DGTalker: Disentangled Generative Latent Space Learning for Audio-Driven Gaussian Talking Heads

Xiaoxi Liang[1], Yanbo Fan[2,*], Qiya Yang[1], Xuan Wang[3], Wei Gao[1], Ge Li[1,*]
[1]School of Electronic and Computer Engineering, Peking University,
[2]Nanjing University, [3]Ant Group
shawks0419.github.io/dgtalker

Abstract

In this work, we investigate the generation of high-fidelity, audio-driven 3D Gaussian talking heads from monocular videos. We present DGTalker, an innovative framework designed for real-time, high-fidelity, and 3D-aware talking head synthesis. By leveraging Gaussian generative priors and treating the task as a latent space navigation problem, our method effectively alleviates the lack of 3D information and the low-quality detail reconstruction caused by the absence of structure priors in monocular videos, which is a longstanding challenge in existing 3DGS-based approaches. To ensure precise lip synchronization and nuanced expression control, we propose a disentangled latent space navigation method that independently models lip motion and talking expressions. Additionally, we introduce an effective masked cross-view supervision strategy to enable robust learning within the disentangled framework. We conduct extensive experiments and demonstrate that DGTalker surpasses current state-of-the-art methods in visual quality, motion accuracy, and controllability.

1. Introduction

Building an audio-driven 3D head avatar from a monocular video is a valuable research topic and has many different applications, such as gaming, filmmaking, holographic communication, etc. It often has high requirements, including the visual quality from different viewpoints, the rendering speed, and the synchronization between audio and video.

Previous works [2, 8, 15, 16, 23, 24, 42, 44] build the 3D head avatar with neural radiance field (NeRF) [28], utilizing its high-fidelity view synthesis capabilities. However, despite various improvements, the NeRF-based methods are still limited by relatively low rendering efficiency, making it difficult for them to be applied to tasks with real-time rendering requirements. Recently, a few works [6, 11, 25, 43] have started to use 3D Gaussian Splatting (3DGS) [21] for head avatar modeling. While 3DGS provides efficient and flexible explicit 3D representation, it typically requires a wide range of viewpoints to learn 3D geometry. These viewpoints, which are difficult to obtain from monocular videos, pose significant practical challenges. An insightful solution is to leverage the generative prior learned from large-scale data to construct a more expressive representation with viewpoint generalization capability, as explored in NeRF-based methods [2, 23]. However, a key challenge arises from the fact that the latent space of generative methods is inherently entangled with multiple identities and global facial expression variations, whereas the audio primarily correlates with local lip movements. This mismatch hampers effective model learning when the full facial image is used as supervision, as in previous methods, and results in suboptimal audio-visual synchronization.

To this end, we introduce DGTalker, a novel framework leveraging Gaussian generative priors [18, 22] for modeling the audio-driven 3D Gaussian head avatar. To seamlessly integrate with generative priors, we reformulate the task as a latent space navigation problem. We propose a novel *dis-*

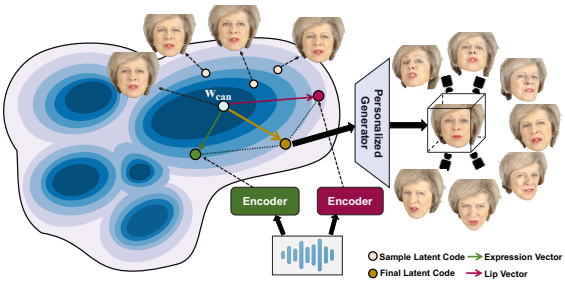

Figure 1. A brief illustration of DGTalker. We reformulate the problem as disentangled latent space navigation. We adopt dual encoders to extract a lip vector and an expression vector from the audio input. These two vectors, together with the canonical code, are then linearly combined to form the final latent code, which is then sent to the personalized generator for synthesis.

*Corresponding authors

entangled navigation method, leveraging the linear interpolation property of the latent space [7, 18, 19, 22], which enables continuous and smooth synthesis transitions between two latent codes, thereby alleviating the unequal mapping between audio and lip/expression movements. For a certain identity, we propose to learn an *anchor vector* with a canonical expression, along with an *expression vector*, and a *lip vector* which independently model the variations of expressions (upper-face motions) and lip motions (lower-face motions). In particular, we adopt a dual-encoder architecture and decompose the lip and expression vectors into two sets of learnable, orthogonal blendshape bases, to facilitate more effective training. For audio-driven talking head synthesis, we first map the audio signals to coefficients corresponding to these two orthogonal groups of blendshapes. These components are then fused with a global canonical latent vector via linear interpolation to obtain the final latent vector, which is subsequently fed into the generator to synthesize the 3D talking head avatar. A brief overview of the principle of DGTalker is illustrated in Fig. 1.

To effectively learn disentangled expression and lip components, we further introduce a novel *masked cross-view supervision strategy*. Specifically, during training, we prompt the generator to synthesize a non-existent 3DGS head by combining the lip latent code from one audio and the expression latent code from another. Then, we render the 3DGS head under each audio-correlated viewpoint, and apply region-specific supervision focusing only on the upper face or the lower face, respectively. Based on the above designs, DGTalker enables high-fidelity rendering across a wide range of viewpoints, achieves superior motion synchronization accuracy, and provides each component with well-defined semantic meaning, which leads to stronger controllability. We conduct extensive experiments and demonstrate that DGTalker outperforms several state-of-the-art methods in both quantitative and qualitative evaluations. Additionally, the ablation studies further validate the effectiveness of the proposed tailored designs. The main contributions of our paper are summarized as follows:

- We propose to leverage Gaussian generative priors and formulate the audio-driven Gaussian talking head reconstruction as a latent space navigation task.
- We propose a novel disentangled framework that decomposes the high-dimensional latent vector into a global canonical vector, an expression vector, and a lip vector.
- We propose an effective masked cross-view supervision strategy that enables the learning of disentangled expression and lip components.
- Experiments demonstrate that the proposed DGTalker outperforms state-of-the-art methods and exhibits superior disentangled controllability.

2. Related Work

3D Talking Head Synthesis. Reconstructing and animating talking heads by arbitrary audio is an active research topic [2, 16, 26, 32, 39]. In the early stages, NeRF [28] was introduced as a 3D representation of the talking head, enabling photorealistic rendering and personalized talking style through person-specific training. Earlier NeRF-based works [16, 26, 35, 42] suffer from the expensive cost of vanilla NeRF. Although RAD-NeRF [15] and ER-NeRF [24] have improved efficiency with grid-based NeRF [29], real-time rendering of 3D talking head remains challenging. Recently, 3DGS [21] have been explored for this field as a novel 3D representation. GaussianTalker [11] improves the responsiveness of Gaussian points to audio by integrating cross-attention mechanisms at the cost of slightly reduced inference speed. TalkingGaussian [25] improves motion accuracy via a face-mouth decomposition module but introduces novel views with more undesired artifacts. Concurrently, EmoTalkingGaussian [6] achieves diverse emotional talking heads utilizing a curated speech audio dataset, still leaving the issue of novel view reconstruction challenges unexplored. GaussianSpeech [1] achieves finer-grained modeling quality by utilizing their proposed multi-view talking head dataset. However, it is limited to only a few identities, which restricts its applicability.

GANs Editing and 3D-aware GANs. In response to the recent emergence of GANs, a wide range of studies have explored various approaches for tasks such as editing, super-resolution, and image inpainting. For facial editing, InterfaceGAN [36] leverages the smooth variation of output images through interpolation in the latent space, enabling controllable editing by identifying hyperplanes corresponding to specific attributes. Subsequent works further explore identity-specific editing [30, 45] and temporally consistent editing for videos [14]. However, these methods typically focus on coarse attribute manipulations (e.g., age or smile intensity) and struggle to achieve precise fine-grained facial dynamics from audio signals. Meanwhile, 3D-aware GANs, which incorporate 3D representations, have emerged with similar latent space properties. EG3D [7] first introduced NeRF [28] as a 3D representation and applied a super-resolution network to the rendered features, resulting in 3D inconsistencies. Although subsequent works [10, 31, 38, 41] proposed various methods to improve 3D consistency, they are still constrained by the slow inference speed inherent to NeRF representation. Recently, GSGAN [18] and GGhead [22] improved rendering speed and ensured 3D consistency by utilizing 3DGS as the generator's output, enabling direct rendering without the need for a neural network during rendering, which presents greater potential for real-time, speech-driven 3D talking head synthesis.

Talking Head With Generative Priors. Recently, two

methods have emerged that leverage generative priors for synthesizing talking heads, which are most closely related to our work. HFA-GP [2] learned a low-dimensional latent subspace spanned by audio in the whole latent space of a 3D-aware GAN. They neither disentangle identity-specific region from the universal latent space nor address the inherent mismatch between audio and the latent representation, leading to suboptimal audio-lip synchronization. Subsequently, Talk3D [23] improved the audio-lip synchronization by introducing an audio-guided attention U-Net to perform motion compensation on the 3D representations generated by 3D-aware GANs. However, it not only increased the computational burden but also failed to ensure multi-view consistency, as the motion compensation is learned independently from the generator. How to better leverage generative priors for achieving audio-lip synchronization across diverse viewpoints remains an open research question.

3. Method

In this section, we first briefly review the Gaussian generative priors in Sec. 3.1. Then, we explain the proposed disentangled navigation design, along with the corresponding two sets of blendshapes and dual encoders in Sec. 3.2 and detail our training strategy, which includes the personalized generator, canonical code learning, and the Masked Cross-view Supervision scheme, in Sec. 3.3. Finally, we demonstrate our superior controllability in Sec. 3.4.

3.1. Gaussian Generative Priors

Our approach builds upon 3D-aware Gaussian GANs, which enable the generation of real-time, high-fidelity, and multi-view consistent Gaussian heads from randomly sampled latent vectors. The state-of-the-art method, GGhead [22] employs 3DGS primitives as the generator's output to alleviate the computational bottleneck and 3D inconsistencies introduced by the super-resolution module. In particular, given a latent vector $w \in \mathbb{R}^{512}$ and a standard front-view pose π_0 as pose-conditioning, the GGhead generator $\mathcal{G}$ predict a sufficient and compact 3DGS head representation. An image I is then rendered using the tile-based rasterizer $\mathcal{R}$ under the desired camera viewpoint π,

$$I = \mathcal{R}(\mathcal{G}(w, \pi_0), \pi). \tag{1}$$

Given a sequence of N video frames $\mathcal{V} = \{I_n\}_{n=1}^{N}$ of a specific identity, each frame I_n is associated with an audio feature f_n and a camera parameter π_n. As π_n is the ground truth, our goal is to seek a function $\mathcal{F}$ conditioned on the audio feature f_n that predicts the latent code w_n. To simplify notation, we omit the subscript n in the subsequent section.

3.2. Disentangled Navigation Design

An overview of DGTalker is shown in Fig. 2. Leveraging the smooth variation of the output through interpolation in the latent space, we define an anchor w_{can} and two orthogonal vectors Δw_{exp}, Δw_{lip} in the latent space. These components independently encode a global specific identity with a canonical expression, dynamic talking expression variations (upper-face motion), and dynamic lip motion variations (lower-face motion), respectively. This can be written as,

$$w = w_{\text{can}} + \Delta w_{\text{exp}} + \Delta w_{\text{lip}}, \tag{2}$$

where Δw_{exp} and Δw_{lip} should be controlled by audio, determining how far to move from w_{can}. To further facilitate effective training, we design two sets of learnable orthogonal latent blendshapes, $B_{\text{exp}}, B_{\text{lip}}$ instead of vectors. Specifically, for B_{lip} with k_l vectors as $B_{\text{lip}} = [b^l_1, \cdots, b^l_{k_l}] \in \mathbb{R}^{k_l \times 512}$ in the latent space, where each vector represents distinct lip motion variations, the final corresponding lip vector Δw_{lip} is obtained by linear blending. The same process applies to $B_{\text{exp}} = [b^e_1, \cdots, b^e_{k_e}] \in \mathbb{R}^{k_e \times 512}$ with k_e vectors, and final formulation can be written as,

$$w = w_{\text{can}} + \sum_{k=1}^{k_e} \alpha^e_k b^e_k + \sum_{k=1}^{k_l} \alpha^l_k b^l_k, \tag{3}$$

where α^*_k denotes the coefficient corresponding to the basis vector b^*_k with $* \in \{e, l\}$, indicating the intensity variations of expression and lip blendshapes, respectively. We denote the coefficient vector as $\boldsymbol{\alpha}_* = [\alpha^*_1, \cdots, \alpha^*_{k_*}]^\top \in \mathbb{R}^{1 \times k_*}$ for simplicity. We then utilize dual-encoders to regress two coefficient vectors from audio features f. To better control the audio-independent action of eye blinking, we follow previous works [24, 25] using AU45 [34] to describe the degree of eye closure, denoted as $\mathcal{E}$. This is further incorporated into the expression encoder, leading to the final formulation $\boldsymbol{\alpha}_e = E_e(f, \mathcal{E}), \boldsymbol{\alpha}_l = E_l(f)$.

3.3. Disentangled Learning

Personalized Generator and w_{can} Learning. Directly achieving high-fidelity reconstruction from the large-scale pretrained generator is challenging due to the distribution discrepancy between real images and generative priors [14, 23, 33]. Hence, we first learn a personalized generator by designing a variant of pivotal fine-tuning [33]. Specifically, for each training image, we decompose the corresponding latent vector into a global w_{can} and a frame-specific w_{frame}. To obtain a neutral w_{can} with more accurate geometry, we jointly optimize these latent vectors under the ground-truth camera parameters. Then, we fix the latent vectors and fine-tune the generator using the same objectives as in PTI. Finally, we discard the w_{frame}, and retain and freeze the personalized generator along with the w_{can} for subsequent use.

Masked Cross-view Supervision Strategy. To ensure that our disentangled components can be effectively learned, we propose a novel Masked Cross-View Supervision(MCS)

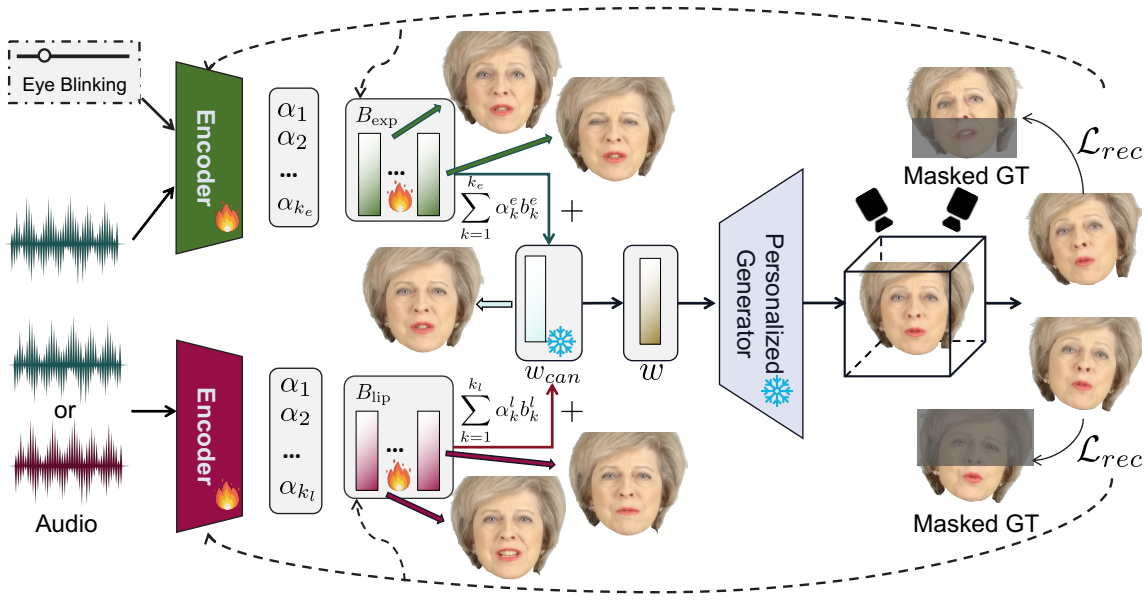

Figure 2. Overall Framework of DGTalker. We design a disentangled navigation framework consisting of an anchor w_{can}, which encodes a global canonical expression for a specific identity, and two sets of learnable, orthogonal blendshapes $B_{\text{exp}}, B_{\text{lip}}$ containing k_e and k_l vectors, respectively. Each vector corresponds to a disentangled variation in upper/lower face expressions. The input audio is used to regress the coefficients of these blendshapes. To ensure effective learning, we randomly feed the encoder with different audio inputs and render the output images from two viewpoints. The corresponding masked ground-truth (GT) images are then used for supervision.

training strategy. Specifically, at each training iteration, with probability p, we randomly sample a frame pair, denoted as I_1, I_2, along with different viewpoints π_1, π_2, audio features f_1, f_2 and eye actions $\mathcal{E}_1, \mathcal{E}_2$. The non-existent w_{12} in the training set is derived as,

$$w_{12} = w_{\text{can}} + E_e(f_1, \mathcal{E}_1)B_{\text{exp}} + E_l(f_2)B_{\text{lip}}. \quad (4)$$

We then render the generated 3DGS head from w_{12} twice, under viewpoints π_1 and π_2, respectively, and supervise the two renderings using only the upper-face region from I_1 and only the lower-face region from I_2. To avoid distributional shift between the learned model and real face data, we disable MCS with a probability of $1 - p$. The Training objectives $\mathcal{L}_{rec}$ are composited by pixel-wise $\mathcal{L}_2$ loss and perceptual loss $\mathcal{L}_{\text{LPIPS}}$.

3.4. Controllability

Thanks to our carefully designed disentangled latent space navigation method and the masked cross-view supervision training strategy, the learned components possess well-defined semantic meanings, which will be discussed in detail in Sec. 4. This enables our DGTalker to generate diverse talking expressions for the same speech content, offering superior controllability over other methods. Fig. 3 shows an example where the same utterance is spoken by the same identity with different talking expressions. In the "May" example, we present normal speech alongside speech with closed eyes. In the "Macron" example, we demonstrate normal speech contrasted with speech with a shocked expression.

(a) The first row shows normal speech, while the second row features the same speech with closed eyes.

(b) The first row shows normal speech, while the second row shows shocked speech for the same utterance.

Figure 3. Our DGTalker shows excellent controllability on two different identities.

4. Experiment

In this section, we first introduce three experimental settings. Then, we show quantitative and qualitative evaluation for these experimental settings. Finally, we perform analysis and ablation study to analyze the key elements of our approach.

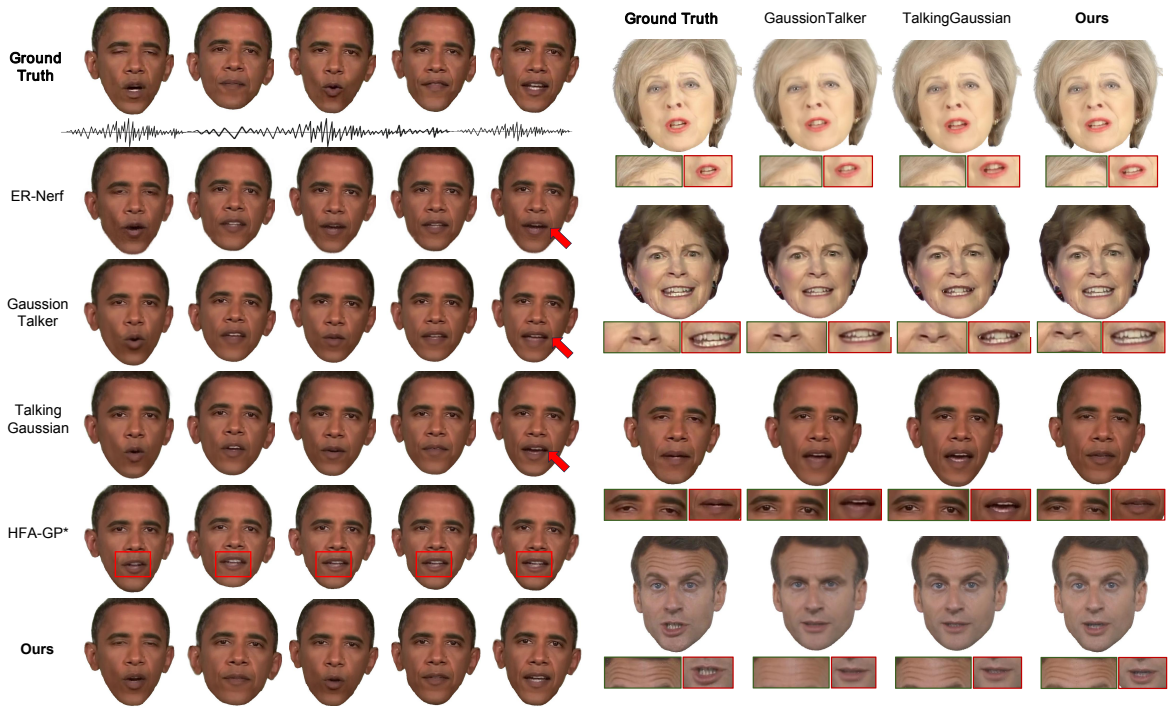

Figure 4. Qualitative comparison of reconstruction quality and visual-audio synchronization under the *self-reconstruction setting*. We remove the background to allow for more intuitive and direct comparisons. Our method generates more precise and complete details compared to recent state-of-the-art approaches [2, 11, 24, 25]. Please **zoom in for details**.

4.1. Experimental Settings

Dataset and Pre-processing. We evaluate our method on 4 publicly available videos used in prior works [16, 24, 42], as well as 4 videos from the HDTF dataset [47], resulting in a total of 8 clips. Each video contains approximately $6,500$ frames at 25 FPS, with a balanced gender distribution. Each video is cropped and resized to 512×512. For each frame of the video, we follow HFA-GP [2] to obtain camera parameters, ensuring consistency with the generator scale. Finally, we employ the off-the-shelf background matting network MODNet [20] to remove the background following GGhead [22]. We split each video into training and testing sets with a ratio of 10:1, following the same protocol as previous works.

Implementation Details. For the personalized generator and canonical code training, both the optimization and fine-tuning stages are performed with a batch size of 8 for $15k$ iterations. For disentangled learning stage, we empirically set p to 50%, the number of blendshapes k_l and k_s to 20, and train for $60k$ iterations using the batch size 16. We simply use facial landmarks [5] to divide the face into upper and lower parts and our dual-encoders share the same structure as previous works [24, 25]. We enforce the orthogonality of the blendshapes via QR decomposition. All training stages are optimized using the AdamW [27] optimizer with a learning rate of 3e-4.

Comparison Baselines. We evaluate our proposed DGTalker against recent 3DGS-based approaches, GaussianTalker [11] and TalkingGaussian [25]. To further demonstrate that the success stems from our design rather than the 3D-aware GAN, we also compare a method with a similar concept, HFA-GP [2], and replace the EG3D [7] used in the original paper with GGHead [22]. We refer to the modified version as HFA-GP*. Moreover, we included a classic NeRF-based approach, ER-NeRF [24], to provide a more comprehensive set of comparisons.

4.2. Quantitative Evaluation

Comparison settings. To evaluate the 3D-aware reconstruction quality and lip-audio synchronization ability, our quantitative comparison contains three settings:

1) The *self-reconstruction setting*, where each of the eight videos is split into training and test sets. The audio, eye blink, and pose sequences from the unseen test set are used to reconstruct the talking head in a self-driven manner, following the same protocol as previous methods [11, 24, 25].

2) The *novel-view self-reconstruction setting*, where the reconstruction and motion quality are evaluated under a wider range of viewpoints, while the audio and eye blink

Figure 5. We present two sets of visualizations, each rendered from five viewpoints—including a front view and four views gradually shifted in yaw and pitch—under two different settings, respectively. The left-side images are generated under the *novel-view self-reconstruction setting*, where the corresponding test set audio is used. The right-side images are produced under the *generalized 3D-aware audio-lip synchronization setting*, using an unseen audio clip from SynObama [37] to demonstrate the audio generalization capability. The topmost images show the ground-truth lip shapes.

sequences are still derived from the corresponding test set. Specifically, we selected two extreme viewpoints with yaw and pitch angles $(\pm 30°, \pm 30°)$ and two shifted viewpoints $(\pm 20°, \pm 20°)$ from the canonical viewpoint. Additionally, we incorporated a spiral camera trajectory to eliminate suspicion of cherry-picking.

3) The *generalized 3D-aware audio-lip synchronization setting*, where we follow previous work using two unseen audio tracks, called test audio A and B, from SynObama [37] to drive the models. To further emphasize the overall accuracy of audio-lip synchronization across multi-views, we render the evaluation videos using a spiral camera trajectory with a fixed eye blink signal.

For a fair comparison, we retrained all methods using our camera parameters and rendered all heads onto a white background and masked out the torso region to highlight the performance of the algorithms.

Metrics and Measurements. In the aspect of image quality, we employ PSNR for the overall quality, LPIPS [46] for high-frequency details, and SSIM [40] to evaluate face structure. For dynamic motions, we utilize the landmark distance (LMD) [9] and the confidence score (Sync-C) and error distance (Sync-E) of SyncNet [12] for lip synchronization. We also compute the Action Unit Error (AUE) by estimating the action units [34] of the videos using OpenFace [3, 4]. Additionally, we record the inference FPS to evaluate real-time performance.

For the second setting, due to the absence of ground-truth images, we employ Fréchet Inception Distance (FID) [17] and Identity Similarity (IDSIM) derived from ArcFace [13] to evaluate reconstruction quality and use AUE extracted from GT video, non-comparison-based Sync-C and Sync-E to evaluate motion quality. For the last setting, we only employ Sync-C and Sync-E following [25].

Evaluation Results. We report the results of the three settings in Tab.1, Tab.2 and Tab.3, respectively. In the *self-reconstruction setting*, our method achieves the best quality in all image quality metrics. While TalkingGaussian achieves the highest synchronization scores thanks to its well-designed face and inside-mouth branches, its overall performance is hindered by the insufficient number of Gaussians. This limitation stems from the vanilla 3DGS reconstruction system, which, in the absence of sufficient viewpoints to provide facial structure priors, ultimately leads to insufficient Gaussians and lower image quality. On the other hand, HFA-GP* achieves higher image quality, it overlooks the latent-audio unequal mapping problem, resulting in lower synchronization scores.

In the *novel-view self-reconstruction setting*, we report the average results over four specific viewpoints with yaw $(\pm 30°, \pm 20°)$ and pitch $(\pm 30°, \pm 20°)$, along with a spiral camera trajectory. While most methods achieve comparable performance on frontal view rendering, all methods without generative priors experience a significant performance drop

Methods	PSNR ↑	LPIPS ↓	SSIM ↑	FID ↓	LMD ↓	Sync-E ↓	Sync-C ↑	AUE ↓	FPS ↑
Ground Truth	N/A	0	1	0	0	6.859	8.468	0	N/A
ER-NeRF [24]	24.021	0.155	0.810	75.316	4.413	10.425	4.299	1.465	27
GaussianTalker [11]	26.727	0.127	0.854	24.157	4.386	9.915	5.207	1.432	83
TalkingGaussian [25]	26.249	0.094	0.847	21.737	**3.944**	**8.460**	**6.629**	1.254	**94**
HFA-GP* [2]	27.417	0.104	0.838	31.215	4.558	11.483	2.469	1.842	75
Ours	**28.943**	**0.065**	**0.863**	**15.149**	3.997	8.936	6.295	**1.209**	71

Table 1. The quantitative results of the *self-reconstruction setting*. The best and second-best methods are in **bold** and underline, respectively. Ours achieves the best image quality and competitive motion quality.

Methods	FID ↓	IDSIM ↑	AUE ↓	Sync-E ↓	Sync-C ↑
Ground Truth	0	1	0	6.859	8.468
ER-NeRF [24]	228.740	0.306	2.753	11.141	2.887
GaussianTalker [11]	138.332	0.337	2.601	10.785	3.773
TalkingGaussian [25]	137.914	0.363	2.621	9.624	5.198
HFA-GP* [2]	99.601	0.373	2.745	12.455	1.627
Ours	**80.011**	**0.436**	**2.525**	**9.565**	**5.255**

Table 2. To provide a comprehensive evaluation, we test four specific viewpoints with yaw ($\pm 30^\circ, \pm 20^\circ$) and pitch ($\pm 30^\circ, \pm 20^\circ$), along with a spiral camera trajectory, resulting in five viewpoint configurations in total. We report the average performance across these views. The best and second-best results are highlighted in **bold** and underline, respectively. Our method achieves the highest scores in both image quality and motion accuracy.

Method	Test Audio A		Test Audio B	
	Sync-C ↑	Sync-E ↓	Sync-C ↑	Sync-E ↓
Ground Truth	8.167	6.808	8.080	7.182
ER-NeRF [24]	2.267	11.669	2.458	11.369
GaussianTalker [11]	3.242	10.903	1.557	12.476
TalkingGaussian [25]	3.922	10.528	3.999	10.202
HFA-GP* [2]	1.098	13.172	0.976	12.966
Ours	**3.947**	**10.433**	**4.069**	**10.106**

Table 3. The quantitative results of the *generalized 3D-aware audio-lip synchronization setting*. The best and second-best methods are in **bold** and underline, respectively. For an unseen audio clip, our method achieves the highest score, indicating strong audio generalization capability and ensuring motion accuracy across multiple viewpoints.

as the viewing angle varies. Notably, although HFA-GP* demonstrates promising results on novel-view reconstruction metrics (e.g., FID and IDSIM), it suffers from inferior motion quality (SYNC-C) compared to other methods. In contrast, our method achieves state-of-the-art performance across all metrics. This underscores the effectiveness of our approach, which benefits from our disentangled design and training.

In the *generalized 3D-aware audio-lip synchronization setting*, for an unseen audio clip, we perform a holistic evaluation of generalization by rendering with a spiral camera trajectory. This setup not only emphasizes motion accuracy but also highlights 3D consistency. Our method achieves the highest scores, demonstrating not only strong generalization to novel audio but also robust 3D consistency across a wide range of viewpoints.

4.3. Qualitative Evaluation

We first present key frames from a reconstructed sequence under the *self-reconstruction setting*, along with detailed views of four subjects used in prior work in Fig. 4. While TalkingGaussian achieves accurate lip motion by explicitly decoupling the facial region and the inside-mouth components, the results on the "Shaheen" example on the right shows hole-like artifacts when the test set includes large motion which is not covered in the training set. Besides, both GaussianTalker and TalkingGaussian still face challenges in capturing fine-grained details due to insufficient Gaussians, steming from the vanilla 3DGS reconstruction method. As for HFA-GP* which also utilize generative priors, we observe that the "Obama" sequence on the left clearly demonstrates that HFA-GP* exhibits limited mouth movement. This is primarily due to its failure to effectively correlate audio with the latent space. In addition, despite employing the same backbone, HFA-GP* still exhibits poor visual quality due to the lack of consideration to disentangle identity from the latent space. In contrast, our approach achieves superior visual quality and motion synchronization performance.

Fig. 5 illustrates examples of the *novel-view self-reconstruction setting* and the *generalized 3D-aware audio-lip synchronization setting*. We present five viewpoint configurations for qualitative evaluation: one frontal view and four shifted views. All previous 3DGS-based methods suffer from significant performance degradation when rendering from camera angles far from the canonical front view, often exhibiting inconsistent geometry and color artifacts. Although HFA-GP* demonstrates the capability to capture correct head geometry, it struggles to find the optimal solution in the latent space, resulting in inaccurate motion and color. Our method achieves superior visual quality and motion accuracy across multi-views.

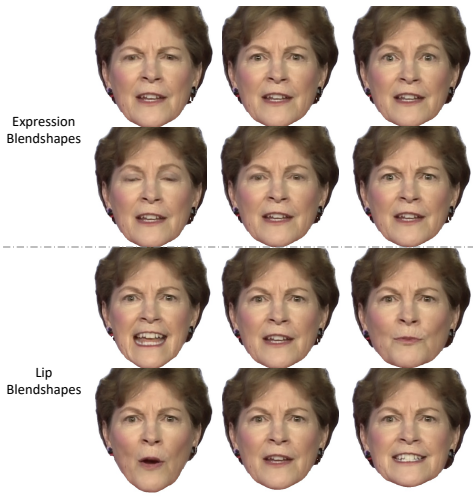

Figure 6. We selected two examples each from the expression blendshapes and the lip blendshapes. The first two rows illustrate the expression blendshapes, while the last two rows present the lip blendshapes. In each row, as the blendshape coefficient varies, the corresponding facial region undergoes noticeable changes.

4.4. Analysis and Ablation Study

Method	PSNR ↑	LPIPS ↓	FID ↓	LMD ↓	Sync ↑
Ground Truth	N/A	0	0	0	8.468
w/o Disentangled Design	27.741	0.101	19.951	4.547	3.869
w/o Dual-Encoders	28.473	0.073	16.208	4.127	5.870
w/o Blendshapes	28.868	0.070	15.156	4.191	6.189
w/o MCS	28.559	0.072	15.742	4.551	4.547
All	**28.943**	**0.065**	**15.149**	**3.997**	**6.295**

Table 4. Ablation study of each component under the self-reconstruction setting. The complete result(All) are highlighted in **bold** for better reference.

To demonstrate the effectiveness of our contributions, we analyze the visualization results of key components and conduct an ablation study under the *self-reconstruction setting* using all eight identities.

Visualization of Disentangled Blendshapes. Fig. 6 presents the visualization of the learned blendshapes. Each blendshape carries well-defined semantic meaning. As the coefficients vary, the upper-face changes in the talking expression blendshapes, while the lower-face changes in the lip blendshapes. This demonstrates the effectiveness and disentanglement of our two sets of blendshapes.

Effectiveness of MCS. We also present the visualization results of a learned blendshape obtained without employing MCS in Fig.7. In the first row, without MCS, the learned lip blendshape entangles both the upper-face expression and lip motion. In contrast, in the second row, the lip blendshape captures only the variations specific to the lips.

Ablation Study. We also conducted quantitative ablation studies, as presented in Tab. 4. First, we conducted experiments without the disentangled design, employing a single encoder, a canonical code, and a unified blendshape set without MCS to demonstrate the importance of latent space disentanglement. Next, we removed the dual-encoder architecture and replaced it with a single encoder that simultaneously regresses both expression and lip blendshape coefficients with MCS. Subsequently, we eliminated the blendshapes, requiring the audio to directly regress two orthogonal deformation vectors. Both variants highlight the effectiveness of promoting disentangled training. Finally, we also quantitatively presented the results of training without MCS to further underscore its importance in achieving disentangled learning.

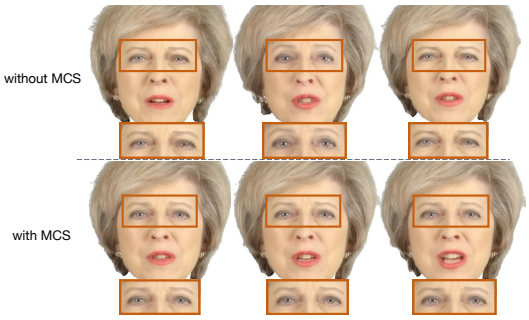

Figure 7. Visualizations of a set of lip blendshapes trained with and without MCS. In the first row, without MCS, the results exhibit an entangled representation. In contrast, the second row, with MCS applied, captures only the lip motion while preserving a consistent expression.

5. Conclusion

In this work, we introduce DGTalker, a novel framework leveraging Gaussian generative priors for real-time, high-fidelity audio-driven Gaussian talking head synthesis. To address the challenge of the unequal mapping between the latent space and audio, we propose a disentangled latent space navigation method by decomposing the high-dimensional latent vector into a canonical vector, an expression vector, and a lip vector, and employing a dual-encoder architecture aligned with corresponding blendshapes. Furthermore,we introduce an effective mask cross-view supervision mechanism to achieve the disentangled learning. We conduct extensive experiments that not only demonstrate superior performance over existing methods in both quantitative and qualitative evaluations, but also showcase remarkable controllability.

Acknowledgment

This work was supported by the National Science and Technology Major Project (2024ZD01NL00101).

References

[1] Shivangi Aneja, Artem Sevastopolsky, Tobias Kirschstein, Justus Thies, Angela Dai, and Matthias Nießner. Gaussianspeech: Audio-driven gaussian avatars, 2024. 2

[2] Yunpeng Bai, Yanbo Fan, Xuan Wang, Yong Zhang, Jingxiang Sun, Chun Yuan, and Ying Shan. High-fidelity facial avatar reconstruction from monocular video with generative priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4541–4551, 2023. 1, 2, 3, 5, 7

[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66, 2018. 6

[4] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6, 2015. 6

[5] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 5

[6] Junuk Cha, Seongro Yoon, Valeriya Strizhkova, Francois Bremond, and Seungryul Baek. Emotalkinggaussian: Continuous emotion-conditioned talking head synthesis. *arXiv preprint arXiv:2502.00654*, 2025. 1, 2

[7] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022. 2, 5

[8] Aggelina Chatziagapi, ShahRukh Athar, Abhinav Jain, Rohith MV, Vimal Bhat, and Dimitris Samaras. Lipnerf: What is the right feature space to lip-sync a nerf? In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2023. 1

[9] Lele Chen, Zhiheng Li, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII*, page 538–553, Berlin, Heidelberg, 2018. Springer-Verlag. 6

[10] Xingyu Chen, Yu Deng, and Baoyuan Wang. Mimic3d: Thriving 3d-aware gans via 3d-to-2d imitation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[11] Kyusun Cho, Joungbin Lee, Heeji Yoon, Yeobin Hong, Jaehoon Ko, Sangjun Ahn, and Seungryong Kim. Gaussiantalker: Real-time high-fidelity talking head synthesis with audio-driven 3d gaussian splatting, 2024. 1, 2, 5, 7

[12] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016. 6

[13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 6

[14] Anna Frühstück, Nikolaos Sarafianos, Yuanlu Xu, Peter Wonka, and Tony Tung. VIVE3D: Viewpoint-independent video editing using 3D-Aware GANs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[15] Lidong Guo, Xuefei Ning, Yonggan Fu, Tianchen Zhao, Zhuoliang Kang, Jincheng Yu, Yingyan Celine Lin, and Yu Wang. Rad-nerf: Ray-decoupled training of neural radiance field. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 2

[16] Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 5

[17] 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. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. 6

[18] Sangeek Hyun and Jae-Pil Heo. Gsgan: Adversarial learning for hierarchical generation of 3d gaussian splats. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 1, 2

[19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020. 2

[20] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson W.H. Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *AAAI*, 2022. 5

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

[22] Tobias Kirschstein, Simon Giebenhain, Jiapeng Tang, Markos Georgopoulos, and Matthias Nießner. GGHead: Fast and Generalizable 3D Gaussian Heads. In *SIGGRAPH Asia 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 1, 2, 3, 5

[23] Jaehoon Ko, Kyusun Cho, Joungbin Lee, Heeji Yoon, Sangmin Lee, Sangjun Ahn, and Seungryong Kim. Talk3d: High-fidelity talking portrait synthesis via personalized 3d generative prior. *arXiv preprint arXiv:2403.20153*, 2024. 1, 3

[24] Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, and Lin Gu. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7568–7578, 2023. 1, 2, 3, 5, 7

[25] 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, pages 127–145. Springer, 2024. 1, 2, 3, 5, 6, 7

[26] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. *arXiv preprint arXiv:2201.07786*, 2022. 2

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[28] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2

[29] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 2

[30] Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. *ACM Trans. Graph.*, 41(6), 2022. 2

[31] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. StyleSDF: High-Resolution 3D-Consistent Image and Geometry Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13503–13513, 2022. 2

[32] Ziqiao Peng, Yanbo Fan, Haoyu Wu, Xuan Wang, Hongyan Liu, Jun He, and Zhaoxin Fan. Dualtalk: Dual-speaker interaction for 3d talking head conversations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21055–21064, 2025. 2

[33] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021. 3

[34] Mike Seymour. Facs at 40: facial action coding system panel. In *ACM SIGGRAPH 2019 Panels*, New York, NY, USA, 2019. Association for Computing Machinery. 3, 6

[35] Shuai Shen, Wanhua Li, Zheng Zhu, Yueqi Duan, Jie Zhou, and Jiwen Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. In *European conference on computer vision*, 2022. 2

[36] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *TPAMI*, 2020. 2

[37] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Trans. Graph.*, 36(4), 2017. 6

[38] Alex Trevithick, Matthew Chan, Towaki Takikawa, Umar Iqbal, Shalini De Mello, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano and. Rendering every pixel for high-fidelity geometry in 3d gans. In *arXiv*, 2023. 2

[39] Yinuo Wang, Yanbo Fan, Xuan Wang, Guo Yu, and Fei Wang. Diffusion-based realistic listening head generation via hybrid motion modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15885–15895, 2025. 2

[40] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 6

[41] Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2195–2205, 2023. 2

[42] 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*, 2023. 1, 2, 5

[43] Hongyun Yu, Zhan Qu, Qihang Yu, Jianchuan Chen, Zhonghua Jiang, Zhiwen Chen, Shengyu Zhang, Jimin Xu, Fei Wu, Chengfei Lv, and Gang Yu. Gaussiantalker: Speaker-specific talking head synthesis via 3d gaussian splatting. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 3548–3557, New York, NY, USA, 2024. Association for Computing Machinery. 1

[44] Wangbo Yu, Yanbo Fan, Yong Zhang, Xuan Wang, Fei Yin, Yunpeng Bai, Yan-Pei Cao, Ying Shan, Yang Wu, Zhongqian Sun, et al. Nofa: Nerf-based one-shot facial avatar reconstruction. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–12, 2023. 1

[45] Libing Zeng, Lele Chen, Yi Xu, and Nima Khademi Kalantari. Mystyle++: A controllable personalized generative prior. 2023. 2

[46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[47] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 5