FA-LAM: Focus-Aware Large Avatar Model for One-Shot 4D Animatable Gaussian Head

Yingdong Hu[*]
HKUST
Hong Kong SAR, China
yhudj@connect.ust.hk

Yisheng He[*✉]
Alibaba Token Hub, Alibaba Group
Hangzhou, China
ethanheysh@gmail.com

Yiming Jiang
Alibaba Token Hub, Alibaba Group
Hangzhou, China

Zehong Lin
Lingnan University
Hong Kong SAR, China

Steven Hoi
Alibaba Token Hub, Alibaba Group
Hangzhou, China

Jun Zhang
HKUST
Hong Kong SAR, China

Abstract

We propose FA-LAM, a **F**ocus-**A**ware **L**arge **A**vatar **M**odel for one-shot animatable Gaussian head creation, while simultaneously enabling static 3D and dynamic 4D full-head recovery. The core of our method lies in a thorough analysis of the attention mechanisms and the entangled reconstruction and animation training pipeline adopted by prior state-of-the-art approaches. Our analysis identifies two main factors that compromise the quality of 3D full-head generation: (1) incorrect and noisy attention activations, and (2) conflicts between the tasks of reconstruction and animation. To address the first issue, we introduce a symmetric and semantic attention regularization strategy that leverages the inherent semantics and structural symmetry of human heads. To disentangle the objectives of reconstruction and animation, we develop a novel dual-phase training pipeline that separates the model's capabilities for large-view hallucination and animation into distinct modules. Moreover, we enhance our model to support multi-view and streaming 4D reconstruction in an efficient and memory-friendly manner through a core autoregressive modification with tailored visibility-aware token fusion. Collectively, these innovations enable FA-LAM to reconstruct animatable Gaussian full heads with superior quality, particularly in fine facial regions and large viewing angles. Project page: https://zhenliuzju.github.io/huyingdong/FA-LAM/.

1 Introduction

The generation of animatable human avatars from a single image and the reconstruction of dynamic 4D human heads from monocular video streams are fundamental challenges in digital human modeling. These tasks are critical for emerging applications, such as virtual try-on, telepresence, AR/VR, and live performance digitization. Nevertheless, it is extremely difficult to solve both tasks simultaneously due to the severe ambiguity inherent in monocular inputs and the need to balance high-fidelity appearance reconstruction and controllable expression animation.

[1]* indicates equal contribution. Work done while Yingdong and Yiming were interns at Alibaba Group.
[2]✉ Corresponding author. Contact:ethanheysh@gmail.com
[3]This work was supported by Alibaba Research Intern Program.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Traditional methods address these tasks by overfitting Gaussian avatars [1] or neural radiance fields (NeRFs) [2, 3] to monocular or multi-view videos [1, 4–6]. While effective, the per-subject optimization is computationally heavy and yields poor novel view synthesis for viewpoints far from the training cameras. Another line of work leverages large multi-view, multi-expression datasets [7] to learn strong human priors that enable few-shot inference [8–10]. However, the limited model capacity often leads to degraded quality when generalizing to unseen identities or extreme poses.

Recently, feed-forward transformer-based reconstruction models [11–15] have demonstrated compelling inference speed and generalization. They embed head representations, such as FLAME [16] tokens, and apply cross-attention to image features to directly regress 3D Gaussian assets, bypassing test-time optimization. Some works [14, 17] extend this idea to 4D modeling by learning a direct mapping from images to dynamic representations. Despite promising progress, existing methods exhibit several critical limitations. First, attention mechanisms are operated without explicit regularization on the relationship between avatar tokens and image regions, often ignoring the inherent bilateral symmetry of human heads and causing blurry textures in fine detail areas such as hair. Second, these methods suffer from inherent conflicts between reconstruction fidelity and animation quality, making it difficult to generate avatars that simultaneously achieve high-quality novel view rendering and accurate expression transfer under joint supervision. Third, naively extending such models to multi-view [18] or streaming scenarios [19] results in quadratic computational growth with the number of input frames, making them impractical for real-time or resource-constrained settings.

In this work, we present a unified framework to address these challenges with three key innovations. First, to address the lack of explicit attention guidance in existing methods, we provide an in-depth analysis of the attention mechanisms within large avatar reconstruction networks. Based on this analysis, we **propose a semantic and symmetric attention regularization method** that explicitly encourages the model to focus on semantically meaningful and spatially symmetric image regions, leveraging the natural bilateral symmetry of human heads. This regularization significantly improves the clarity of the rendering of reconstructed head avatars, particularly around fine facial features. Second, to resolve the common frontal-view bias and significant large-view distortion observed in prior works, we identify that the underlying cause is gradient conflict between two supervision objectives: novel-view synthesis (reconstruction) and novel-expression transfer (animation). Through empirical gradient analysis, we show that jointly training with both objectives leads to competing updates that degrade performance, especially for large viewing angles. To resolve this, we **propose a simple yet effective two-stage training pipeline:** the first stage focuses purely on static 3D reconstruction from multi-view images, while the second stage specializes in refining expression-driven animation only on highly dynamic regions from single-view or multi-view videos. This decoupling enables our model to excel at two distinct tasks within a unified architecture: static full-head 3D reconstruction and animatable avatar creation. Third, to support efficient multi-view and streaming scenarios, we extend our framework to handle sequential input streams via an autoregressive reconstruction design. Specifically, we **introduce a visibility-gated Gaussian fusion mechanism** that selectively updates the canonical avatar representation based on the visibility map of each incoming frame. This design ensures constant memory consumption regardless of sequence length and allows our model to temporally align 4D Gaussian assets from a monocular stream, updating the head avatar incrementally as new observations arrive. To facilitate effective training of our model, we curate a multi-view extension of the VFHQ dataset [20]. We augment it with high-quality side-view images generated by Qwen-Image-Edit [21], providing stronger supervision for wide viewing angles.

In summary, our contributions are as follows:

- We propose a large avatar reconstruction model for one-shot animatable Gaussian full-head avatars, achieving improved attention activation, enhanced detail reconstruction, and superior novel view synthesis quality. The model is further extended into an autoregressive pipeline with visibility-gated Gaussian fusion, enabling streaming and multi-view reconstruction under constant memory consumption.
- We introduce a semantic and symmetric attention regularization method that directs attention toward semantically meaningful and spatially relevant image regions according to human head symmetry, improving overall reconstruction fidelity and fine-grained rendering.
- We identify that performance degradation in large-view synthesis arises from the gradient entanglement between reconstruction and animation objectives, and resolve it with a dual-phase training strategy that decouples these two learning signals.

2 Related Work

2.1 Images to Head Avatars

Early 2D-based methods for image-to-head avatar reconstruction typically rely on CNNs and GANs [22] to synthesize morphable faces. Some inject pose and expression features into generators [23–26], often structured as UNet or StyleGAN [27]. Others warp expressions and head poses as motion fields on source views [28–30]. Recent efforts have turned to diffusion models for high-quality face generation [31–33]. Notably, CAP4D [34] and MVP4D [35] incorporate 3D Morphable Models (3DMMs) [16, 36, 37] to guide novel view synthesis with novel expressions. Despite their strengths, these 2D models suffer from multi-view consistency issues and inefficient inference. Another line of research lifts images to 3D head avatars, starting with animatable mesh- or point-cloud-based avatars from image sequences [4, 16]. Subsequent work have shifted to more powerful representations, including NeRFs [2, 3, 38–41] and 3DGS [1, 42, 43]. Recent works [8, 44] reduce input requirements by leveraging diffusion models for novel view augmentation and reconstruct high-quality avatars from augmented images. However, all these methods require costly iterative optimization per case, lacking the temporal efficiency of feed-forward reconstruction.

2.2 Feed-Forward 3D Head Avatar Reconstruction

Feed-forward reconstruction [11, 12, 17, 18] has recently gained attention for its efficiency in unconstrained scenarios, enabled by transformer scaling. In head avatar modeling, recent works [12–14, 45] incorporate human geometry priors into transformers for efficient feed-forward reconstruction. Some approaches [14, 46, 47] reconstruct static representations from inputs, achieving high-quality reconstruction and novel view synthesis, but cannot animate novel expressions or poses. Conversely, other works [12, 13, 15, 48, 49] introduce parametric priors like FLAME [16] into feed-forward frameworks for animatable avatar reconstruction from single or multi-view images. Specifically, LAM [12] and MeshLAM [13] use FLAME priors with transformer attention to reconstruct animatable Gaussians and meshes from a single image. Avat3r [15] employs transformers as an expression driver to animate DUSt3R [50] outputs. Concurrent works improve expression driving with transformers [49] or UNet [48], enhance multi-view capability with large-scale datasets [48, 51], support multi-view inputs [51], and utilize UV space [51, 52]. However, these methods neglect bilateral facial symmetry and the conflict between reconstruction and animation, leading to poor rendering quality under novel expressions and views. In contrast, we systematically examine two common issues in LRM-based avatar models: detail blurring and large-view degradation. To address them, we propose bilateral semantic attention and a dual-phase training strategy. Furthermore, to handle out-of-memory issues with multiple or streaming frames, we introduce a visibility-aware autoregressive approach.

3 Method

Given a single portrait image $\mathbf{I}_{in}$, **FA-LAM** aims to construct an animatable 3D Gaussian asset $\mathcal{G}_{3D} = \{\boldsymbol{\mu}_k, \boldsymbol{s}_k, \boldsymbol{r}_k, o_k, \boldsymbol{c}_k\}_{k=1}^{M}$ that can be controlled via expression codes $\mathbf{z}_{exp}$ and rendered from arbitrary viewpoints with minimal image distortion. This process can be formalized as:

$$\mathcal{G}_{3D} = \mathbf{D}_{\theta_2}(\mathbf{E}_{\theta_1}(\mathbf{I}_{in}), \mathbf{T}_{UV}, \boldsymbol{\phi}_{FLAME}), \tag{1}$$

where $\mathbf{E}_{\theta_1}$ denotes a pretrained Vision Transformer (ViT) network that extracts features from the input image $\mathbf{I}_{in}$, $\mathbf{T}_{UV}$ is a learnable token map defined in UV space, and $\mathbf{D}_{\theta_2}$ represents a transformer-based decoder that maps learnable tokens to 3D Gaussian attributes. $\boldsymbol{\phi}_{FLAME}$ refers to the parametric 3D morphable model (FLAME [16]).

In the following, we first introduce the architecture of the encoding and decoding networks in Sec. 3.1. We then detail our proposed attention regularization strategy in Sec. 3.2. Finally, we present a novel two-stage training pipeline in Sec. 3.3 to disentangle the reconstruction and animation objectives.

3.1 UV Space Encoding and Decoding

To improve spatial reasoning of 3D Gaussian tokens, we perform reconstruction in the UV space, as illustrated in Fig. 1. Given an input RGB image $\mathbf{I}_{in}$, FA-LAM first extracts a coarse RGB UV image $\mathbf{I}_{UV}$ using a FLAME model animated by the tracked pose, expression, and camera parameters. The extracted UV image $\mathbf{I}_{UV}$, together with the original RGB image $\mathbf{I}_{in}$, are passed through separate

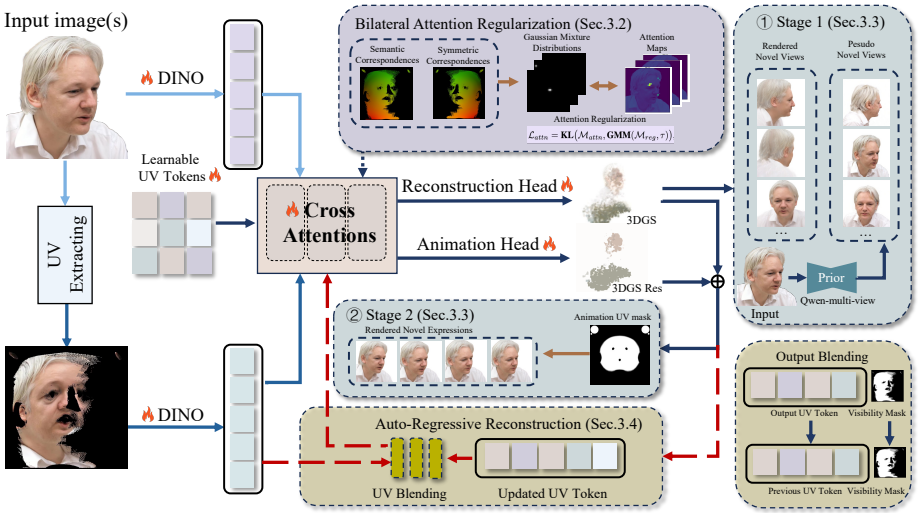

Figure 1: **Overview of FA-LAM.** Given an input RGB image, we first perform learning-free coarse UV extraction to obtain a coarse RGB UV map. The UV map and the original image are separately encoded by two DINO encoders. A learnable UV token then queries cross-attention over both UV and RGB features. The resulting UV tokens are used to decode canonical 3D Gaussians and animation-specific Gaussian residuals.

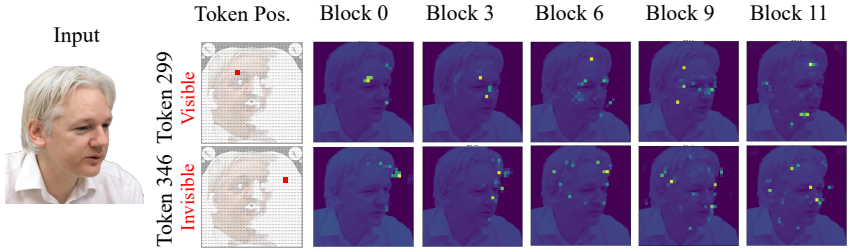

Figure 2: **Attention map activations** in our network w/o regularization. Token 299 (visible in the input RGB image) exhibits attention mostly on relevant facial regions with some noise, whereas token 346 (not visible) produces largely random activations without coherent semantic focus. This motivates our attention regularization method to improve attention utilization. Please zoom in for detailed views and corresponding UV map token positions.

DINO encoders $\mathbf{DINO}_{\theta_{in}}$, $\mathbf{DINO}_{\theta_{UV}}$ to extract image features and UV features, respectively. This process can be formulated as:

$$\mathbf{F}_{in}, \mathbf{F}_{UV} = \mathbf{DINO}_{\theta_{in}}(\mathbf{I}_{in}),\ \mathbf{DINO}_{\theta_{UV}}\big(\mathbf{ExtractUV}(\mathbf{I}_{in})\big), \tag{2}$$

where the UV extraction operation **ExtractUV** is detailed in the Appendix.

Concurrently, a token map $\mathbf{T}_{UV}$ defined in UV space is initialized with learned semantic embeddings and undergoes cross-attention with the extracted image and UV features:

$$\mathbf{T}^{'}_{UV} = \mathbf{CrossAttention}\big(q = \mathbf{T}_{UV},\ kv = \mathbf{F}_{in} \oplus \mathbf{F}_{UV}\big). \tag{3}$$

The resulting UV tokens are then passed to a 3D Gaussian decoder, which decodes the semantic 3D tokens into a set of 3D Gaussians as $\mathcal{G}_{3D} = \mathbf{D}_{3DGS}(\mathbf{T}^{'}_{UV})$.

3.2 Semantic Attention Regularization with Human Symmetric Prior

A limitation of the vanilla model in Sec. 3.1 is that its cross-attention is not spatially well constrained even when trained with large-scale data. To demonstrate this, we train our framework on a large-scale dataset as in LAM [12]. We then visualize the activation to analyze how the attention mechanism works. As shown in Fig. 2, UV tokens corresponding to visible facial regions usually activate around

semantically relevant image patches, but the responses remain noisy and spatially dispersed. The failure is more severe for tokens corresponding to invisible regions, whose attention often collapses onto irrelevant facial parts or even background clutter. As a result, the model uses its attention budget inefficiently, which degrades reconstruction quality, especially for unseen regions.

This behavior stems from an ambiguity inherent to single-view reconstruction. For visible regions, the model receives direct evidence but still lacks an explicit constraint on where attention should concentrate. For invisible regions, the ambiguity is even worse as no direct image evidence exists, making the learned attention prone to drifting toward spurious but locally plausible patterns. In other words, standard end-to-end training does not provide sufficient structures for the model to distinguish informative correspondences from arbitrary activations.

We address this with semantic attention regularization using a human symmetry prior. Human heads are approximately bilaterally symmetric, especially in UV space, where facial layout is canonically aligned. When one side of the face is occluded, its appearance can often be inferred from the opposite symmetric region. We inject this prior directly into cross-attention layers. For each UV token, we determine its semantic correspondence in the input RGB image by reprojecting the tracked FLAME template into the image plane. Its correspondence in the extracted UV image is obtained by matching the token to the UV patch at the same spatial location. These correspondences provide explicit supervision on attention targets. To handle self-occlusion, we also define symmetric correspondences in both the RGB image and UV map, encouraging tokens of invisible regions to attend to their visible symmetric counterparts instead of drifting.

We implement this constraint by matching the predicted attention map to a target Gaussian mixture distribution centered at the semantic and symmetric correspondences. The regularization loss is

$$\mathcal{L}_{attn} = \mathbf{KL}\big(\mathcal{M}_{attn}, \mathbf{GMM}(\mathcal{M}_{reg}, \tau)\big), \tag{4}$$

where $\mathcal{M}_{attn}$ denotes the predicted attention map, $\mathcal{M}_{reg}$ denotes the target correspondence locations, $\mathbf{GMM}(\cdot, \tau)$ converts these sparse locations into a Gaussian mixture with fixed standard deviation τ, and $\mathbf{KL}$ denotes the Kullback–Leibler divergence. This regularization encourages attention to be spatially concentrated, semantically aligned, and symmetry-aware.

3.3 Disentangling Reconstruction and Animation

Direct training on monocular talking head videos (LAM [12]) yields strong animation but poor generalization to unseen views. In contrast, training on static multi-view images (PanoLAM [14]) enables high-quality wide-view reconstruction but cannot animate new expressions. A naive combination of both objectives produces blurry results (Fig. 7), indicating a negative correlation between reconstruction and animation. To investigate this, we analyze the gradient cosine similarity and the t-SNE feature distributions following PCGrad [53] and t-SNE [54], comparing novel expression and novel view gradients, and visualizing features from networks, where PCGrad is calculated with joint objective training and t-SNE with separate training.

As shown in Fig. 3, attention layer gradients exhibit negative correlation (cosine similarity < -0.1) for nearly half of the pairs, with the remainder nearly orthogonal (cosine similarity between -0.1 and 0.1). Such negative correlations hinder multi-task optimization by causing one task to dominate training [53]. To address this issue, we thus propose a two-stage strategy. First, we train for reconstruction only on multi-view portrait data, including our augmented VFHQ dataset, to learn a canonical neutral full-head Gaussian representation with strong cross-view consistency. Second, we fine-tune on animation data for controllable expression dynamics and long-sequence motion. To preserve reconstruction quality, expression deformations are modeled as residual corrections on the canonical Gaussians (Fig. 1). Specifically, we predict residual Gaussian attributes only for regions that require large non-rigid motion (periocular and

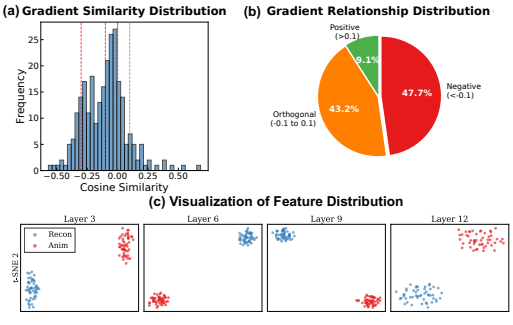

Figure 3: **Gradient and feature analysis** reveals strong negative correlation between reconstruction and animation objectives, causing performance degradation under joint training.

perioral areas), decoded from the same UV tokens via an additional DPT head. This design separates identity-preserving canonical geometry from expression-specific deformation, improving both animation quality and large-view synthesis.

3.4 Autoregressive Reconstruction with Visibility-Gated Feature Fusion

We further extend our framework to multi-view and sequential inputs in an autoregressive manner. Unlike full-attention designs, whose memory footprint grows quadratically with the number of input images, our autoregressive model stores only current features and an accumulated state, achieving constant memory per additional frame. The key challenge is that each frame provides only partial evidence: some regions are newly revealed, while others are occluded or unreliable. Directly fusing the accumulated state with incoming predictions thus tends to overwrite stable geometry with inferior observations. To address this, we introduce *visibility-gated feature fusion*, which updates each UV-space Gaussian feature according to its reliability in the current frame.

Given a sequence $\{\mathbf{I}_{in}^{(1)}, \ldots, \mathbf{I}_{in}^{(N)}\}$, FA-LAM processes the first frame $\mathbf{I}_{in}^{(1)}$ (Sec. 3.1) to obtain UV tokens $\mathbf{T}_{UV}^{(1)}$ and Gaussian representation $\mathcal{G}_{3D}^{(1)}$. For each $t \geq 2$, we update the latent UV state by fusing the previous tokens with the current UV features:

$$\mathbf{F}_{UV}^{'(t)} = \mathbf{MLP}_{\phi}\left(\mathbf{T}_{UV}^{'(t-1)} \oplus \mathbf{F}_{UV}^{(t)}\right), \tag{5}$$

where $\mathbf{F}_{UV}^{(t)}$ denotes the UV features extracted from frame t, and $\mathbf{MLP}_{\phi}$ is a lightweight fusion module. The UV tokens are then updated with image features and the new UV features $\mathbf{F}_{UV}^{'(t)}$ following the same process as in Sec. 3.1. The updated tokens are fused with the output tokens of the current frame using the following equation:

$$\begin{aligned}\mathbf{T}_{UV}^{'(t)} =&(\mathbf{C}^{'(t-1)} \wedge \neg\mathbf{C}^{(t)}) \odot \mathbf{T}_{UV}^{'(t-1)} + (\neg\mathbf{C}^{'(t-1)} \wedge \mathbf{C}^{(t)}) \odot \mathbf{T}_{UV}^{t} \\ &+(\mathbf{C}^{'(t-1)} \wedge \mathbf{C}^{(t)}) \odot \mathbf{MLP}_{\psi}(\mathbf{T}_{UV}^{'(t-1)} \oplus \mathbf{T}_{UV}^{t}).\end{aligned} \tag{6}$$

Here, $\mathbf{C}^{(t)}$ measures the visibility of each token in the current view. Visible regions $\mathbf{C}^{'(t-1)}$ are updated by incorporating visible areas from the current frame into the historical visibility map. This prevents degradation caused by partial observations and allows the representation to improve progressively as new frames arrive. The fused tokens are then decoded into a 3D Gaussian representation for the current frame as $\mathcal{G}_{3D}^{(t)} = \mathbf{D}_{3DGS}(\mathbf{T}_{UV}^{'(t)})$. In this way, our method remains practical for both sparse multi-view fusion and long streaming sequences.

4 Experiment

4.1 Experiment Details

Dataset. Our method, **FA-LAM**, is trained on the VFHQ dataset [20], the Nersemble-v2 dataset [55], the Ava256 dataset [56], and our self-curated MV-VFHQ dataset. To construct the MV-VFHQ dataset, we employ the Qwen-Image-Edit [21] image generator to produce a per-frame multi-view image set, comprising four distinct novel views for each input image. The detailed procedure for generating and annotating the MV-VFHQ dataset is provided in the Appendix.

Training. We conduct all of the training processes on $8\times$H20 GPUs. The training of **FA-LAM** is decomposed into three stages. First, we train a static reconstruction variant of **FA-LAM** under a per-frame static setting using the MV-VFHQ, Nersemble-v2, and Ava256 datasets. Second, we adapt this pipeline into an autoregressive reconstruction framework by further fine-tuning under a multi-view setting, again on the MV-VFHQ, Nersemble-v2, and Ava256 datasets. Finally, we fine-tune the model to enable smooth animation, supervised with sequential images from the VFHQ, Nersemble-v2, and Ava256 datasets. The model is trained with a composite loss function comprising MSE loss, LPIPS loss, mask loss, and attention regularization loss, formulated as:

$$\mathcal{L}_{all} = \mathbf{MSE}(\boldsymbol{I}_{render}, \hat{\boldsymbol{I}}_{gt}) + \mathbf{LPIPS}(\boldsymbol{I}_{render}, \hat{\boldsymbol{I}}_{gt}) + \mathbf{MSE}(\boldsymbol{M}_{render}, \hat{\boldsymbol{M}}_{gt}) + \mathcal{L}_{attn}. \tag{7}$$

Metrics. To evaluate rendered novel-view image quality, we adopt the most commonly used image metrics: PSNR, SSIM, and LPIPS. For identity similarity (CSIM), we compute the cosine distance of face recognition features following [57]. To assess expression and pose fidelity, we employ the Average Expression Distance (AED) and Average Pose Distance (APD) as estimated by a 3D

Morphable Model (3DMM) [45]. Additionally, we measure the Average Keypoint Distance (AKD) using a facial landmark detector [58]. These metrics provide insights into the accuracy of driving control in our animations. For cross-identity reenactment, where ground truth images are unavailable, we rely on CSIM, AED, and APD for evaluation. These metrics align with those used in prior work [12], enabling consistent comparisons across different approaches.

Table 1: **Overview of experimental results**. We evaluate **FA-LAM** on 4 different tasks and 3 different datasets. e, v, and i represents novel expression, view and identity respectively.

Task	#Inputs	Evaluation Data	Evaluation Type	Tabs.& Figs.
Single View Avatar Creation (Self-reenactment)	1	VHFQ	e	Tab.2, Fig.4
- Novel View Assessment	1	Nersemble-v2	e, v	Tab.3, Fig.5
Single View Avatar Creation (Cross-reenactment)	1	VHFQ	e, i	Tab.2, Fig.4
Multi-View Avatar Creation (Self-reenactment)	4	Ava256	e, v	Tab.4, Fig.5
4D Human Head Reconstruction	stream	Nersemble-v2	v	Tab.5

4.2 Main Results

Overview. We carry out experiments on different settings with different datasets to evaluate the performance of our model under various situations. We list our experimental settings in Tab.1.

Baselines. We ensure fair comparisons with all state-of-the-art, open-source head avatar reconstruction models that are directly comparable. For the monocular setting, we compare against LAM [12], GAGAvatar [45], Portrait4D-v2 [3], Real3DPortrait [59], and GPAvatar [60]. In the multi-view setting, we select InvertAvatar [61], GPAvatar [60] DiffusionRig [62], and FastAvatar [63] as baselines. For 4D reconstruction, we include Forge4D [17], PanoLAM [14], and LAM [12]. Notably, Forge4D requires four input views to achieve good reconstruction, whereas PanoLAM and LAM take only a single RGB image as input, consistent with our approach.

Single-Image Avatar Creation. We evaluate on VFHQ for self and cross-reenactment (Tab. 2, Fig. 4). Our model matches prior methods on self-reenactment and surpasses all on cross-reenactment, with maximum gains of +0.27 dB (1.28%) in PSNR and 0.029 (2.37%) in CSIM. On NeRSemble-v2 (Tab. 3), our model achieves best novel-view quality for novel views. Inference time is 0.4644s on an H20 GPU, comparable to feed-forward methods and faster than optimization-based approaches.

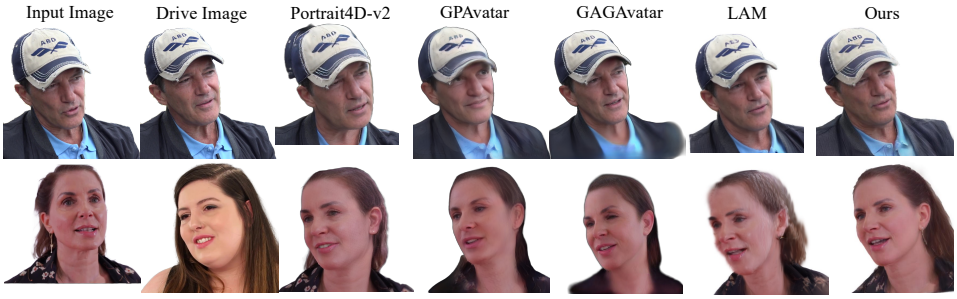

Figure 4: **Qualitative visualizations on the VFHQ dataset.**

Table 2: Comparisons for single-image avatar creation on the VFHQ dataset. This experiment focuses on evaluating models' ability in self-renactment and cross-renactment.

	Self-Renactment							Cross-Renactment		
	PSNR↑	SSIM↑	LPIPS↓	CSIM↑	AED↓	APD↓	AKD↓	CSIM↑	AED↓	APD↓
GPAvatar	21.04	0.807	0.150	0.772	0.132	0.189	4.226	0.564	0.255	0.328
Real3DPortrait	20.88	0.780	0.154	0.801	0.150	0.268	5.971	<u>0.663</u>	0.296	0.411
Portrait4D-v2	21.34	0.791	0.144	0.803	0.117	0.187	3.749	0.656	0.268	0.273
GAGAvatar	21.83	0.818	0.122	0.816	<u>0.111</u>	0.135	3.349	0.633	0.253	<u>0.247</u>
LAM	<u>22.65</u>	<u>0.829</u>	<u>0.109</u>	<u>0.822</u>	**0.102**	<u>0.134</u>	**2.059**	0.651	<u>0.250</u>	0.356
Ours	**22.91**	**0.857**	**0.096**	**0.844**	0.117	**0.112**	<u>2.464</u>	**0.692**	**0.229**	**0.234**

Few-Image Avatar Creation. We evaluate few-shot avatar creation performance on the Ava256 dataset. Quantitative comparisons are summarized in Tab. 4, and qualitative results are illustrated

Table 3: Comparisons on self-renactment for single-image avatar creation on the Nersemble-v2 dataset. This experiment focuses on evaluating models' ability in novel view synthesis.

	PSNR↑	SSIM↑	LPIPS↓	CSIM↑	AED↓	APD↓	AKD↓
GPAvatar	21.71	0.871	0.319	0.543	0.539	0.541	12.537
Real3DPortrait	20.71	0.885	0.317	0.586	0.511	0.573	11.749
Portrait4D-v2	21.98	0.854	0.301	0.554	0.534	0.610	37.738
GAGAvatar	23.46	0.906	0.157	0.693	0.433	0.217	7.874
LAM	18.32	0.809	0.131	0.538	0.319	0.301	31.479
Ours	**25.96**	**0.954**	**0.087**	**0.766**	**0.102**	**0.128**	**7.034**

Table 4: Comparisons for single- and few-image avatar creation on the Ava256 dataset. This experiment focuses on evaluating models' flexibility in changing the input settings. FFD, GAN, and DIF mean the method is based on feed-forward network, GAN inversion, or Diffusion, respectively.

One-Shot	Type	PSNR↑	SSIM↑	LPIPS↓	CSIM↑	AED↓	APD↓	AKD↓
Portrait4D-v2	FFD	18.58	0.738	0.186	**0.598**	0.172	0.391	15.672
GAGAvatar	FFD	19.86	0.688	0.254	0.506	**0.157**	0.428	19.875
LAM	FFD	18.45	0.621	0.289	0.463	0.192	0.557	16.338
InvertAvatar	GAN	19.72	0.695	0.228	0.397	0.189	0.491	12.731
GPAvatar	FFD	19.89	0.699	0.277	0.386	0.201	0.530	24.941
DiffusionRig	DIF	20.71	0.718	0.268	0.389	0.223	0.445	9.360
FastAvatar	FFD	16.92	0.536	0.349	0.379	0.205	0.569	20.842
Ours	FFD	**22.18**	**0.733**	**0.156**	0.517	0.171	**0.380**	**6.141**
Few-Shot	Type	PSNR↑	SSIM↑	LPIPS↓	CSIM↑	AED↓	APD↓	AKD↓
InvertAvatar	GAN	20.73	0.724	0.213	0.441	0.183	0.555	12.403
GPAvatar	FFD	18.91	0.675	0.294	0.349	0.204	0.509	22.499
DiffusionRig	DIF	22.09	0.756	0.251	0.487	0.198	0.401	10.431
FastAvatar	FFD	16.97	0.530	0.332	0.446	0.209	0.596	20.911
Ours	FFD	**22.46**	**0.766**	**0.120**	**0.655**	**0.153**	**0.316**	**4.183**

Table 5: Comparisons for 4D streaming reconstruction on the Nersemble-v2 dataset. This experiment demonstrates **FA-LAM**'s ability in 4D full-head reconstruction.

	Temporal Align	Input Views	PSNR↑	SSIM↑	LPIPS↓
Forge4D	Yes	4	19.32	0.712	0.146
PanoLAM	No	1	20.17	0.917	0.214
LAM	No	1	19.96	0.799	0.145
Ours	Yes	1	25.33	0.942	0.096

in Fig. 5. Our model consistently outperforms previous methods across all metrics. Additionally, creating an avatar from few-shot images yields a noticeable performance gain of +0.42 dB in PSNR.

Static Full-Head Reconstruction. We demonstrate the static per-frame full-head reconstruction capability of **FA-LAM** by comparing novel view image qualities of reconstructed human heads on the NeRSemble-v2 test set. We show metrical reports in Tab. 5. Our method outperforms other methods in reconstructed static human head quality.

4.3 Ablation Study

Effectiveness of Attention Regularization. We evaluate our attention regularization by comparing four variants: 1) no regularization; 2) semantic regularization only; 3) full regularization; and 4) direct DINO feature warping (replacing cross-attention). Attention activation visualizations (Fig. 6) show that our method achieves better activation with correct focus. Quantitative results (Tab. 7) and qualitative comparisons (Fig. 8) demonstrate that facial details are significantly improved, with bilateral symmetry issues persisting only under semantic regularization alone.

Effectiveness of the Dual-Phase Training Strategy. We demonstrate the effectiveness of our dual-phase training strategy by comparing the following training variants: 1) joint supervision of novel views and novel expressions; 2) training with novel expressions only; 3) training with novel views only; and 4) separating training into a reconstruction stage followed by an animation stage. Quantitative results and qualitative comparisons are presented in Tab. 6 and Fig. 7, respectively. The results indicate that our proposed attention regularization achieves better symmetrical hallucination, while the full semantic attention regularization leads to improved recovery of fine facial details.

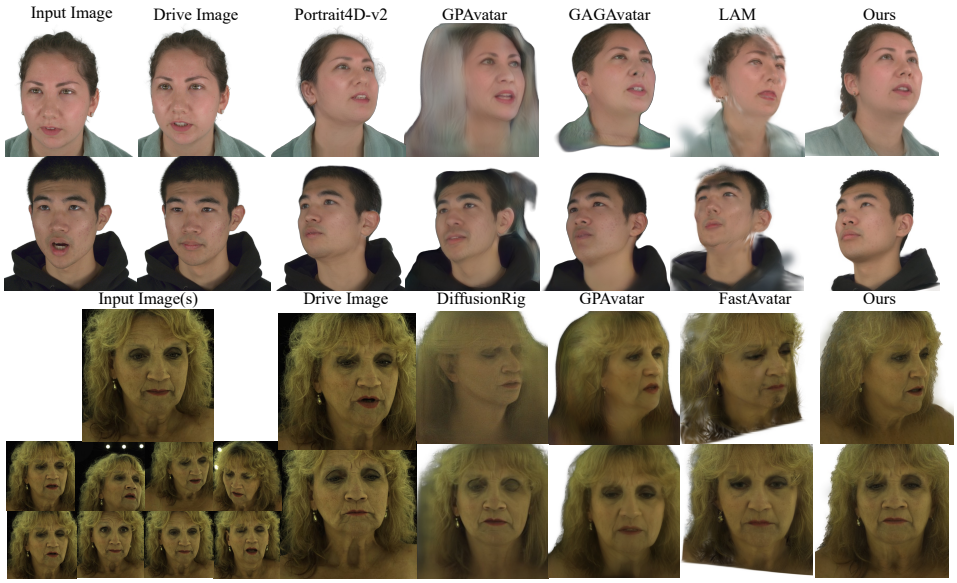

Figure 5: **Qualitative visualizations on Nersemble-v2 and Ava256 datasets.**

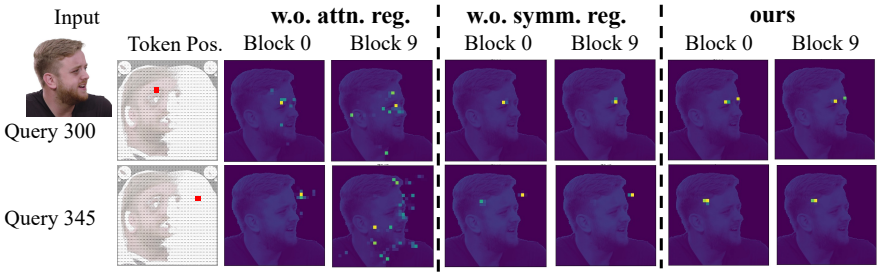

Figure 6: Visualizations on the attention activation maps of different attention regularization methods.

Table 6: Ablation studies on different training strategies.

	PSNR↑	SSIM↑	LPIPS↓
Joint training	24.99	0.938	0.103
Animation Only	24.93	0.940	0.111
Reconstruction Only	25.65	0.939	0.099
Recon-then-Ani (Ours)	25.80	0.949	0.091

Table 7: Ablation studies on different attention regularization strategies.

	PSNR↑	SSIM↑	LPIPS↓
direct feature unwarping	23.34	0.871	0.157
w.o. attn-reg	25.30	0.934	0.101
w. sym. attn-reg	25.68	0.932	0.096
w. full attn-reg (Ours)	25.80	0.949	0.091

Table 8: Quantitative results of 4D head reconstruction on Nersemble-v2 dataset with streaming inputs. We use TL-STD and TI to measure the temporal performance.

	TL-STD↓	TI↓	PSNR↑	SSIM↑	LPIPS↓	GPU Memory↓
Full-Attention	-	-	-	-	-	OOM
Framewise Reconstruction	8.141	13.731	24.46	0.940	0.106	8751MB
Auto-regressive (No Token Fusion)	8.253	13.760	24.53	0.940	0.099	8755MB
Auto-regressive (Visibility-Aware Fusion)	7.253	12.805	25.33	0.942	0.096	8771MB

Efficiency of Autoregressive Adaptation. We evaluate our autoregressive adaptation by comparing animated novel-view image quality and GPU memory consumption in Tab. 8. Directly reconstructing 4D heads from a large number of streaming frames (e.g., 20 frames) causes out-of-memory issues. In contrast, our autoregressive fusion model reconstructs 4D heads with minimal and constant additional GPU cost (+20 MB, 0.228%) and achieves more temporally stable results, as evidenced by gains in Temporal Luminance Standard Deviation (TL-STD) [64] (+0.888) and Temporal Information (TI, ITU-T P.910 [65]) (+0.926).

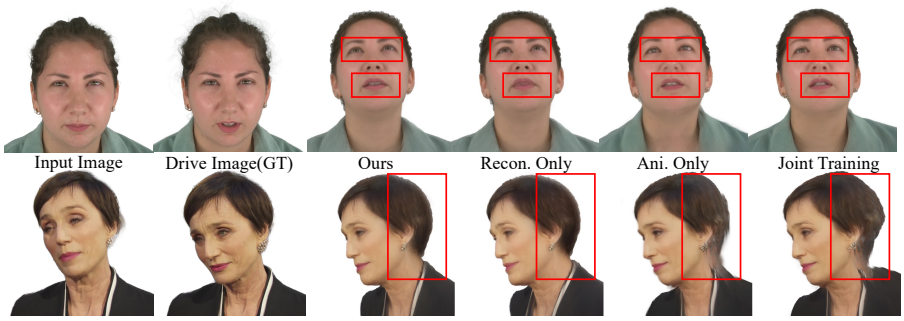

Figure 7: Qualitative visualization of different training strategies.

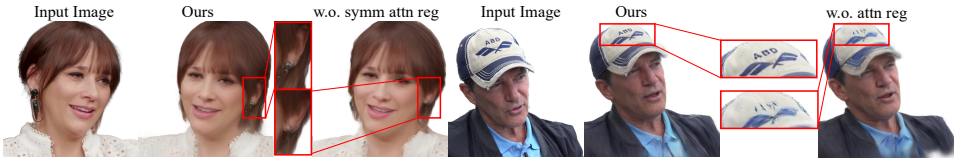

Figure 8: Qualitative visualization on different attention regularization supervisions.

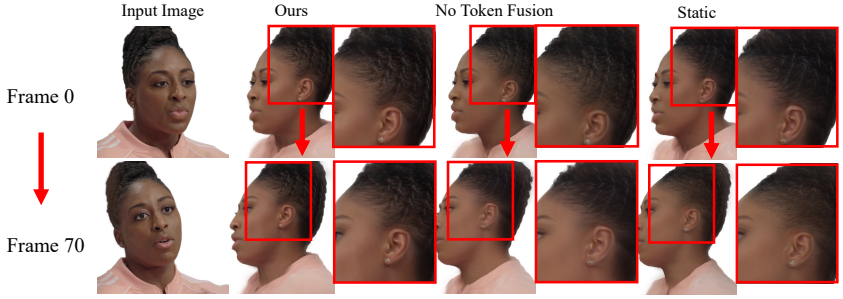

Figure 9: Qualitative evaluation on our auto-regressive token fusion designed. Our design enables model to generate temporal consistent hallucinations in invisible areas.

5 Conclusion

In this paper, we propose a 3D Gaussian-based focus-aware large full-head avatar reconstruction model that places greater emphasis on the semantic and symmetric properties of the human head. In addition, we present a comprehensive analysis of the degradation in large-view reconstruction performance when the model is supervised simultaneously with both novel-view and novel-time images, identifying this as an entanglement between the tasks of reconstruction and animation. Furthermore, we extend the proposed pipeline to support streaming inputs through an efficient autoregressive formulation. Future works may focus on improving the quality at occlusions and extending this pipeline to full-body avatars.

Limitations. FA-LAM relies on tracked FLAME parameters, so tracking accuracy directly determines reconstruction and animation quality; it also cannot capture details FLAME cannot model (e.g., dynamic wrinkles, sticking out tongue), which requires more expressive 3DMMs and tracking algorithms. Additionally, the model assumes bilateral head symmetry, so asymmetric attributes (e.g., single-sided decorations, hairstyles, or asymmetric expressions) may cause unreliable hallucinations on the unobserved side that deviate significantly from the ground truth. Finally, under large head rotations, the occiput exhibits blurriness and geometric distortion due to scarce rear-head observations and insufficient supervision in training data; we explicitly identify this as a future direction, including collecting more diverse occiput data or using 2D diffusion priors to hallucinate plausible back-of-head textures.

References

[1] 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*, pages 20299–20309, 2024.

[2] Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, and Baoyuan Wang. Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[3] Yu Deng, Duomin Wang, and Baoyuan Wang. Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. In *European Conference on Computer Vision*, pages 316–333. Springer, 2024.

[4] Jiawei Zhang, Zijian Wu, Zhiyang Liang, Yicheng Gong, Dongfang Hu, Yao Yao, Xun Cao, and Hao Zhu. Fate: Full-head gaussian avatar with textural editing from monocular video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5535–5545, 2025.

[5] Jaeseong Lee, Taewoong Kang, Marcel Buehler, Min-Jung Kim, Sungwon Hwang, Junha Hyung, Hyojin Jang, and Jaegul Choo. Surfhead: Affine rig blending for geometrically accurate 2d gaussian surfel head avatars. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=1x1gGg49jr`.

[6] Yuelang Xu, Lizhen Wang, Xiaochen Zhao, Hongwen Zhang, and Yebin Liu. Avatarmav: Fast 3d head avatar reconstruction using motion-aware neural voxels. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023.

[7] Yuelang Xu, Lizhen Wang, Zerong Zheng, Zhaoqi Su, and Yebin Liu. 3d gaussian parametric head model. In *European Conference on Computer Vision*, pages 129–147. Springer, 2024.

[8] Xiaozheng Zheng, Chao Wen, Zhaohu Li, Weiyi Zhang, Zhuo Su, Xu Chang, Yang Zhao, Zheng Lv, Xiaoyuan Zhang, Yongjie Zhang, Guidong Wang, and Lan Xu. HeadGAP: Few-Shot 3D Head Avatar via Generalizable Gaussian Priors . In *2025 International Conference on 3D Vision (3DV)*, pages 946–957, Los Alamitos, CA, USA, March 2025. IEEE Computer Society. doi: 10.1109/3DV66043.2025.00092. URL `https://doi.ieeecomputersociety.org/10.1109/3DV66043.2025.00092`.

[9] Zhixuan Yu, Ziqian Bai, Abhimitra Meka, Feitong Tan, Qiangeng Xu, Rohit Pandey, Sean Fanello, Hyun Soo Park, and Yinda Zhang. One2avatar: Generative implicit head avatar for few-shot user adaptation, 2024. URL `https://arxiv.org/abs/2402.11909`.

[10] Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao, and Yebin Liu. Monogaussianavatar: Monocular gaussian point-based head avatar. In *ACM SIGGRAPH 2024 conference papers*, pages 1–9, 2024.

[11] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3d. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=sllU8vvsFF`.

[12] Yisheng He, Xiaodong Gu, Xiaodan Ye, Chao Xu, Zhengyi Zhao, Yuan Dong, Weihao Yuan, Zilong Dong, and Liefeng Bo. Lam: Large avatar model for one-shot animatable gaussian head. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–13, 2025.

[13] Yisheng He. Feed-forward one-shot animatable textured mesh avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4145–4156, 2026.

[14] Peng Li, Yisheng He, Yingdong Hu, Yuan Dong, Weihao Yuan, Yuan Liu, Siyu Zhu, Gang Cheng, Zilong Dong, Steven Hoi, et al. Panolam: large avatar model for gaussian full-head synthesis from one-shot unposed image. *IEEE Transactions on Visualization and Computer Graphics*, 2026.

[15] 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 (ICCV)*, pages 12089–12100, October 2025.

[16] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. URL `https://doi.org/10.1145/3130800.3130813`.

[17] Yingdong Hu, Yisheng He, Jinnan Chen, Weihao Yuan, Kejie Qiu, Zehong Lin, Siyu Zhu, Zilong Dong, Steven Hoi, and Jun Zhang. Forge4d: Feed-human reconstruction and interpolation from uncalibrated sparse-view videos. In *European Conference on Computer Vision*, pages 297–315. Springer, 2026.

[18] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[19] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming visual geometry transformer. In *International Conference on Learning Representations*, volume 2026, pages 88055–88072, 2026.

[20] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.

[21] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL `https://arxiv.org/abs/2508.02324`.

[22] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[23] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13786–13795, 2020.

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

[25] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9459–9468, 2019.

[26] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3397–3406, 2022.

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[28] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2663–2671, 2022.

[29] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024.

[30] Bowen Zhang, Chenyang Qi, Pan Zhang, Bo Zhang, HsiangTao Wu, Dong Chen, Qifeng Chen, Yong Wang, and Fang Wen. Metaportrait: Identity-preserving talking head generation with fast personalized adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22096–22105, 2023.

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

[32] 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 *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=rkzabmWl5k.

[33] 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.

[34] Felix Taubner, Ruihang Zhang, Mathieu Tuli, and David B. Lindell. CAP4D: Creating animatable 4D portrait avatars with morphable multi-view diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5318–5330, June 2025.

[35] Felix Taubner, Ruihang Zhang, Mathieu Tuli, Sherwin Bahmani, and David B Lindell. Mvp4d: Multi-view portrait video diffusion for animatable 4d avatars. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–11, 2025.

[36] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.

[37] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[38] 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. *Communications of the ACM*, 65(1):99–106, 2021.

[39] Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. GPAvatar: Generalizable and precise head avatar from image(s). In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hgehGq2bDv.

[40] Tobias Kirschstein, Simon Giebenhain, Jiapeng Tang, Markos Georgopoulos, and Matthias Nießner. GGHead: Fast and Generalizable 3D Gaussian Heads. In *SIGGRAPH Asia 2024 Conference Papers*, SA '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400711312. doi: 10.1145/3680528.3687686. URL https://doi.org/10.1145/3680528.3687686.

[41] Weichuang Li, Longhao Zhang, Dong Wang, Bin Zhao, Zhigang Wang, Mulin Chen, Bang Zhang, Zhongjian Wang, Liefeng Bo, and Xuelong Li. One-Shot High-Fidelity Talking-Head Synthesis with Deformable Neural Radiance Field . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17969–17978, Los Alamitos, CA, USA, June 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.01723. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01723.

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

[43] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[44] Tingting Liao, Yujian Zheng, Yuliang Xiu, Adilbek Karmanov, Liwen Hu, Leyang Jin, and Hao Li. Soap: Style-omniscient animatable portraits. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025.

[45] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=gVM2AZ5xA6.

[46] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[47] Yingdong Hu, Zhening Liu, Jiawei Shao, Zehong Lin, and Jun Zhang. Eva-gaussian: 3d gaussian-based real-time human novel view synthesis under diverse camera settings, 2024. URL https://arxiv.org/abs/2410.01425.

[48] Cheng Peng, Zhuo Su, Liao Wang, Chen Guo, Zhaohu Li, Chengjiang Long, Zheng Lv, Jingxiang Sun, Chenyangguang Zhang, and Yebin Liu. Flexavatar: Flexible large reconstruction model for animatable gaussian head avatars with detailed deformation, 2025. URL https://arxiv.org/abs/2512.17717.

[49] Tobias Kirschstein, Simon Giebenhain, and Matthias Nießner. Flexavatar: Learning complete 3d head avatars with partial supervision. *arXiv preprint arXiv:2512.15599*, 2025.

[50] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024.

[51] Zijian Wu, Boyao Zhou, Liangxiao Hu, Hongyu Liu, Yuan Sun, Xuan Wang, Xun Cao, Yujun Shen, and Hao Zhu. Uika: Fast universal head avatar from pose-free images, 2026. URL https://arxiv.org/abs/2601.07603.

[52] Jiawei Zhang, Lei Chu, Jiahao Li, Zhenyu Zang, Chong Li, Xiao Li, Xun Cao, Hao Zhu, and Yan Lu. Bringing your portrait to 3d presence, 2025. URL https://arxiv.org/abs/2511.22553.

[53] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *arXiv preprint arXiv:2001.06782*, 2020.

[54] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

[55] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 42 (4), jul 2023. ISSN 0730-0301. doi: 10.1145/3592455. URL https://doi.org/10.1145/3592455.

[56] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venshtain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah,

James Booth, Taylor Koska, Kayla Haidle, Matt Andromalos, Joanna Hsu, Thomas Dauer, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. *NeurIPS Track on Datasets and Benchmarks*, 2024.

[57] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019. doi: 10.1109/CVPR.2019.00482.

[58] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d and 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1021–1030, 2017. doi: 10.1109/ICCV.2017.116.

[59] Zhenhui Ye, Tianyun Zhong, Yi Ren, Jiaqi Yang, Weichuang Li, Jiawei Huang, Ziyue Jiang, Jinzheng He, Rongjie Huang, Jinglin Liu, et al. Real3d-portrait: One-shot realistic 3d talking portrait synthesis. *arXiv preprint arXiv:2401.08503*, 2024.

[60] Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. GPAvatar: Generalizable and precise head avatar from image(s). In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hgehGq2bDv.

[61] Xiaochen Zhao, Jingxiang Sun, Lizhen Wang, Jinli Suo, and Yebin Liu. Invertavatar: Incremental gan inversion for generalized head avatars. In *ACM SIGGRAPH 2024 Conference Papers*, SIGGRAPH '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705250. doi: 10.1145/3641519.3657478. URL https://doi.org/10.1145/3641519.3657478.

[62] Zheng Ding, Cecilia Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[63] Yue Wu, Xuanhong Chen, Yufan Wu, Wen Li, Yuxi Lu, and Kairui Feng. Fastavatar: Towards unified and fast 3d avatar reconstruction with large gaussian reconstruction transformers. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=P7zBSCs4Xt.

[64] Lawrence E Arend Jr. Response of the human eye to spatially sinusoidal gratings at various exposure durations. *Vision Research*, 16(11):1311–1315, 1976.

[65] ITU-T. Subjective video quality assessment methods for multimedia applications. ITU-T Recommendation P.910, 2022. URL https://handle.itu.int/11.1002/1000/15005. Accessed on: 2026-05-05.

[66] Radek Danecek, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20311–20322, 2022.

A More Visualizations

In this section, we provide additional visualizations. Specifically, we show attention maps obtained without attention regularization in Fig. 10, attention maps with different attention regularizations in Fig. 11, and present more qualitative results for both self-reenactment and cross-reenactment on the VFHQ dataset in Fig. 12.

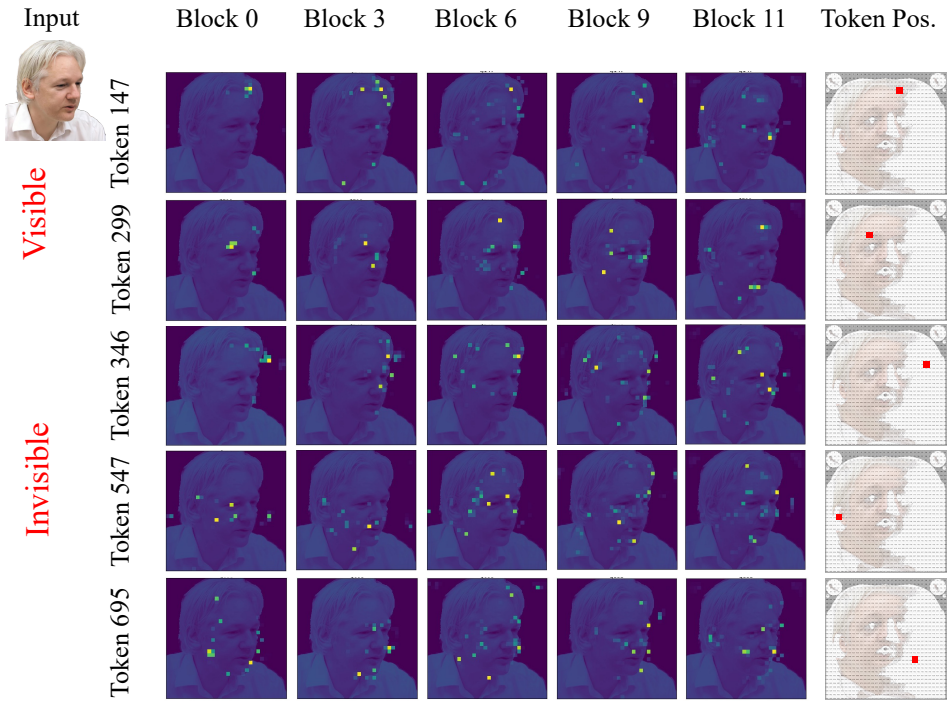

Figure 10: **Visualization of attention map activations** in a vanilla-trained network. Among the randomly selected tokens, token 147 and token 299 are visible in the input RGB image, whereas token 346, token 547, and token 695 are not. Please zoom in for detailed views and the corresponding token positions on the UV map.

B Details on the MV-VFHQ Dataset

We provide further details on the construction of our MV-VFHQ dataset. For each frame in every video sequence, we feed the image into Qwen-Image-Edit [21] with four camera control prompts: (1) turn the camera right to $45°$; (2) turn the camera right to $90°$; (3) turn the camera left to $45°$; (4) turn the camera left to $90°$. This process augments the original monocular VFHQ dataset with four additional novel views per frame.

We additionally annotate each frame with FLAME shape parameters, expression parameters, and camera pose parameters. To obtain these annotations, we employ an optimization-based FLAME tracker that minimizes the distance between labeled human face keypoints and the corresponding FLAME-derived keypoints, with initializations provided by EMOCA [66]. A sample visualization of the resulting dataset is shown in Fig. 13.

C Experimental Details

For Nersemble-v2 dataset, we leave sequences 017 EXP-5, 037 EXP-8, 055 EMO-4, 074 EXP-5, 134 EMO-1, 165 EXP-8, 221 EXP-5, 251 EMO-1, 264 EMO-1, 304 EMO-1 for evaluation, while all sequences from other ids for training. For Ava256 dataset, we leave all sequence for id PGO261,

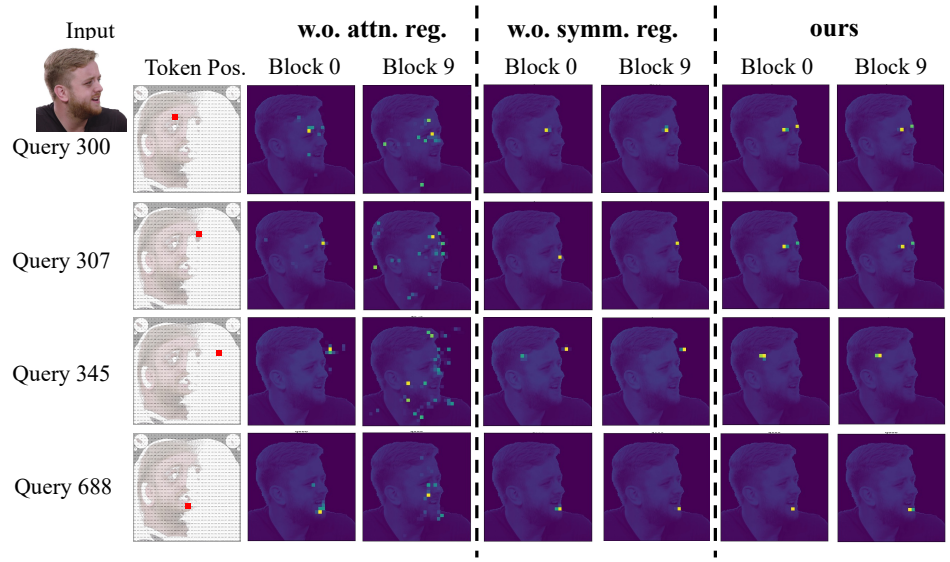

Figure 11: Visualizations on the attention activation maps of different attention regularization methods.

UHV563, BGR645, IBQ026, LCJ763, PDG961, APP152, PSV686, TCE049, YJF815, INQ807, KWL586 for evaluation and all others for training. For VFHQ dataset, we follow the authors' instruction for training and evaluation. We adopt Adam as optimizer, with a initialized learning rate of 2×10^{-4}, which linearly decreases to 1×10^{-7} at the end of training.

D Rendering Speed

We provide a comparison with main baselines in Sec. 4 on the peak GPU cost and time for creating avatars in Tab. 9, the results of which show that our method achieves an acceptable computational cost for real-world usages. Note that our setup follows LAM, which focus on fast animation and rendering speed without network post processing and are capable to deployed on mobile devices, while getting quite faster reconstruction speed compared with LAM.

Table 9: Comparisons with computational efficiency with baseline methods.

RTX A6000	Ours	LAM-20k	GAGAvatar	Portrait4D-v2	InvertAvatar	GPAvatar
Animation and Rendering Speed (FPS) ↑	<u>243.9</u>	**250.1**	47.6	8.1	4.3	4.8
Reconstruction Speed (ms) ↓	495.9	2780.7	85.2	244.5	164.2	255.6
FLOPs (G)	1701	7667	570	838	122	443
GPU Memory (GB)	6.675	3.613	2.885	3.261	4.605	5.028

E More Ablation on Token Fusion Strategy

We additionally ablate a learnable fusion logic switch (rather than explicitly controlled by the visibility map), fitting our model with a learnable mask to determine the fusion strategy (adopt current tokens/historical token/both fused via MLP). Results in Tab. 10 below show learnable gating is harder to optimize and underperforms our visibility-based fusion.

Table 10: Ablation study on the fusion module.

	TL-STD↓	TI↓	PSNR↑	SSIM↑	LPIPS↓
No Fusion	8.253	13.760	24.53	0.940	0.099
Learnable Gated Fusion	8.317	13.233	24.10	0.939	0.099
Ours	**7.253**	**12.805**	**25.33**	**0.942**	**0.096**



Figure 12: **Qualitative visualizations on VFHQ dataset.**



Figure 13: Samples of Multi-view VFHQ dataset.

F Broader Impacts

Positive Impacts. **FA-LAM** democratizes animatable 3D/4D head avatar creation from a single or few images, enabling applications in telepresence, VR/AR, creative industries, and assistive technology. Our analysis of attention mechanisms and task disentanglement also provides insights for future research in generative 3D vision.

Negative Impacts & Safeguards. Our method could be misused for deepfake generation or unauthorized identity reconstruction. To mitigate these risks: (1) all training data are from public datasets with informed consent; (2) code and models are released under a Non-Commercial Academic License prohibiting deceptive or privacy-violating uses.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions by proposing a symmetrical and semantic attention regularization (Section 3.2), a two stage training method (Section 3.3) and auto-regressive reconstruction (Section 3.4).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Please see Sec. 5.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper does not cotain any theoretical result.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experimental details are provided in Sec. 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: This article involves commercial collaboration; the code and data will be released upon review and approval.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Please see Appendix Sec. C and Sec. 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Please see Sec. 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see Sec. 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed and obeyed the NeurIPS Code of Ethics..

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please see Appendix. F.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We will properly give access to users with license only.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All of the used datasets are properly credited

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification:

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: This paper only includes public datasets with proper approvals.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM is only used for polish writing in this research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.