A More Implementation Details

We provide the detailed network structure of Flow-Audio-Attention, the principle of the Progressive Adaptive Densification method, and the details of the loss function in Sec. A.1, Sec. A.2, and Sec. A.3.

A.1 Flow Audio Attention Network

As shown in Fig. 1, we employ a cross-attention fusion module to facilitate cross-modal interaction between audio and facial features. This module consists of a multi-head attention layer and a feed-forward neural network (FNN), with residual connections incorporated to enhance optimization stability, This allows the model to focus on regions with significant facial movements and changes, such as the eyes in Fig. 2.

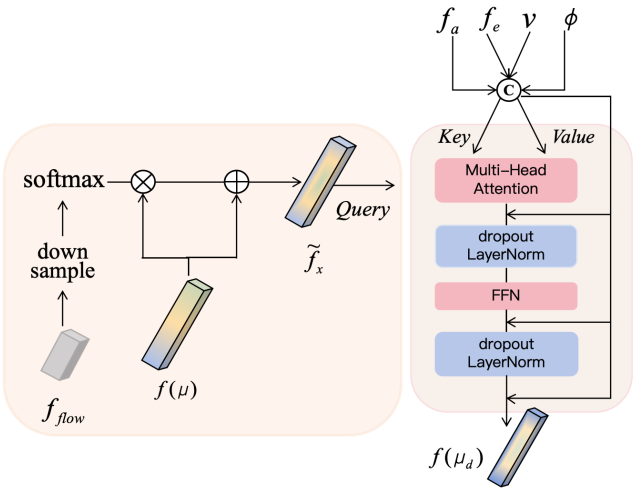

Fig. 1 The structure of Flow Audio Attention Network.

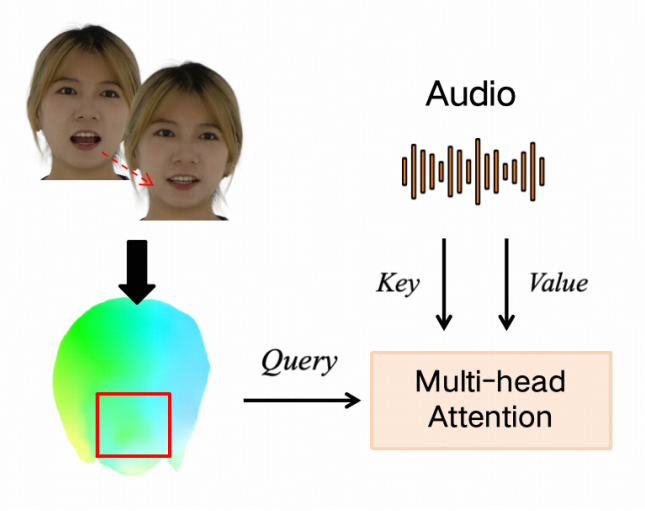

Fig. 2 Optical flow recognition of facial dynamic regions.

A.2 Progressive Adaptive Densification

As shown in the visual results in Fig. 3, τ_d densification threshold control has a significant impact on detail preservation. When the τ_d are high, the model retains only the Gaussian points that contribute strongly to the overall structure, and facial details tend to be discarded. As the τ_d gradually decrease, the model begins to include more weak-contribution Gaussian points that capture subtle features. For instance, in the green box, the color transition and texture granularity of the eye shadow are clearly rendered under a lower threshold; in the red box, the muscle folds at the mouth corner and the contours of the teeth are significantly better preserved. These observations directly demonstrate the crucial role of dynamic threshold control in generating fine details. We divide the training into two phases: coarse and fine, as shown in Fig. 4. In the coarse phase, we set a higher densification threshold τ_d, allowing Gaussian points to quickly cover the entire 3D geometry. In the fine phase, we dynamically adjust the densification threshold τ_d as the iteration progresses, transitioning from the initial values to the final ones:

$$\tau_d = \tau_{d_{init}} - \frac{t}{T}\left(\tau_{d_{init}} - \tau_{d_{final}}\right). \tag{1}$$

A.3 Training Details

In the coarse phase, we optimize the 3DGS attributes and the spatial position of multi-resolution three-

Fig. 3 Visualization under different densification thresholds. Lowering the threshold τ_d reveals finer details, with clearer eyelid and eyeshadow features (green box) and more accurate mouth contours and tooth structures (red box).

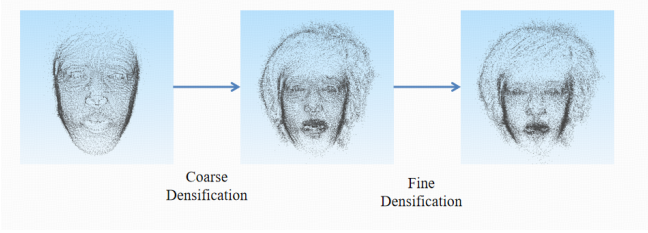

Fig. 4 Progressive Adaptive Densification. Coarse densification builds the overall facial geometry, while fine densification fills in the facial details.

plane encoding through a rough densification strategy. Rasterization techniques are used to reconstruct the coarse facial head $\hat{I}_{\text{coarse}}$:

$$\hat{I}_{\text{coarse}} = R(G_0; \pi), \tag{2}$$

where π is the extrinsic camera pose. In addition, we use pixel-level reconstruction loss (L_1) and SSIM loss to measure the error between the original image I_{gt} and the generated image $\hat{I}$ within the facial mask:

$$\mathcal{L}_{\text{coarse}} = \lambda \mathcal{L}_1(I_{\text{gt}}M, \hat{I}M) + (1-\lambda)\mathcal{L}_{\text{SSIM}}(I_{\text{gt}}, \hat{I}). \tag{3}$$

In the fine stage, the spatial deformation of the position is learned through FAAN, and the following rasterization generates each dynamic frame:

$$\hat{I}_{\text{fine}} = R(G_{deform}; \pi), \tag{4}$$

the overall model is fine-tuned by adding LPIPS loss to enhance lip details based on the coarse stage. The optical flow loss between the original image and the generated image is computed to stabilize facial motion trends and reduce frame jitter:

$$\mathcal{L}_{\text{fine}} = \mathcal{L}_{\text{flow}}(I_{\text{gt}}M, \hat{I}M) + \lambda_{\text{lpips}}\mathcal{L}_{\text{lpips}}(I_{\text{gt}}, \hat{I}). \tag{5}$$

Finally, the overall loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{coarse}} + \mathcal{L}_{\text{fine}} + \mathcal{L}_{\text{VGG}}, \tag{6}$$

where M represents the facial region mask, L_{VGG} is the perceptual loss function.

B. Additional Experiments

B.1 Qualitative evaluation

In the cross-driven setting, as shown in Fig. 5, PD-GauTalk generates accurate lip movements across different target individuals and performs excellently even with out-of-domain audio.

Fig. 5 Comparative results under the cross-driven setting. Each line represents the generated results for the same audio but different target individuals, specifically aligning with the phonemes in the words ”year”, ”look”, ”store”, and ”pass”.

B.2 Ablation study

In this section, we conduct ablation experiments under the self-driven setting to validate the effectiveness of each component, providing both quantitative and qualitative demonstrations. Specifically,

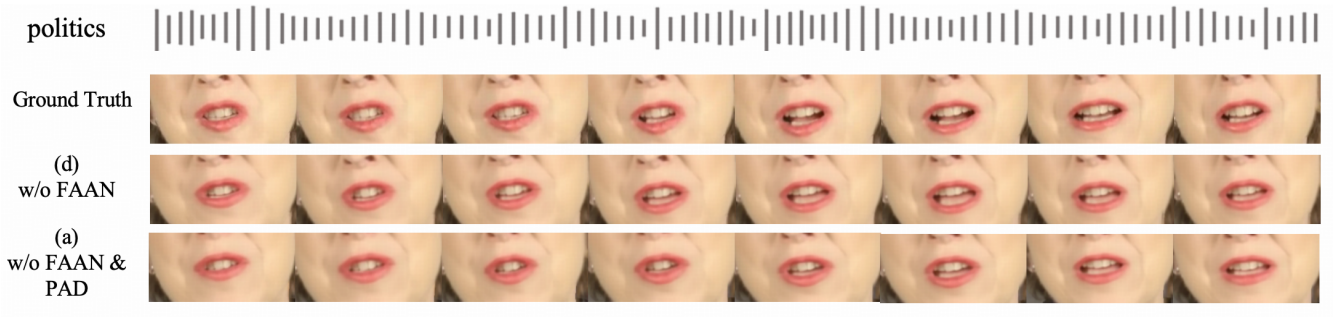

Fig. 6 Ablation study of PAD. The comparison demonstration of continuous frame results for the PAD model was conducted when the target person says "politics" within the self-driven framework.

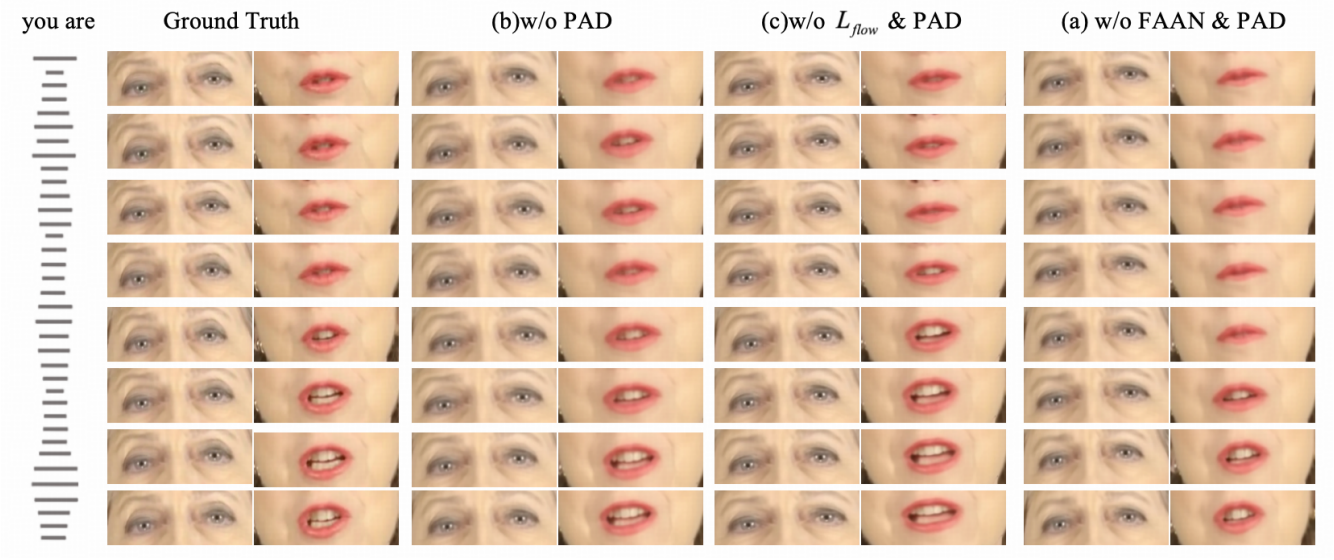

Fig. 7 Ablation study of FAAN. The comparison demonstration of continuous frame results for the the FAAN and FAAN (without L_{flow}) models was conducted when the target person says "you are" within the self-driven driving framework.

baseline (a) represents the baseline without the FAAN and PAD modules; (b) denotes the addition of the FAAN module to the baseline (a); (c) indicates the removal of the optical flow loss L_{flow} based on (b); and (d) represents the addition of only the PAD module to the baseline (a). Finally, we present the complete PD-GauTalk model for comprehensive evaluation.

B.2.1 Flow Audio Attention Network

In the setting (b), FAAN uses the RAFT model to extract the optical flow fields of talking head videos, learning inter-frame facial motion trends to accurately capture facial dynamics. As shown in Fig. 7, by integrating optical flow information with audio through spatial-temporal attention fusion, FAAN enhances lip-sync consistency, improves inter-frame continuity, and effectively reduces inter-frame jitter and abrupt changes. As shown in Table 1, all metrics show improvement compared to the baseline (a). In the setting (c), FAAN does not include optical flow loss, making the model focus more on minimizing pixel-level errors and improving PSNR. However, due to the lack of optimization for dynamic regions, the generated facial dynamics appear less natural, leading to a decline in

Table 1 The quantitative evaluation of ablation study. ↑ indicates higher is better while ↓ indicates lower is better. The best and second-best results are in **bold** and underline.

Setting	FAAN	FAAN w/o L_{flow}	PAD	PSNR↑	LPIPS↓	LMD↓	FID↓	Sync↑
Ground Truth	-	-	-	-	-	-	-	8.115
baseline (a)	✗	✗	✗	30.5324	0.0666	2.8852	6.7775	5.716
b	✓	✗	✗	30.8590	0.0599	2.9366	6.2746	6.015
c	✗	✓	✗	**30.9311**	0.0643	2.9354	6.6022	5.984
d	✗	✗	✓	30.9214	0.0649	**2.7399**	7.0018	6.049
PD-GauTalk (Ours)	✓	✓	✓	30.8629	**0.0596**	2.7621	**6.1249**	**6.058**

LPIPS, FID, and Sync metrics.

B.2.2 Progressive Adaptive Densification

As shown in Fig. 6, the coarse-to-fine PAD strategy effectively guides the model to generate more realistic facial details, such as forehead wrinkles. Table 1 shows that compared to the baseline (a), PAD allows the model to focus more on generating facial details while reconstructing the overall geometry, significantly improving PSNR and LPIPS. Additionally, PAD enhances the details of lip movements, further improving the Sync metric. However, the increased focus on local details causes a slight decline in the FID metric. Nevertheless, from an overall perspective, the results show significant improvements in detail quality and dynamic consistency.