

# FADA: Fast Diffusion Avatar Synthesis with Mixed-Supervised Multi-CFG Distillation

## Supplementary Material

Table 1. Additional ablation study of mixed-supervised distillation on HDTF test set. All methods in this table are mixed supervised distillation with fixed loss weight and without multi-CFG distillation.

Method	NFE-D	IQA $\uparrow$	Sync-D $\downarrow$	FVD-R $\downarrow$	FID $\downarrow$	E-FID $\downarrow$
+ Fixed-0.0	18	3.762	7.909	25.06	21.47	1.462
+ Fixed-0.1	18	<b>3.875</b>	7.997	<u>13.93</u>	<b>19.01</b>	1.478
+ Fixed-0.2	18	<u>3.863</u>	7.934	<b>13.71</b>	19.94	<b>1.434</b>
+ Fixed-0.3	18	3.858	7.975	15.75	<u>19.54</u>	1.523
+ Fixed-0.5	18	3.844	<u>7.907</u>	16.41	19.98	1.500
+ Fixed-0.75	18	3.821	7.959	16.679	20.38	1.465
+ Fixed-1.0	18	3.812	<b>7.903</b>	24.53	21.42	1.494

Table 2. Counting numbers of different kinds of samples. Large movement(**Large Move**) and large pose(**Large Pose**) samples are worth learning while low image quality(**Low IQ**) and long silence(**Long Silence**) samples should be prevented from ground-truth supervision.

	Large Move $\uparrow$	Large Pose $\uparrow$	Low IQ $\downarrow$	Long Silence $\downarrow$
Small- $\mathcal{R}$	<u>2</u>	<u>3</u>	<b>1</b>	<b>4</b>
Medium- $\mathcal{R}$	<b>4</b>	<b>5</b>	<b>1</b>	<b>0</b>
Large- $\mathcal{R}$	0	1	3	6

## A. More Analysis of Mixed-Supervised Distillation

### A.1. Additional Hyper-Parameter Ablation

As shown in Table 1, we conducted additional hyper-parameter ablation experiments on mixed supervised distillation with fixed loss weight. The results indicate that setting the loss weight to 0.1 or 0.2 results in superior and comparable overall performance. We selected the loss weight of 0.2 since its E-FID shows good expressiveness.

### A.2. Sample Analysis for Adaptive Strategy

In the analysis presented in Section 3.2, we observed that within the moderate-quality training dataset  $\mathcal{B}$ , there are some samples that are worth learning, while others should be prevented from ground-truth supervision. As depicted in Figure 1, we identified samples with large movements and large poses as two categories of cases worth learning, while those with low image quality or extended periods of silence were classified as undesirable cases. Specifically, a sample is considered a large movement sample if its maximum movement distance exceeds half of the frame width, and a large pose sample if the rotation angle relative to the frontal face is greater than approximately  $45^\circ$ . A case is deemed

to have low image quality if its mean HyperIQA[7] value is below 35, and a long silence sample if the silence duration exceeds 75

Furthermore, we investigated the relationship between the ratio of ground-truth loss to distillation loss  $\mathcal{R} = L_{\text{gt}}/L_{\text{teacher}}$  and the sample distribution. We recorded the  $\mathcal{R}$  values of approximately 5000 samples during training, sorted them accordingly, and then randomly sampled Small- $\mathcal{R}$ , Medium- $\mathcal{R}$ , and Large- $\mathcal{R}$  samples. Specifically, Small- $\mathcal{R}$  corresponds to  $\mathcal{R}$  values ranging from 0 to 10, while Medium- $\mathcal{R}$  and Large- $\mathcal{R}$  encompass  $25 \sim 35$   $\mathcal{R}$  and  $> 80$   $\mathcal{R}$ , respectively. Each category consists of 10 sampling cases. As shown in Table 2, Medium- $\mathcal{R}$  demonstrates the most favorable learning outcomes, whereas Small- $\mathcal{R}$  is also valuable for learning due to the presence of large movement and large pose cases with minimal instances of low image quality or long silence. Conversely, the Large- $\mathcal{R}$  samples are deemed less beneficial for learning. This observation underscores the importance of the peak and dead threshold in our proposed adaptive method.

## B. Limitation

Our proposed FADA method primarily explores improvements in model inference speed, but it also has some objective limitations. The first limitation arises when the acceleration setting is more aggressive (e.g., Ours-Fast setting), there is still a slight decline in model performance. Two factors mainly contribute to this. Although CFG distillation significantly reduces computational complexity, it also brings about a certain decrease in expressiveness and accumulation of visual errors. Therefore, a careful balance between speed and performance is required in practical applications. The second limitation is that the multi-CFG distillation technique will increase training duration. It requires the teacher model to perform several inference processes (e.g., three times) during model training, resulting in a decrease in training speed. For example, in our implementation, using CFG distillation would reduce the training speed from approximately 4 seconds per step to around 10 seconds per step.

## C. More Analysis of Multi-CFG Distillation

Due to the limitations of static image presentation, please refer to our demo video **different\_cfg\_comparisons.mp4** in the supplementary materials for visualizing the multi-CFG distillation as the CFG scale changes. From the re-

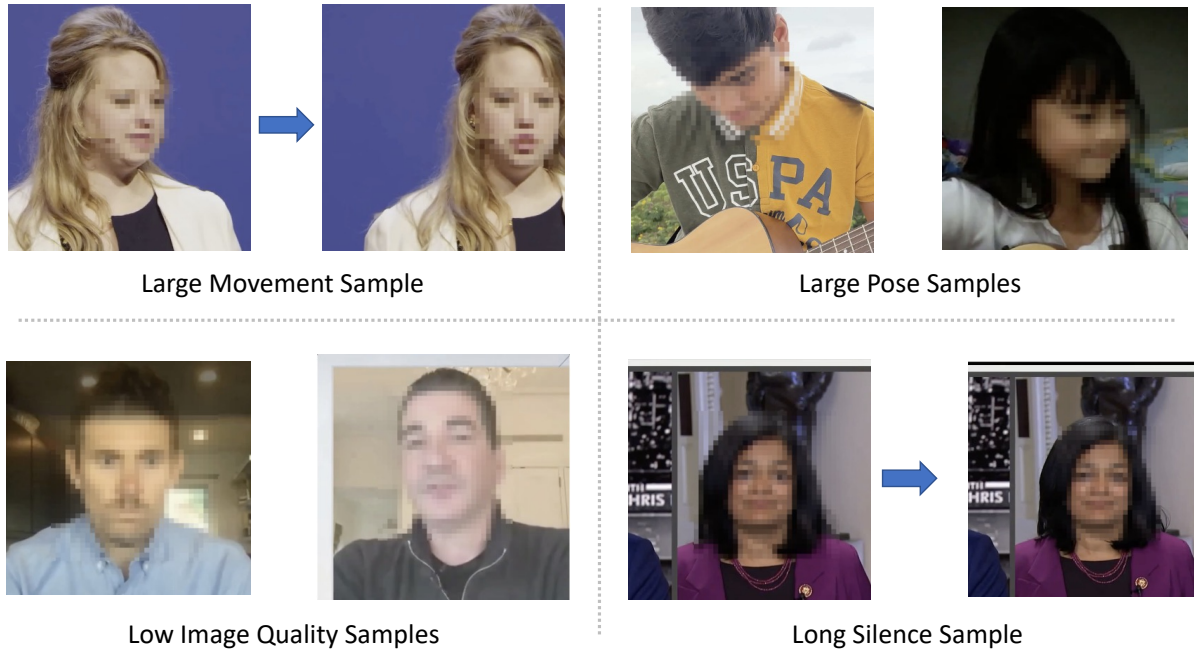


Figure 1. Visualised examples of different kinds of samples.

Table 3. Degradation of CFG distillation in different scenes.

Method	Normal		Long Duration		Emotional	
	FVD-R↓	E-FID↓	FVD-R↓	E-FID↓	FVD-R↓	E-FID↓
Ours-Balanced	<b>33.67</b>	<b>2.202</b>	<b>39.78</b>	<b>3.386</b>	<b>97.45</b>	<b>3.171</b>
Ours-Fast	35.01	2.604	49.79	3.596	105.4	3.530

sults obtained with different reference CFGs, we observe that as the reference CFG increases, the model’s identity-preserving capability strengthens. However, simultaneously, the model’s error accumulation becomes more noticeable, and the model’s expressive movements weaken. One possible reason for this is that both the reference and motion frames pass through a shared reference network, and the two are not entirely decoupled. Therefore, when the model enhances its reliance on the reference, it also reinforces its dependence on the motion frames, leading to increased error accumulation. Hence, we selected 2.0 as a compromise reference CFG scale for FADA with multi-CFG distillation.

Regarding the results obtained with different audio CFGs, we observe that the model shows sensitivity to this parameter change, but the sensitivity is somewhat reduced. Overall, as the audio CFG increases, the model’s lip movement intensity, lip amplitude, and head movement intensity exhibit a certain increase. After the audio CFG exceeds 6.5, the rising trend becomes less pronounced. Therefore, we chose 6.5 as the audio CFG scale value.

As shown in Table.3, we show the metric differences under different scenes, with changes in other unlisted metrics being less than 3%. For the long-duration and emotional scenes, we manually selected 20 samples for evaluation. The Ours-Fast model shows minimal loss when generating short ( $\leq 15$ s) clips but benefits from faster performance, making it suitable for such applications. We believe the fast model can be further improved by increasing the amount of moderate-quality data and optimizing it with our proposed training strategy.

## D. Details of Training Dataset Construction

### D.1. Data Collection Procedure

First, we obtained raw videos from several video platforms and by referencing publicly available dataset papers such as Panda 70M[1]. We manually defined tags, such as *Talk* and *Singing*, and categorized and filtered the video content. Videos meeting the criteria were retained. We strictly remove any personal information during the video collection process to comply with privacy requirements, utilizing only the RGB frames of the videos and the audio data itself.

Next, we performed simple preprocessing on these raw videos. First, we used the Mediapipe[5] tool for face detection, filtering out video frames that did not contain faces. In our experiments, only continuous face video segments longer than 2 seconds were included in the dataset, and we limited the segment length to no more than 10 seconds by

Table 4. Filtering rate of different filter processes

	Head	Move	SyncNet	Background	HyperIQA	Total
Filter Rate	17.7%	65.6%	27.0%	68.7%	87.7%	

Table 5. Some statistics of the training dataset

Gender		Age		Audio Type		Race			
M	F	Young	Mid	Old	Speech	Sing	African	Asian	Caucasian Others
52%	48%	34%	55%	11%	85%	15%	18%	26%	52% 4%

slicing the videos. Finally, we cropped the video segments to obtain portrait-framed face video clips, where the face position was determined based on the face location information provided by Mediapipe.

## D.2. Data Filter Strategy

Our data filtering strategy comprises multiple dimensions, including head movements, audio-visual sync, static background, image quality, and more. It is noteworthy that each filtering criterion has its own threshold, where stricter thresholds result in a high-quality dataset  $\mathcal{A}$  and looser thresholds lead to a moderate-quality dataset  $\mathcal{B}$ . Below, we will introduce some important filtering strategies utilized in FADA and filter rates for each filter process in Table.4:

1. **Head Movements Filter:** We leverage a pre-trained DWPose [11] keypoint extractor to obtain facial keypoints, computing the maximum relative movement distance of the nose keypoint,  $lmk_{nose}^{maxmove}$ , across the entire video clip. Data exceeding the threshold of  $lmk_{nose}^{maxmove}$  will be filtered out.

2. **Audio-Visual Sync Filter:** SyncNet [6] is employed to assess the audio-visual synchronization score, Sync-D, within the video clip. Data with Sync-D surpassing the threshold will be filtered out.

3. **Static Background Filter:** Initially, we use the Mediapipe [5] segmentation toolkit to segregate foreground and background within the video clip, generating a background mask  $\mathcal{M}_i$  for each frame. Subsequently, considering the background mask common to all frames in the video clip,  $\mathcal{M}_{share} = \bigcup \{\mathcal{M}_i\}$ , we calculate temporal frame differences based on the background segments corresponding to  $\mathcal{M}_{share}$ , denoted as  $I_i^{diff} = |I_{i+1}^{bg} - I_i^{bg}|$ . If any  $I_i^{diff}$  exceeds the predefined threshold, the respective video clip will be filtered out.

4. **Image Quality Filter:** HyperIQA [7] is utilized to extract the image quality level of each frame within the video. Data with an average image quality below the specified threshold will be filtered out.

## D.3. Dataset Meta Infomation

The quantity of our moderate-quality training dataset is about 1300 hours, which is less than Hallo3’s[3] 3500+ hours and Vlogger[14]’s 2000+ hours. We should note

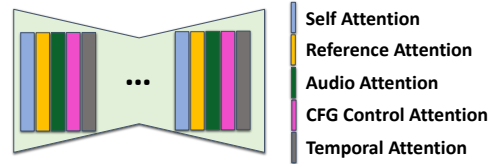


Figure 2. Structure graph of FADA student denoising network

that existing diffusion-based methods also commonly collect and filter data to build their private datasets. For instance, Loopy has about 176 hours, Echomimic 540 hours, and Hallo3 134 hours. We obtained about 160 hours of high-quality data, while the discarded data will be used for our proposed mixed supervised distillation.

Additionally, we provide statistics of our training dataset, including gender, age, race, and audio type in Table.5. These statistics are estimated by manually labeling 100 random samples in our moderate-quality training dataset. We can discover that the gender rate is basically balanced but there are some biases about age, audio type, and human race. It is worth noting that FADA focuses on utilizing non-selected filtered data, which precisely helps reduce data acquisition costs.

## E. More Implementation Details

We have introduced the design of the base model in Section 3.1 in the main paper, which is a streamlined model based on existing methods. It can be considered as built from Loopy by removing TSM and Audio2Latents, or from EMO by removing the Speed Layer and Face Locator. We provide a structure graph of the denoising U-Net in our student model in Figure.2. Each layer of the FADA student denoising network consists of one self attention layer, reference attention layer, audio attention layer, CFG control attention layer, and temporal attention layer sequentially.

During teacher pre-training with the high-quality dataset  $\mathcal{A}$ , the teacher model first underwent image pre-training to learn the relationship between reference images and target images with random movements. Subsequently, audio conditions and motion frames were incorporated for video training to learn audio-visual synchronization and temporal consistency. All training is conducted on 8 A100 GPUs, with 100,000 training steps per stage. The batch size is set to 8, and the gradient accumulation is 4. The length of motion frames is set to 4, with a motion frame dropout probability of 40%. The training optimizer used is AdamW [4], with a learning rate of  $1e-5$ . Our training and inference are conducted at 12.5 FPS for speedup, which will not influence the NFE-D in this paper. The NFE-D is simply the product of the denoising step number and the CFG runs number.

Table 6. Subjective evaluation results on openset.

Method	NFE-D	Quality $\uparrow$	Consistency $\uparrow$	Expressiveness $\uparrow$
Sadtalker[13]	-	0.02	0.02	0.02
Hallo[9]	80	0.08	0.15	0.07
EchoMimic[2]	60	0.10	0.11	0.16
V-Express[8]	50	0.07	0.03	0.02
Ours-Balanced	18	<b>0.53</b>	<b>0.43</b>	<b>0.66</b>
Ours-Fast	6	<u>0.20</u>	<u>0.26</u>	0.07

Table 7. Comparisons between DMD-v2 and PerFlow

Method	NFE-D	IQA $\uparrow$	Sync-D $\downarrow$	FVD-R $\downarrow$	FID $\downarrow$	E-FID $\downarrow$
PerFlow	18	<b>3.823</b>	<b>8.001</b>	<b>21.91</b>	21.26	<b>1.477</b>
DMD-v2	18	2.205	9.101	106.4	<b>19.62</b>	1.490

## F. Subjective Evaluation

For subjective evaluation, we randomly selected 20 test samples in the open set, encompassing diverse styles (real people, anime, humanoid crafts, and side faces) and various types of audio (speech, singing, rap, and emotional audio). Five participants with a certain level of knowledge and experience in the talking avatar synthesis task were invited to participate in the evaluation. Three metrics were assessed: **Quality**, **Consistency**, and **Expressiveness**, representing video quality, temporal consistency, and audio-visual expressiveness respectively. They were asked to select the best result number from the six compared results. All videos were independently shuffled beforehand.

As shown in Table 6, Ours-Balanced, with an inference speed of 18 NFE-D, demonstrated a remarkable advantage across all subjective metrics. On the other hand, Ours-Fast, with a speed of only 6 NFE-D, achieved the second-best results in Quality and Consistency. While the expressiveness pattern of Ours-Fast was similar but slightly weaker than that of Ours-Balanced, it is reasonable that it did not achieve an outstanding result in the pick-the-best evaluation.

## G. Comparisons with Adversarial Distillation Method

We did our best to re-implement DMD-v2[12] for the diffusion avatar tasks. However, the training procedure of the discriminator is unstable and often fails, leading to poor video quality as shown in Table 7. We suspect that more modifications or tricks are needed for a diffusion-based audio-driven model when using GAN loss. Hence, we finally select PerFlow[10] as our basic distillation method.

## H. Future Work and Application

The Ours-Fast model achieves about 3.1 RTF on a single A100 GPU, which can be further accelerated using TensorRT and model quantization. It can also be sped up

through tensor parallelism on an 8-A100 machine. We believe that real-time performance will be achievable with these engineering techniques. In terms of the application scope, FADA is not dependent on a specific backbone, so it can be extended to the latest 3D DiT methods.

## I. Ethical Concerns

With the rapid advancement of talking avatar synthesis technology, increasingly realistic generated faces have brought about an undeniable deepfake problem. We emphasize that, currently, FADA is solely intended only for research purposes. We suggest that efforts to mitigate the deepfake issue should focus on the following aspects: Firstly, incorporating clearly visible explicit AIGC watermarks in generated videos to make users acutely aware of the source of the generated content; secondly, integrating robust invisible watermarks into generated videos, enabling technological methods to identify the content even after explicit watermarks have been removed.

## References

- [1] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 2
- [2] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024. 4
- [3] Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with diffusion transformer networks. *arXiv preprint arXiv:2412.00733*, 2024. 3
- [4] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3
- [5] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Ubaweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 2, 3
- [6] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492, 2020. 3
- [7] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqui Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3667–3676, 2020. 1, 3



- [8] Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-express: Conditional dropout for progressive training of portrait video generation. *arXiv preprint arXiv:2406.02511*, 2024. 4
- [9] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Luc Van Gool, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024. 4
- [10] Hanshu Yan, Xingchao Liu, Jiachun Pan, Jun Hao Liew, Qiang Liu, and Jiashi Feng. Perflow: Piecewise rectified flow as universal plug-and-play accelerator. *arXiv preprint arXiv:2405.07510*, 2024. 4
- [11] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 3
- [12] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in Neural Information Processing Systems*, 37: 47455–47487, 2024. 4
- [13] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2023. 4
- [14] Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. Vlogger: Make your dream a vlog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8806–8817, 2024. 3