to investigate the effects of the temperature parameter τ in the fusion loss and the hierarchical weight coefficient λ_l in the alignment loss. By varying these

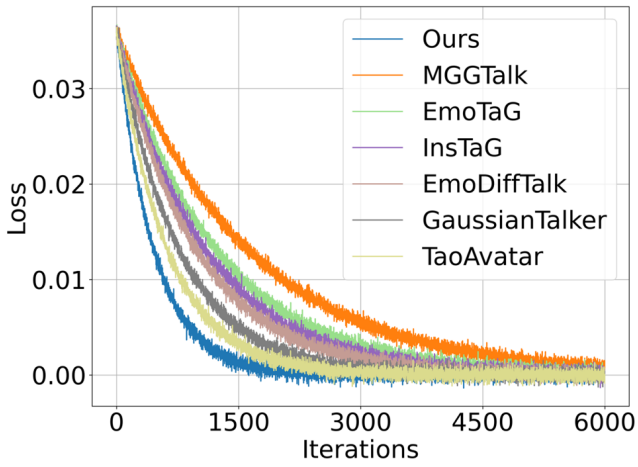

Figure 8: Model performance.

parameters, we analyze their effects on fine-grained dynamic representation and the coordination between global and local motions, and determine the final parameter settings accordingly.

Temperature Parameter τ : In the static-dynamic interaction loss, the temperature parameter τ controls the smoothness of the correlation distribution between static and dynamic features. As illustrated in Figure 9, when τ is small, the model places greater emphasis on highly correlated local matches. Although this can enhance sensitivity to fine-grained details, it may also make the model overly sensitive to local dynamics, resulting in temporal fluctuations and detail jitter. As τ increases, the correlation distribution becomes smoother, which helps maintain more stable overall motion. However, an excessively large τ weakens the distinction among different dynamic features, causing fine-grained time-varying details to become overly smoothed.

Our experiments show that the model achieves the best performance with $\tau \in [0.05, 0.1, 0.2]$, indicating that an appropriate temperature provides a favorable balance between sensitivity to local dynamics and overall motion stability.

Hierarchical Weighting Coefficients λ_k: In the rigid-flexible factor alignment loss, the hierarchical weighting coefficients λ_k balance the contributions of different motion levels during feature alignment. Lower levels mainly capture global and rigid motion trends, whereas higher levels focus more on local and flexible dynamic variations.

To investigate the effect of different hierarchical weight allocations, we considered five weighting patterns: Uniform (U), Gradual-Increasing (GI), Sharp-Increasing (SI), Gradual-Decreasing (GD), and Sharp-Decreasing (SD). All configurations were normalized under the constraint:

$$\sum_{k=1}^{K} \lambda_k = 1. \tag{89}$$

Among the five configurations, the Gradual-Increasing (GI) distribution achieves the best performance, with:

$$\lambda_k = [0.067, 0.133, 0.200, 0.267, 0.333]. \tag{90}$$

Compared with the uniform distribution $[0.200, 0.200, 0.200, 0.200, 0.200]$, the gradual-increasing scheme appropriately strengthens the contribution of higher-level local dynamics. Compared with the sharp-increasing distribution $[0.018, 0.055, 0.127, 0.218, 0.382]$, it avoids placing excessive emphasis on fine-grained local variations. In contrast, the gradual-decreasing $[0.333, 0.267, 0.200, 0.133, 0.067]$ and sharp-decreasing $[0.455, 0.291, 0.164, 0.073, 0.018]$ configurations assign greater importance to lower-level global motion, which may suppress local dynamic variations.

These results in Figure 10 suggest that the hierarchical weights should not be excessively biased toward either global or local motion. Instead, a moderately increasing weighting scheme provides a

better balance between global motion trends and fine-grained dynamic variations, such as lip movements and facial expressions.

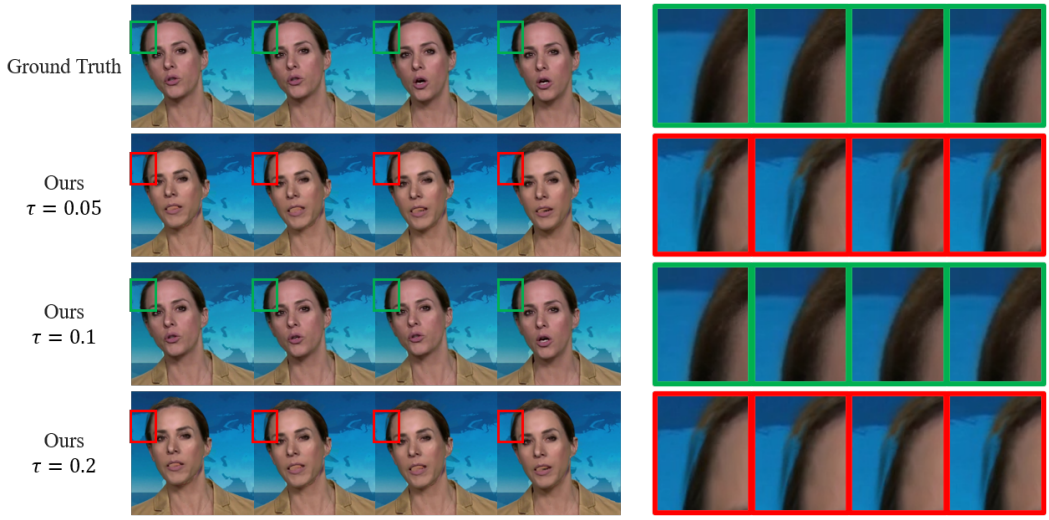

Figure 9: Comparison of parameters of τ. When $\tau = 0.1$, the visual performance is optimal, as shown in the green box.



Figure 10: Comparison of parameters of λ_k. When $\lambda_k = [0.067, 0.133, 0.200, 0.267, 0.333]$, the visual performance is optimal, as shown in the green box.

C.4 LIMITS

Our method can effectively model interactions among different motion factors within a portrait and, as discussed in the Discussion section, supports explicit user control over factors such as gestures,

pose, and emotion. However, the current framework is mainly limited to motion factors within a single portrait and has not yet been extended to more complex external interaction scenarios.

First, regarding human–environment interaction, the current method mainly focuses on the dynamic changes of the person and does not jointly model dynamic objects in the surrounding environment. As shown in Figure 11, the clock hands in the background move over time in the Ground Truth, whereas both our method and the representative comparison methods fail to recover this environmental motion, leaving the clock hands largely stationary. This indicates that although our method can predict the state evolution of the person, it cannot yet predict the evolution of other dynamic factors in the scene.

Second, regarding human–human interaction, the current framework is primarily designed for single-person portrait generation and does not yet support multi-person talking or dynamic interactions among multiple subjects. Consequently, when multiple people participate in a conversation, their pose changes, facial expressions, and audio-driven dynamics are difficult to model and coordinate in a unified manner.

Overall, the main limitation of the current method is that its state evolution mechanism is currently confined to motion factors within a single portrait. In the future, this mechanism could be extended to interactions between people and their environment, as well as between multiple people. This would enable the model to jointly predict the evolution of a person's internal state and external dynamic factors, providing a potential direction toward a more comprehensive interactive digital human generation system.

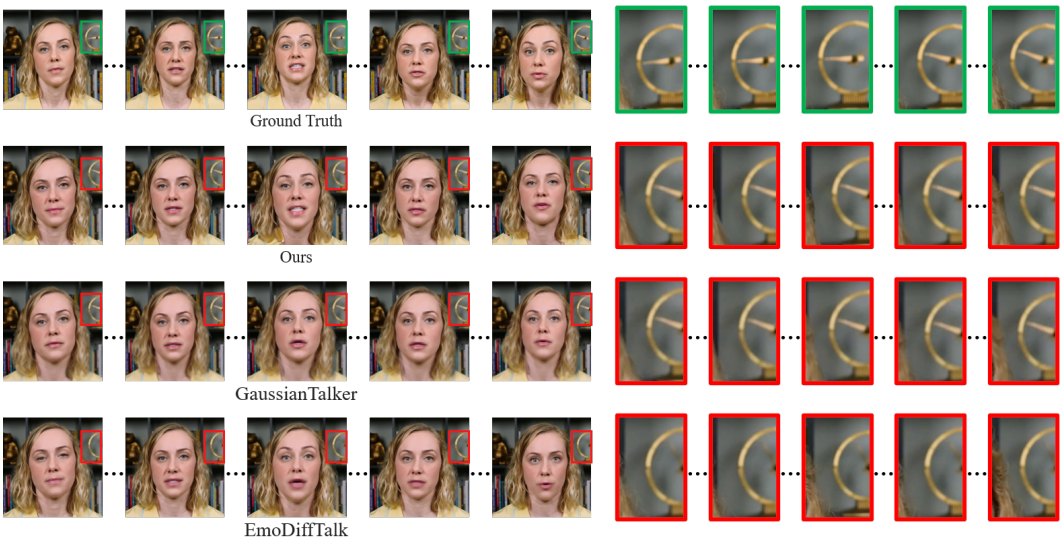

Figure 11: Limitations of this method. Our method fails to capture interactions between the portrait and the environment.