FantasyTalking2: Timestep-Layer Adaptive Preference Optimization for Audio-Driven Portrait Animation

Mengchao Wang*, Wang Qiang*, Fan Jiang † ‡, Mu Xu

AMAP, Alibaba Group

{wangmengchao.wmc, yijing.wq, frank.jf, xumu.xm}@alibaba-inc.com

Abstract

Recent advances in audio-driven portrait animation have demonstrated impressive capabilities. However, existing methods struggle to align with fine-grained human preferences across multiple dimensions, such as motion naturalness, lip-sync accuracy, and visual quality. This is due to the difficulty of optimizing among competing preference objectives, which often conflict with one another, and the scarcity of large-scale, high-quality datasets with multidimensional preference annotations. To address these, we first introduce Talking-Critic, a multimodal reward model that learns human-aligned reward functions to quantify how well generated videos satisfy multidimensional expectations. Leveraging this model, we curate Talking-NSQ, a large-scale multidimensional human preference dataset containing 410K preference pairs. Finally, we propose Timestep-Layer adaptive multi-expert Preference Optimization (TLPO), a novel framework for aligning diffusion-based portrait animation models with fine-grained, multidimensional preferences. TLPO decouples preferences into specialized expert modules, which are then fused across timesteps and network layers, enabling comprehensive, fine-grained enhancement across all dimensions without mutual interference. Experiments demonstrate that Talking-Critic significantly outperforms existing methods in aligning with human preference ratings. Meanwhile, TLPO achieves substantial improvements over baseline models in lip-sync accuracy, motion naturalness, and visual quality, exhibiting superior performance in both qualitative and quantitative evaluations.

1 Introduction

Audio-driven portrait image animation aims to synthesize realistic human speech videos from a reference image and driving audio. Recent advances (Chen et al. 2025a; Wang et al. 2025a; Gan et al. 2025; Lin et al. 2025b; Ji et al. 2025) have achieved notable improvements in facial expressions, motion diversity, and visual quality. However, critical challenges persist in generating high-fidelity animations, including difficulties in achieving perceptually natural lip-sync, the presence of obvious artifacts in complex local features (e.g., facial attributes and hand structures), and a failure to generate human motions aligned with user preferences. In language modeling and image generation, learning from human preferences (Ouyang et al. 2022; Rafailov et al. 2023; Zhou et al. 2023) has proven highly effective for enhancing generation quality and aligning models with user expectations.

However, applying such preference-driven alignment strategies to audio-driven portrait animation faces challenges. A key obstacle is the lack of large-scale, high-quality multidimensional preference data and reward models. Current methods like Hallo4 (Cui et al. 2025) and AlignHuman (Liang et al. 2025) primarily rely on manually annotated preference data, a costly and time-consuming process that severely limits data scale. This consequently constrains model generalization for complex motions, special pronunciations, and expression variations.

Another critical barrier is the difficulty in fine-grained preference alignment due to conflicts between multi-preference objectives. Most multidimensional preference optimization methods use linear scalarization (Li, Zhang, and Wang 2020; Liu et al. 2025b,a) to combine multidimensional rewards into a composite score, enabling reuse of standard Direct Preference Optimization (DPO) (Rafailov et al. 2023). Yet human preferences are complex and diverse, often involving conflicting goals (Wu et al. 2025): a sample with better motion may exhibit poorer lip alignment, while one with superior lip alignment may demonstrate inferior motion. This makes linearly combined rewards inadequate for addressing all preferences (Zhou et al. 2023).

To address these limitations, we first introduce Talking-Critic, a multidimensional video reward model designed to learn fine-grained human preferences, enabling the construction of large-scale, multidimensional preference datasets. Building upon this model, we evaluate outputs from four state-of-the-art portrait animation methods and curate Talking-NSQ, a large-scale preference dataset containing approximately 410k annotated samples. The dataset includes detailed annotations on Motion Naturalness (MN), Lip Synchronization (LS), and Visual Quality (VQ), capturing the key factors that users consider when assessing generated portrait videos.

Meanwhile, many studies (Liang et al. 2024; Wang et al. 2024a) reveal that diffusion models exhibit distinct inherent biases across denoising timesteps. The initial timesteps de-

*These authors contributed equally.

†Project leader

‡Corresponding author

termine the overall motion dynamics and structure, whereas the later timesteps are responsible for refining fidelity and fine-grained details. Furthermore, different layers of diffusion models contribute to different dimensions of the generated results (Avrahami et al. 2025; Chen et al. 2024). Some critical layers significantly impact content generation, while others affect clarity and detail representation. These observations suggest that both network layers and denoising timesteps are intrinsically linked to human preferences, highlighting the necessity for layer-wise and timestep-wise preference modeling for diffusion models.

Therefore, we propose a dual-stage Timestep-Layer adaptive multi-expert Preference Optimization (TLPO) for diffusion models. First, we employ a multi-expert approach by training lightweight LoRA (Hu et al. 2022) modules independently. Each module is specialized to optimize a specific dimension of the output, namely motion naturalness, lip-sync, and visual quality. Subsequently, we propose a fusion gate mechanism to dynamically adjust the weight distribution of each expert across timesteps and network layers. This achieves fine-grained multi-objective collaborative optimization, effectively resolving preference conflicts and overfitting to dominant preferences in traditional optimization, while enhancing overall expressiveness and human-likeness in audio-driven portrait animation generation. Figure 1 visually demonstrates the improvements of our method over the baseline across all evaluated dimensions. Our contributions can be summarized as follows:

- We propose Talking-Critic, a unified multimodal reward model that accurately quantifies the alignment between generated portrait animations and multidimensional human expectations.
- We introduce Talking-NSQ, a large-scale portrait animation preference dataset containing 410K samples, which systematically aligns user preferences regarding audio-visual synchronization, visual quality, and motion naturalness.
- We propose a novel preference alignment method, termed TLPO, that adaptively integrates multiple preference objectives across timesteps and network layers. Extensive experiments demonstrate that our approach significantly outperforms existing baselines across multiple metrics.

2 Related Work

2.1 Audio-Driven Portrait Animation

Portrait animation is a highly active research area that utilizes driving signals such as video (Wang et al. 2025b), audio (Chen et al. 2025b; Shen et al. 2023) or pose (Hu 2024) to generate vivid portrait animations from static images. Within the subfield of audio-driven portrait animation, early approaches (Ma et al. 2023; Wei, Yang, and Wang 2024; Zhang et al. 2023) leveraged 3D Morphable Models (3DMMs) (Egger et al. 2020) as an intermediate representation to bridge audio and video. While 3DMMs effectively capture facial geometry and expression variations, their expressive power is inherently limited, constraining the modeling of subtle expressions and the achievement of high realism. Driven by the advancement of diffusion models, recent methods (Kong et al. 2025; Gan et al. 2025; Chen et al. 2025a) bypass intermediate representations, directly synthesizing high-quality temporal sequences from audio and a static image. These methods typically build upon large-scale, pre-trained video generative models (Wan et al. 2025; Kong et al. 2024). By incorporating audio conditioning, they generate videos exhibiting audio-visual synchronization. However, although these methods successfully integrate the audio signal, the domain gap between the data used for pre-training and subsequent fine-tuning can compromise the generative quality of the base model to some extent. Our work aims to enhance audio-driven portrait animation frameworks through reinforcement learning, utilizes a carefully designed critic model and preference data to further improve audio-visual alignment, visual quality, and the naturalness of motion in the generated sequences.

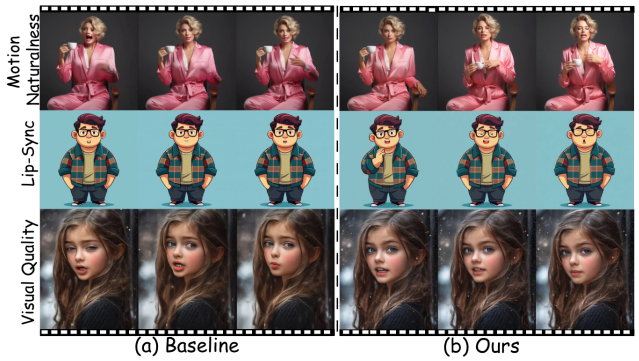

Figure 1: Comparison of the baseline and our method.

2.2 Human Preference Alignment

Human preference alignment, which aims to align model outputs with human preferences, has proven effective in both language (Dubey et al. 2024; Mehta et al. 2024) and vision models (Xu et al. 2023; Liu et al. 2025a). Among alignment methods, DPO (Rafailov et al. 2023) is widely adopted (Liu et al. 2025b, 2024). DPO optimizes the log probability ratio between preferred and non-preferred responses while constraining output deviations from the original distribution using a reference model. In the context of audio-driven video generation, both Hallo4 (Cui et al. 2025) and AlignHuman (Liang et al. 2025) employ preference optimization techniques to enhance portrait animation. However, these approaches rely heavily on large-scale, manually annotated preference data pairs and do not incorporate or optimize a critic model. Consequently, their ability to satisfy complex and diverse user requirements remains limited. Our method construct a critic model to evaluate preference data pairs, eliminating the need for extensive manual annotation. Furthermore, we introduce a dual-perception strategy across timesteps and network layers to more effectively integrate multiple preference objectives.

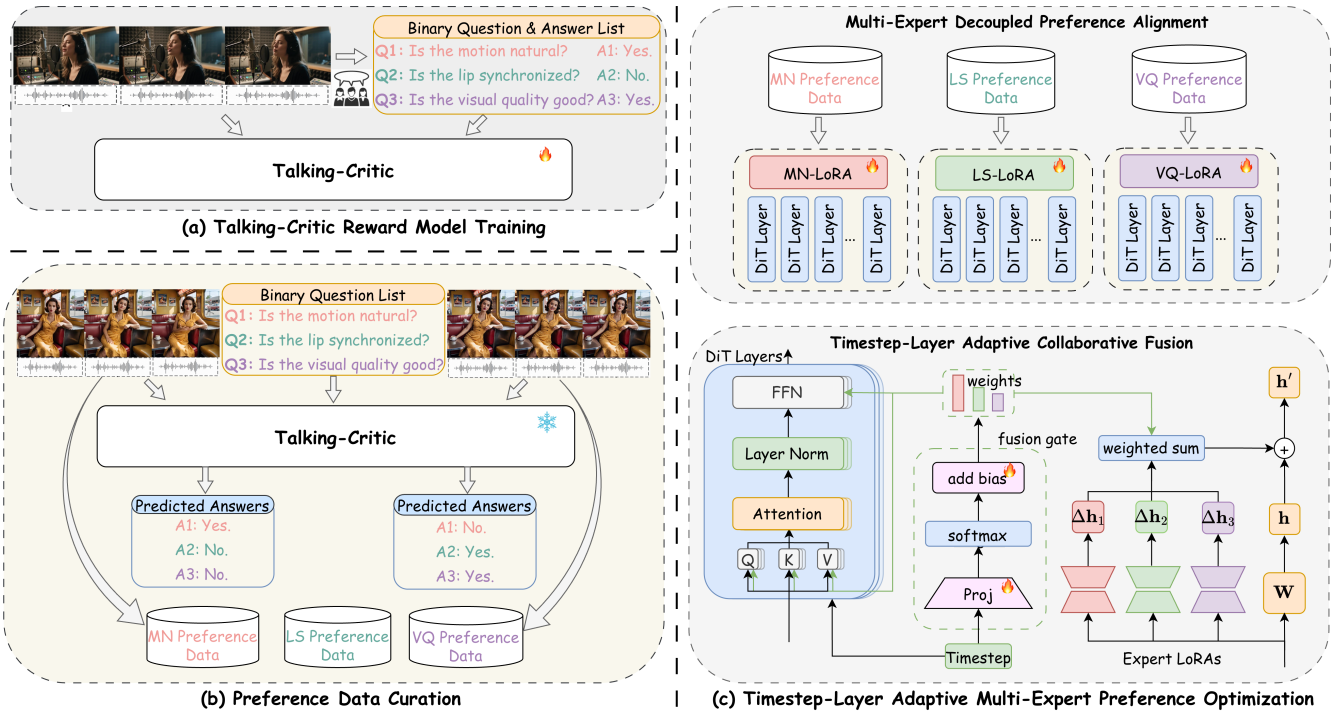

Figure 2: The Overview of FantasyTalking2.

3 Method

3.1 Preliminary

Base Model. We begin by presenting our foundational model for audio-driven portrait animation, known as FantasyTalking (Wang et al. 2025a). Built upon the pre-trained Wan2.1 (Wan et al. 2025) model, it comprises a 3D Variational Autoencoder (VAE) (Kingma, Welling et al. 2013) and a Latent Diffusion Transformer (DiT) (Peebles and Xie 2023). Specifically, the VAE encoder E converts input video data x into latent representations $z = E(x)$. The decoder D reconstructs latent tokens back to video space. During training, Gaussian noise ϵ is incrementally added to z via a forward process, generating noisy latent variables defined as: $z_t = tz + (1-t)\epsilon$, where $t \in [0, 1]$ is sampled from a logit-normal distribution. To adapt this image-to-video foundation model for audio-driven character animation, FantasyTalking utilizes Wav2Vec 2.0 (Baevski et al. 2020) to extract audio features, and inject via cross-attention layers into each DiT block, enabling audio-conditioned portrait driving.

DPO for Flow Matching. While diffusion models learn to generate human-meaningful images through reconstruction, this capability alone does not ensure alignment with human preferences and requirements. DPO provides a framework for aligning generative models with human preferences. By training on pairs of generated samples annotated with positive/negative labels, the model learns to assign higher probabilities to preferred outputs and lower probabilities to dispreferred ones. Flow-DPO (Liu et al. 2025a) applies DPO to flow matching (Lipman et al. 2022), by reformulating the DPO objective, the loss is defined as:

$$\mathcal{L} = -\mathbb{E}\Big[\log\sigma\Big(-\frac{\beta}{2}\big(L(x_t^w, t) - L(x_t^l, t)\big)\Big)\Big] \quad (1)$$

where β is a temperature coefficient, $L(x_t^w, t)$ and $L(x_t^l, t)$ respectively represent the losses for positive and negative parts, and x_t^w and x_t^l denote winning and losing samples respectively. This loss encourages the generation of preference-aligned samples, promoting coherent motion trajectories and producing more natural, realistic expressions.

3.2 Talking-Critic Reward

Previous video reward modeling approaches primarily leveraged vision-language models for training (He et al. 2024; Wang et al. 2024b; Xu et al. 2024). In contrast, our audio-driven portrait animation task requires multimodal inputs encompassing text, video, and audio modalities, rendering conventional vision-language models inadequate for reward modeling. Benefiting from recent advances in unified Vision-Audio-Language Models (VALMs), significant breakthroughs have been achieved in multimodal understanding and alignment. We employ the Qwen2.5-Omni (Xu et al. 2025) as our foundational model, which introduces TMRoPE, this position embedding method organizes audio and video frames into chronologically interleaved structures, enabling exceptional audio-visual alignment.

As shown in Figure 2(a), to harness the full potential of Qwen2.5-Omni for evaluating portrait animations, we adapt it into a reward model through specialized instruction fine-tuning. We construct a preference dataset encompassing three key dimensions: MN, LS, and VQ. During its

construction, we implement a rigorous balanced sampling strategy to ensure an equal number of positive and negative samples for each comparison, thereby enabling the model to learn human preferences without bias. The resulting fine-tuned reward model provides reliable guidance signals for downstream tasks, such as DPO.

3.3 Timestep-Layer adaptive multi-expert Preference Optimization

For multi-objective preference optimization, existing methods (Zhou et al. 2023; Liu et al. 2025b) obtain an aggregate score for each sample through various strategies, creating positive/negative pairs that reflect overall quality. This approach handles all preference objectives uniformly. However, it often leads to over-optimization in some dimensions at the expense of performance in others (Xu et al. 2024; Liang et al. 2025). Specifically, In human portrait animation, a sample ranked best overall may exhibit poor lip-sync accuracy, while a sample with the lowest overall score might excel in this regard. This conflict among fine-grained preferences impedes effective, granular alignment and hinders the model's ability to learn along less prominent dimensions. To address this, we propose a two-stage training strategy. As shown in Figure 2(c), the first stage learns decoupled preferences via a multi-expert alignment approach. Second, we introduce a timestep-layer adaptive fusion mechanism to effectively integrate these diverse preferences for robust multi-objective alignment.

Multi-Expert Decoupled Preference Alignment. To address preference competition conflicts, we first perform independent preference alignment through specialized experts, which involves three lightweight expert LoRA modules. The first is the Motion Naturalness Expert E_m, which aims to ensure fluid and natural body movements. The second is the Lip Synchronization Expert E_l, dedicated to optimizing the coordination between audio and visual cues. The third is the Visual Quality Expert E_v, designed to improve the fidelity of individual frames. Each expert module integrates into all linear layers of every DiT block. Since each specializes in a single dominant preference dimension, they achieve efficient convergence.

Noting lip synchronization exclusively concerns the mouth region, to reduce the difficulty of preference alignment and prevent introducing extraneous irrelevant preferences, we utilize MediaPipe to extract precise lip masks in pixel space, then project them to latent space through trilinear interpolation, forming our lip-focused constraint mask M. Consequently, the training loss $\mathcal{L}$ in Eq. 1 for the lip-sync expert is reweighted as:

$$\mathcal{L}_c = M \odot \mathcal{L} \tag{2}$$

For the motion naturalness expert LoRA and the visual quality LoRA, we perform the preference loss $\mathcal{L}$ across all pixel domains. In the end, we obtained three expert modules.

Timestep-Layer Adaptive Collaborative Fusion. Since each expert undergoes isolated dimension-wise optimization with segregated data, naively integrating them for inference may cause conflicting preferences among experts, degrading overall performance. Prior research has established that generative preferences differ across denoising timesteps (Liang et al. 2024; Wang et al. 2024a) and that DiT layers serve distinct functional roles (Avrahami et al. 2025; Chen et al. 2024). These findings motivate our design of a timestep-layer adaptive fusion strategy, enabling collaborative alignment of multi-expert modules.

Specifically, we employ a timestep-layer adaptive fusion gate to dynamically tune LoRA preference weights across DiT layers and timesteps. We integrate a lightweight, parameter-efficient fusion gate into all linear layers of DiT blocks. This gate modulates the influence of each LoRA module using the current denoising timestep t. As shown in Figure 2(c), for the l-th layer, the fusion gate takes the timestep embedding t_{emb} and projects it to fusion weights:

$$\boldsymbol{w}^l = \text{softmax}(W^l_{\text{gate}} t_{\text{emb}}) + \boldsymbol{b}^l \tag{3}$$

where $W^l_{\text{gate}} \in \mathbb{R}^{k\times d}$, $t_{\text{emb}} \in \mathbb{R}^{d\times 1}$ and $\boldsymbol{b}^l \in \mathbb{R}^{k\times 1}$. k is the number of expert LoRA modules. In our implementation, $k = 3$. W_{gate} and $\boldsymbol{b}^l$ are both learnable parameters. Crucially, since $k \ll d$ and $k \ll r$ (where r is the rank of LoRA), the fusion gate introduces only negligible parameters compared to the LoRA modules themselves.

Once the layer-level and timestep-wise weight vector $\boldsymbol{w}^l$ is produced, it is broadcast to every linear sub-layer inside DiT block l that carries LoRA adapters. The fusion of activations for such a layer is then performed as:

$$\boldsymbol{h}' = \boldsymbol{h} + \Delta\boldsymbol{h}\boldsymbol{w}^l, \tag{4}$$

where $\boldsymbol{h}$ represents the output of the l-th layer in the frozen DiT backbone, $\Delta\boldsymbol{h}$ denotes the delta from each expert LoRA at the same layer. During fusion training, we utilize full-dimension preference pairs, i.e., samples where the positive example is superior to its negative counterpart along all considered dimensions. During inference, the fusion gate dynamically adjusts weights $\{\boldsymbol{w}^l_i\}^k_{i=1}$ for per layer l and timestep t, enabling adaptive coordination of specialized LoRAs throughout the denoising process. This timestep-layer dynamic fusion continuously rebalances expert contributions, resolving conflicts and preventing single-metric dominance. By promoting collaboration over competition, it drives the model toward Pareto-optimal outputs (Deb 2011).

4 Experiments

4.1 Dataset Construction

Multidimensional Reward Data Collection. To train our Talking-Critic reward model, we constructed a high-quality, multi-dimensional human preference dataset. This dataset comprises both real and synthetic data, with binary preference annotations provided by professional annotators across MN, LS, and VQ dimensions. Specifically, we sourced approximately 4K real-world video clips from OpenHuman-Vid (Li et al. 2025). To maximize sample diversity, we also generated 6K synthetic videos using four state-of-the-art (SOTA) audio-driven portrait models (Chen et al. 2025a; Kong et al. 2025; Gan et al. 2025; Wang et al. 2025a) with random classifier-free guidance scales (Ho and Salimans 2022). Subsequently, all videos were evaluated by human

Method	HKC ↑	HKV ↑	SD ↑	Sync-C ↑	FID ↓	FVD ↓	IQA ↑	AES ↑
FantasyTalking	0.838	30.142	13.783	3.154	43.137	483.108	3.685	2.980
HunyuanAvatar	0.883	37.336	14.812	4.370	40.063	475.770	3.758	2.953
OmniAvatar	0.845	30.058	13.860	5.452	36.604	394.099	3.929	3.109
MultiTalk	0.857	40.371	14.683	5.668	37.839	362.591	4.027	3.184
Ours	**0.895**	**41.924**	**15.188**	**5.704**	**35.438**	**341.181**	**4.071**	**3.236**

Table 1: Quantitative comparisons with baselines.

annotators based on dimension-specific, binary-choice questions. Each sample was independently assessed by three annotators. In cases of disagreement, a fourth senior annotator was consulted to arbitrate and make the final decision. This meticulous process yielded a multi-dimensional preference dataset of approximately 10K samples. Furthermore, we created a validation set of 1K samples following the identical procedure.

Preference Data Collection. As shown in Figure 2(b), we propose a fully automated pipeline to construct a large-scale, multi-dimensional preference dataset **Talking-NSQ** for multi-expert preference training, culminating in 410K annotated preference pairs. Specifically, for each input audio clip and reference image, we generate candidate videos using the same set of SOTA models. Each model produces four video variants per input to ensure diversity. We then employ our pre-trained Talking-Critic to score these videos across the three distinct dimensions and construct corresponding positive-negative pairs. This dimensional decoupling allows a single video to contribute to multiple preference sets, significantly enhancing data utilization efficiency. This process yielded 180K pairs for motion naturalness, 100K for lip-sync accuracy, and 130K for visual quality.

Furthermore, for the timestep-layer adaptive fusion training stage, we constructed 18K full-dimension preference pairs. This was achieved by introducing controlled degradations to high-quality real videos. We randomly select the four SOTA models to synthesize new videos based on real videos. Then, we create preference pairs by matching the original high-quality real video as the positive sample with the newly generated, degraded video as the negative sample.

4.2 Reward Learning

Training Setting. We utilize Qwen2.5-Omni (Xu et al. 2025) as the backbone for our reward model, conducting supervised fine-tuning using the multidimensional reward data collected in Sec. 4.1. To adapt the model, LoRA is applied to update all linear layers within Qwen2.5-Omni Thinker while keeping visual and audio encoder parameters fully frozen. The training process employs a batch size of 32 with a learning rate of 2×10^{-6} over three epochs, requiring approximately 48 A100 GPU hours.

Evaluation Protocols and Baseline. We evaluate preference alignment accuracy of Talking-Critic using our curated 1K human-annotated test set, with comparisons against the baseline Qwen2.5-Omni model. We further employ Sync-C (Chung and Zisserman 2016) for lip-sync accuracy assessment, visual quality (IQA) score (Wu et al. 2023) for visual quality evaluation, and employ SAM (Kirillov et al. 2023) to segment foreground characters from frames while separately measuring optical flow scores (Teed and Deng 2020) to evaluate Subject Dynamics (SD) for character motion comparison. For Sync-C, aesthetics, and SD metrics, optimal decision thresholds are automatically determined by maximizing accuracy in distinguishing high-quality from low-quality samples.

Method	MN Acc(%)	LS Acc(%)	VQ Acc(%)
SD	78.42	-	-
Sync-C	-	72.34	-
IQA	-	-	68.85
Base Model	63.15	52.63	61.24
Ours	**92.50**	**86.94**	**94.67**

Table 2: Preference accuracy on test dataset.

Quantitative Results. Table 2 demonstrates that our reward model achieves significantly closer alignment with human preferences across all three dimensions compared to the base model. In contrast, existing quantitative evaluation methods can only be limited to evaluation in a certain dimension and cannot precisely align with human preferences. Especially, Sync-C tends to assign higher confidence to exaggerated lip movements, whereas human annotators consistently favor natural, fluid articulation—resulting in a clear misalignment between Sync-C scores and actual human preference.

4.3 TLPO Preference Optimization

Training Setting. We employ the DiT-based FantasyTalking as the backbone. All training is conducted on 16 A100 GPUs optimized via AdamW. All expert modules are optimized while keeping the backbone model frozen. In the first stage of TLPO, we train each expert LoRA module using single-dimension preference pairs, with a rank of 128. We set the learning rate to 10^{-5} and the β to 5000. The MN and VQ experts undergo 10 training epochs, while the LS expert trains for 20 epochs given its complexity. In the second stage of timestep-layer adaptive multi-expert fusion, we freeze all expert LoRA layers and train minimal-parameter fusion gates using full-dimension preference pairs, with learning rate 10^{-6} and DPO $\beta = 1000$ to balance holistic preference alignment over 5 epochs.

Figure 3: Visualization results.

Evaluation Protocols and Baseline. Evaluation is conducted on a benchmark test set following prior work (Wang et al. 2025a), which cover a wide range of scenes, initial poses, and audio content. For motion naturalness, we assess hand quality and motion richness with HKC and HKV (Lin et al. 2025a) , and quantify overall subject dynamics via the SD metric. Sync-C is used to measure the confidence of lip-sync. For visual quality, we adopt FID (Heusel et al. 2017) and FVD (Unterthiner et al. 2019) to assess overall generation quality and deploy q-align (Wu et al. 2023) to obtain fine-grained scores on visual quality (IQA) and aesthetics (AES). While the above metrics provide only a coarse proxy for motion naturalness, lip-sync, and visual quality, we conduct a user study for a more precise alignment check with human preference. We compare with the latest public state-of-the-art methods, including FantasyTalking (Wang et al. 2025a), HunyuanAvatar (Chen et al. 2025a), OmniAvatar (Gan et al. 2025) and MultiTalk (Kong et al. 2025), using empty prompts during inference for fair comparison.

Quantitative Results. Table 1 shows our method achieves state-of-the-art results across all metrics, generating outputs with more natural motion variations, significantly improved lip synchronization, and superior overall video quality. This performance stems from TLPO preference optimization mechanism, which enables superior understanding of fine-grained human preferences in portrait animation while dynamically determining the scope and weighting of preferences according to video model denoising requirements and DiT layer characteristics. This framework achieves precise alignment with the video model's preference outputs, consequently better satisfying practical application scenarios where comprehensive quality matters.

Qualitative Results. Figure 3 demonstrates comparative results across all methods. On the left, our TLPO model generates natural and dynamic full-body motions, while competing methods either produce static poses or exhibit exaggerated and distorted limb movements. The middle section highlights TLPO's robust lip-sync performance even in challenging long-range shots, where baselines exhibit severe desynchronization and misalignment. On the right, visual quality comparisons reveal rendering flaws in other methods. FantasyTalking produces noticeable artifacts, OmniAvatar suffers from overexposure and blurred details, and both HunyuanAvatar and MultiTalk lose significant facial detail. In contrast, TLPO preserves high visual fidelity and structural integrity, especially in complex facial regions.

User Studies. To further verify the alignment of the method we proposed with human preferences, twenty-four participants were asked to rate each generated video on a 0–10 scale across the three dimensions: MN, LS, and VQ. As shown in Table 4, our method achieves superior ratings compared to baselines, with relative improvements of 12.7% in lip synchronization, 15.0% in motion naturalness and 13.7% in visual quality over the strongest baseline (MultiTalk). This comprehensive evaluation highlights the superiority of our method in generating realistic and diverse human animations that align with human preferences.

4.4 Ablation Study

We explore the contribution of each proposed design through several ablation studies. First, to assess our fusion mechanism, we test a variant **without timestep-wise gating**, relying only on layer-wise fusion. We also compare our proposed fusion granularity against two alternatives:

Method	HKC ↑	HKV ↑	SD ↑	Sync-C ↑	FID ↓	FVD ↓	IQA ↑	AES ↑
w/o Timestep-wise Gating	0.857	38.240	14.733	5.509	39.733	357.963	3.994	3.122
Expert-level Fusion	0.879	37.183	14.871	5.565	37.430	353.243	4.029	3.150
Module-level Fusion	0.866	40.968	15.133	5.649	36.401	349.053	4.064	3.198
DPO	0.848	36.497	14.738	4.381	38.285	350.110	3.970	3.114
IPO	0.852	35.118	14.848	4.375	37.347	351.415	4.016	3.163
SimPO	0.864	37.935	14.937	4.546	37.010	350.935	4.054	3.217
LoRA Rank 32	0.886	40.765	15.163	5.546	36.462	346.629	4.039	3.149
LoRA Rank 64	0.890	40.824	15.169	5.695	35.469	343.494	4.053	3.225
LoRA Rank 256	0.893	**41.930**	15.183	**5.704**	35.440	**341.174**	4.069	3.232
Baseline	0.838	30.142	13.783	3.154	43.137	483.108	3.685	2.980
Ours	**0.895**	41.924	**15.188**	**5.704**	**35.438**	341.181	**4.071**	**3.236**

Table 3: Quantitative results of ablation on key designs.

Method	MN	LS	VQ
FantasyTalking	5.82	6.81	6.71
HunyuanAvatar	6.78	6.29	6.25
OmniAvatar	5.95	7.06	8.09
MultiTalk	6.81	7.14	7.40
Ours	**8.14**	**7.96**	**8.42**

Table 4: User Study Results.

expert-level fusion, which assigns one weight per expert, and **module-level fusion**, which assigns weights to individual linear layers (e.g., query projections). Furthermore, we establish a native **DPO** baseline by training a single LoRA on full-dimension preference pairs. We also substitute **IPO** (Azar et al. 2024) and **SimPO** (Meng, Xia, and Chen 2024) to evaluate alternative preference optimization methods. Finally, we investigate the effect of the LoRA rank by varying its size in the preference modules.

As shown in Table 3 and Figure 4, the variant without the timestep-wise gating shows a slight improvement over the baseline but underperforms our full TLPO method. This is because different timesteps in the diffusion process have distinct optimization requirements, necessitating a flexible adjustment of the corresponding preference injection. Both expert-level and module-level fusion result in suboptimal performance. This is because different DiT layers reside in distinct manifolds and serve divergent generative functions, allowing layer-level fusion to outperform expert-level fusion. In contrast, module-level fusion introduces an excessive number of new parameters, which complicates the training process and leads to suboptimal results.

DPO and its variants achieve comparable performance with moderate visual quality improvements, yet exhibit negligible enhancement in motion naturalness and lip-sync. Although we ensured superior samples in the preference data outperform inferior ones across all dimensions, the disparity in learning difficulty between objectives introduces training ambiguity. Consequently, models prioritize optimizing the more accessible fidelity objective to mitigate synthetic artifacts, while struggling to capture nuanced motion naturalness and lip-sync preferences, resulting in limited improvements. This validates the necessity of decoupling optimization for visual quality, lip synchronization, and motion naturalness due to their inherent competing objectives. The performance improves monotonically with the increase of LoRA rank and reaches saturation at about 128.

Figure 4: Qualitative results of ablation on key designs.

5 Conclusion

In this work, we address the challenge of balancing motion naturalness, visual fidelity, and lip synchronization in audio-driven human animation through TLPO – a novel multi-objective preference optimization framework for diffusion models. Our solution decouples competing preferences into specialized expert modules for precise single-dimension alignment, while a timestep-and-layer dual-aware fusion mechanism dynamically adapts knowledge injection throughout the denoising process. This effectively resolves multi-preference competition, enabling simultaneous optimization of all objectives without trade-offs to achieve comprehensive alignment. Qualitative and quantitative experiments demonstrate that FantasyTalking2 surpasses existing SOTA methods across key metrics: character motion naturalness, lip-sync accuracy, and visual quality. Our work establishes the critical importance of granular preference fusion in diffusion-based models and delivers a robust solution for highly expressive and photorealistic human animation.

References

Avrahami, O.; Patashnik, O.; Fried, O.; Nemchinov, E.; Aberman, K.; Lischinski, D.; and Cohen-Or, D. 2025. Stable flow: Vital layers for training-free image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 7877–7888.

Azar, M. G.; Guo, Z. D.; Piot, B.; Munos, R.; Rowland, M.; Valko, M.; and Calandriello, D. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, 4447–4455. PMLR.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.

Chen, G.; Zhao, X.; Zhou, Y.; Qu, X.; Chen, T.; and Cheng, Y. 2024. Towards Stabilized and Efficient Diffusion Transformers through Long-Skip-Connections with Spectral Constraints. *arXiv preprint arXiv:2411.17616*.

Chen, Y.; Liang, S.; Zhou, Z.; Huang, Z.; Ma, Y.; Tang, J.; Lin, Q.; Zhou, Y.; and Lu, Q. 2025a. HunyuanVideo-Avatar: High-Fidelity Audio-Driven Human Animation for Multiple Characters. *arXiv preprint arXiv:2505.20156*.

Chen, Z.; Cao, J.; Chen, Z.; Li, Y.; and Ma, C. 2025b. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2403–2410.

Chung, J. S.; and Zisserman, A. 2016. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, 251–263. Springer.

Cui, J.; Chen, Y.; Xu, M.; Shang, H.; Chen, Y.; Zhan, Y.; Dong, Z.; Yao, Y.; Wang, J.; and Zhu, S. 2025. Hallo4: High-Fidelity Dynamic Portrait Animation via Direct Preference Optimization and Temporal Motion Modulation. *arXiv preprint arXiv:2505.23525*.

Deb, K. 2011. Multi-objective optimisation using evolutionary algorithms: an introduction. In *Multi-objective evolutionary optimisation for product design and manufacturing*, 3–34. Springer.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Egger, B.; Smith, W. A.; Tewari, A.; Wuhrer, S.; Zollhoefer, M.; Beeler, T.; Bernard, F.; Bolkart, T.; Kortylewski, A.; Romdhani, S.; et al. 2020. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (ToG)*, 39(5): 1–38.

Gan, Q.; Yang, R.; Zhu, J.; Xue, S.; and Hoi, S. 2025. OmniAvatar: Efficient Audio-Driven Avatar Video Generation with Adaptive Body Animation. *arXiv preprint arXiv:2506.18866*.

He, X.; Jiang, D.; Zhang, G.; Ku, M.; Soni, A.; Siu, S.; Chen, H.; Chandra, A.; Jiang, Z.; Arulraj, A.; et al. 2024. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *arXiv preprint arXiv:2406.15252*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Hu, L. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8153–8163.

Ji, X.; Hu, X.; Xu, Z.; Zhu, J.; Lin, C.; He, Q.; Zhang, J.; Luo, D.; Chen, Y.; Lin, Q.; et al. 2025. Sonic: Shifting focus to global audio perception in portrait animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 193–203.

Kingma, D. P.; Welling, M.; et al. 2013. Auto-encoding variational bayes.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; et al. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*.

Kong, Z.; Gao, F.; Zhang, Y.; Kang, Z.; Wei, X.; Cai, X.; Chen, G.; and Luo, W. 2025. Let Them Talk: Audio-Driven Multi-Person Conversational Video Generation. *arXiv preprint arXiv:2505.22647*.

Li, H.; Xu, M.; Zhan, Y.; Mu, S.; Li, J.; Cheng, K.; Chen, Y.; Chen, T.; Ye, M.; Wang, J.; et al. 2025. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 7752–7762.

Li, K.; Zhang, T.; and Wang, R. 2020. Deep reinforcement learning for multiobjective optimization. *IEEE transactions on cybernetics*, 51(6): 3103–3114.

Liang, C.; Jiang, J.; Liao, W.; Yang, J.; Zeng, W.; Liang, H.; et al. 2025. AlignHuman: Improving Motion and Fidelity via Timestep-Segment Preference Optimization for Audio-Driven Human Animation. *arXiv preprint arXiv:2506.11144*.

Liang, Z.; Yuan, Y.; Gu, S.; Chen, B.; Hang, T.; Li, J.; and Zheng, L. 2024. Step-aware preference optimization: Aligning preference with denoising performance at each step. *arXiv preprint arXiv:2406.04314*, 2(5): 7.

Lin, G.; Jiang, J.; Liang, C.; Zhong, T.; Yang, J.; Zheng, Z.; and Zheng, Y. 2025a. Cyberhost: A one-stage diffusion framework for audio-driven talking body generation. In *The Thirteenth International Conference on Learning Representations*.

Lin, G.; Jiang, J.; Yang, J.; Zheng, Z.; and Liang, C. 2025b. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. *arXiv preprint arXiv:2502.01061*.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Liu, J.; Liu, G.; Liang, J.; Yuan, Z.; Liu, X.; Zheng, M.; Wu, X.; Wang, Q.; Qin, W.; Xia, M.; et al. 2025a. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*.

Liu, R.; Wu, H.; Zheng, Z.; Wei, C.; He, Y.; Pi, R.; and Chen, Q. 2025b. Videodpo: Omni-preference alignment for video diffusion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8009–8019.

Liu, Z.; Zang, Y.; Dong, X.; Zhang, P.; Cao, Y.; Duan, H.; He, C.; Xiong, Y.; Lin, D.; and Wang, J. 2024. Mia-dpo: Multi-image augmented direct preference optimization for large vision-language models. *arXiv preprint arXiv:2410.17637*.

Ma, Y.; Zhang, S.; Wang, J.; Wang, X.; Zhang, Y.; and Deng, Z. 2023. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv preprint arXiv:2312.09767*, 2(3).

Mehta, S.; Sekhavat, M. H.; Cao, Q.; Horton, M.; Jin, Y.; Sun, C.; Mirzadeh, I.; Najibi, M.; Belenko, D.; Zatloukal, P.; et al. 2024. Openelm: An efficient language model family with open training and inference framework. *arXiv preprint arXiv:2404.14619*.

Meng, Y.; Xia, M.; and Chen, D. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37: 124198–124235.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.

Shen, S.; Zhao, W.; Meng, Z.; Li, W.; Zhu, Z.; Zhou, J.; and Lu, J. 2023. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1982–1991.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, 402–419. Springer.

Unterthiner, T.; Van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; and Gelly, S. 2019. FVD: A new metric for video generation.

Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Wang, M.; Wang, Q.; Jiang, F.; Fan, Y.; Zhang, Y.; Qi, Y.; Zhao, K.; and Xu, M. 2025a. Fantasytalking: Realistic talking portrait generation via coherent motion synthesis. *arXiv preprint arXiv:2504.04842*.

Wang, Q.; Liu, M.; Hu, J.; Jiang, F.; and Xu, M. 2024a. Controllable Longer Image Animation with Diffusion Models. *arXiv preprint arXiv:2405.17306*.

Wang, Q.; Wang, M.; Jiang, F.; Fan, Y.; Qi, Y.; and Xu, M. 2025b. FantasyPortrait: Enhancing Multi-Character Portrait Animation with Expression-Augmented Diffusion Transformers. *arXiv preprint arXiv:2507.12956*.

Wang, Y.; Tan, Z.; Wang, J.; Yang, X.; Jin, C.; and Li, H. 2024b. Lift: Leveraging human feedback for text-to-video model alignment. *arXiv preprint arXiv:2412.04814*.

Wei, H.; Yang, Z.; and Wang, Z. 2024. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*.

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Liao, L.; Li, C.; Gao, Y.; Wang, A.; Zhang, E.; Sun, W.; et al. 2023. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*.

Wu, Z.; Kag, A.; Skorokhodov, I.; Menapace, W.; Mirzaei, A.; Gilitschenski, I.; Tulyakov, S.; and Siarohin, A. 2025. DenseDPO: Fine-Grained Temporal Preference Optimization for Video Diffusion Models. *arXiv preprint arXiv:2506.03517*.

Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; et al. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Xu, J.; Huang, Y.; Cheng, J.; Yang, Y.; Xu, J.; Wang, Y.; Duan, W.; Yang, S.; Jin, Q.; Li, S.; et al. 2024. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 15903–15935.

Zhang, W.; Cun, X.; Wang, X.; Zhang, Y.; Shen, X.; Guo, Y.; Shan, Y.; and Wang, F. 2023. 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*, 8652–8661.

Zhou, Z.; Liu, J.; Shao, J.; Yue, X.; Yang, C.; Ouyang, W.; and Qiao, Y. 2023. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. *arXiv preprint arXiv:2310.03708*.