REFREE-S2V: Towards Realistic Co-Speech Video Generation via Reward-Free RL and Multilevel Speech Guidance

Salaheldin Mohamed
Télécom Paris, Institut Polytechnique de Paris

M. Hamza Mughal
Max Planck Institute for Informatics

Rishabh Dabral
Max Planck Institute for Informatics

Christian Theobalt
Max Planck Institute for Informatics

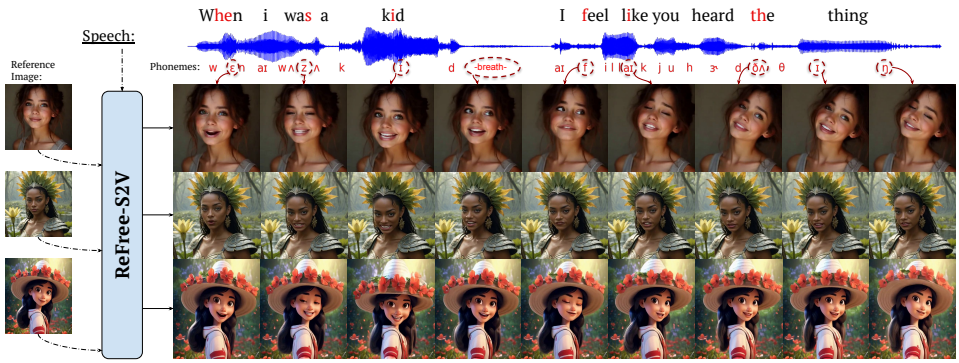


Figure 1: We present a speech-to-video generation framework that produces speech-synchronized, natural-looking portrait animations. Our approach produces accurate lip movements corresponding to speech's phonetic structure, and diverse and natural facial expressions and head movements.[2]

Abstract

Speech-driven talking character animation seeks to generate life-like portrait videos that convey natural conversation behavior, aligning facial motion with spoken audio. Although recent advances in video generation have substantially improved realism in video-based animation, achieving both accurate lip articulation and expressive behavior remains challenging. Existing approaches typically trade off precise phoneme-to-lip synchronization against dynamic facial expressions and head motion, yielding animations that are either accurate yet rigid, or expressive but poorly synchronized. We address this challenge by proposing ReFree-S2V, a flow-matching speech-to-portrait animation framework that builds upon a pretrained video generation model to achieve fine-grained speech articulation and high-level expressive cues in speech-driven portrait animation. This model introduces a multi-level speech representation capturing phonetic and prosodic information at both local and global granularities. These representations are selectively injected into transformer blocks via learnable level selectors, enabling both accurate lip synchronization and natural expressive motion. To achieve natural head movements, we further introduce a novel reward-free reinforcement learning scheme into flow-matching training to discourage perceptually implausible motion without relying

[2]Credit: initial first frame images are from `freepik.com`.

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

on handcrafted synchronization metrics or reward models, or the high cost of human preference annotation. Extensive experiments demonstrate that ReFree-S2V achieves state-of-the-art performance, significantly outperforming existing methods in both quantitative lip-sync accuracy and qualitative human evaluations of naturalness and expressivity. Project page: `https://salaheldinyoussry.github.io/ReFree-S2V/`.

1 Introduction

Virtual character animation has been a long-studied problem in computer vision and computer graphics. More recently, large video generation models Wan et al. [2025], OpenAI [2024], Yang et al. [2025] have spurred renewed interest in animating life-like human characters, directly, in video. Some of these models Gao et al. [2025] aim not only to generate human videos but also to align the animation with underlying speech to achieve talking character animation. Such speech-driven video generation enables a variety of applications, ranging from content creation, video conversational agents, and human-computer interaction. Given a reference portrait image and an input speech, the primary goal of talking character animation in video is to animate a life-like virtual character. To achieve this, every framework must address two main challenges: (1) speech-to-lip synchronization, which entails that lip movements not only match the phonetic structure of speech but also look perceptually natural, and (2) human-like expressivity, which requires generating dynamic and natural facial expressions with head motions driven by prosodic stress and semantic context in speech.

Recent diffusion-based works Cui et al. [2025a], Ji et al. [2025], Wang et al. [2025a] tackle speech-to-portrait animation using data-driven generative approaches. These methods typically build upon pretrained video transformers Yang et al. [2025], Blattmann et al. [2023a] and rely on supervised fine-tuning (SFT) over large-scale datasets Wang et al. [2025a], Cui et al. [2025a] to learn expressive speech-to-portrait animation. The resulting animations try to emulate natural head motions and eye movements, but struggle to transfer the rich phonetic structure of speech to lip movements. This is due to limitations in capturing both long- and short-term speech-animation dependencies, as well as the fact that SFT alone falls short of explicitly guiding models to learn fine-grained phoneme-to-lip synchronization and context-aware head movements. Therefore, existing approaches to speech-driven portrait animation struggle to offer a holistic solution to the aforementioned challenges.

With these considerations, we present REFREE-S2V — a speech-to-portrait animation framework designed for realistic generation . In particular, accurate lip synchronization with speech requires the model to focus on the low-level acoustic structure surrounding each spoken word. Complementing this, natural head movements and expression changes are driven by prosodic stresses and the overall semantic meaning of speech. To achieve both goals, we create a multi-level speech representation to condition the video generation process. This representation captures speech features at multiple levels of granularity, ranging from global prosodic cues to fine-grained phonetic structure. We inject these multi-level features to different transformer blocks through *level selectors*, which choose the appropriate contextual level for each block from among all available levels. This enables our approach to generate accurate lip movements driven by the speech structure and yield realistic and plausible head motions driven by prosodic cues.

Traditionally, RL-based generative modeling methods have relied on Proximal Policy Optimization (PPO) Ren et al. [2024], Wu et al. [2025], while more recent works have demonstrated strong performance using Group Relative Policy Optimization (GRPO) in vision tasks Xue et al. [2025], Wu et al. [2025]. In parallel, Direct Preference Optimization (DPO) has emerged as an alternative paradigm for preference learning. However, these approaches typically rely on either human annotations Rafailov et al. [2023], Liu et al. [2025] or pretrained reward models Wu et al. [2023], Wang et al. [2025b]. Leveraging human annotations for preference learning with (DPO) is expensive and difficult to scale. On the other hand, designing custom reward functions requires strong, often ill-defined assumptions about what constitutes realism, and typically emphasizes only specific aspects of animation, potentially introducing bias toward those isolated factors. Moreover, there is no single, holistic reward function that can reliably capture perceptual realism in generated videos; in practice, such definitions are inherently task-dependent and fail to generalize across diverse scenarios.

Our approach addresses these limitations by introducing a reward-free, RL-inspired fine-tuning strategy that improves perceptual realism and naturalness in generated videos. The model learns

realism in a self-supervised manner, without relying on human annotations or explicitly defined reward functions. Instead, it constructs a curriculum of negative samples alongside ground-truth data, isolating key realism factors and enabling learning through direct comparisons between real and synthesized samples. This formulation removes the need for both overly specified reward functions and costly annotation processes, relying instead on semantic differences between real data distributions and generated unrealistic samples.

Combining the base generative framework with the proposed multi-level speech representation and reward-free RL training strategy yields natural talking portrait videos with improved speech-to-lip synchronization and high expressivity.

2 Related Works

2.1 Diffusion-based Video Generation.

Diffusion models have recently become the dominant paradigm for high-fidelity video generation. Early methods primarily leverage UNet architectures; for instance, AnimateDiff Guo et al. [2023] inflates text-to-image models by incorporating temporal layers to capture motion dynamics, effectively exploiting pretrained image generators for spatial quality. Other approaches, such as Stable Video Diffusion Blattmann et al. [2023b], are trained as dedicated video generation models. More recent approaches based on DiT architectures advance the field by integrating 3D VAEs with transformer-based sequence modeling Gao et al. [2025], Yang et al. [2025], enabling fine-grained control over identity, expressions, and motion. Parallel efforts in preference-aware generation, such as Diffusion-DPO Wallace et al. [2024] and VideoDPO Liu et al. [2025], extend the DPO paradigm to align outputs with human judgments, optimize motion dynamics, and refine perceptual quality. While these methods demonstrate impressive video synthesis capabilities, they are largely agnostic to audio inputs and the specific challenges of speech-driven portrait animation. Our work builds on these advances by integrating multi-level speech conditioning and reward-free human preference alignment to generate realistic and expressive talking portraits.

2.2 Audio-driven Portrait Animation.

Audio-driven talking head generation focuses on synthesizing realistic facial motion from speech. Early methods Zhou et al. [2020], Gururani et al. [2023], Ng et al. [2022] rely on 3D intermediate representations, such as facial landmarks and 3D Morphable Models, to guide lip synchronization, but often produce limited expressions and constrained head motions. Subsequent end-to-end approaches, including Wav2Lip Prajwal et al. [2020], SadTalker Zhang et al. [2023], and FantasyTalking Wang et al. [2025a], leverage audio-visual learning to improve lip alignment and overall expressivity. Transformer- and diffusion-based frameworks, such as Hallo3 Cui et al. [2025a], VASA-1 Xu et al. [2024], and FantasyTalkingWang et al. [2025a], further enhance realism by modeling temporal context and prosodic cues through cross-attention mechanisms to achieve precise audio-to-face alignment. Nevertheless, existing approaches face persistent challenges: maintaining fine-grained articulatory synchronization, capturing high-frequency facial dynamics that extend beyond the lip region, while modeling both local and global speech-motion correlations. To this end, our method introduces multi-level speech representations and temporal motion modulation within a flow-matching framework, allowing concurrent lip-sync accuracy and natural expressive motion in high-fidelity portrait videos.

3 Method

Our framework REFREE-S2V builds on the Wan 5B pretrained video generation model Wan et al. [2025]. Similar to prior approaches such as fantasytalking (which uses Wan 14B) Wang et al. [2025a] and Hallo3 (based on CogVideoX-5B Yang et al. [2025]), we adopt models from a similar class.

We introduce our multi-level speech representation in Sec. 3.1, which is then injected into the DiT blocks through multi-head FiLM-based speech injection blocks (Sec. 3.2). At this stage, the resulting framework is trained with a flow-matching objective by applying low-rank adaptation (LoRA) on the pretrained DiT blocks and fully optimizing Speech Injection blocks and the multi-level speech encoder. We then introduce a second optimization stage based on the reward-free RL finetuning

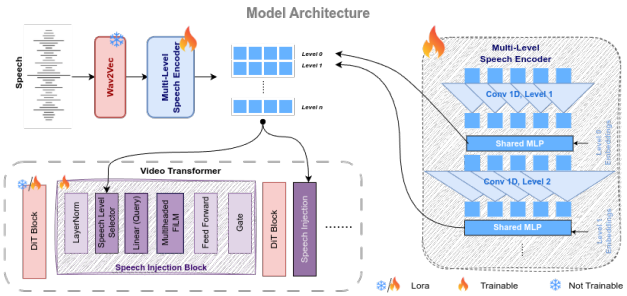


Figure 2: We present a speech-to-video generation framework that produces speech-synchronized, natural-looking portrait animations. The model incorporates multi-level speech guidance to jointly leverage local and global cues for richer neighborhood representations.

strategy (Sec. 3.3), which enhances perceptual realism and natural speech-to-animation transfer. The complete framework is illustrated in Fig. 2 and Fig 3.

3.1 Multi-level speech Representation

We observe that the input speech signal inherently carries both global and local information. For example, when posing a question or saying 'no', humans tend to perform a preparatory gesture like head-shaking. Here, the head-turning motion begins before the speech is articulated. This implies that video frames preceding the onset of speech must anticipate and initiate the motion, meaning that future speech information is necessary for accurate generation (refer to Fig. 6). Relying solely on a simple frame-to-speech alignment can often constrain the naturalness and expressiveness of motion. Additionally, not all transformer layers require access to the full global context; some layers may focus on local features, while others capture global patterns or a combination of both. Motivated by these observations, we develop the proposed multi-level speech representation.

First, the speech input is processed by a multi-level encoder, as shown in Fig. 2, Each level consists of a CNN layer with kernel size K_l and a shared MLP across all L levels, which is additionally conditioned on an embedding for the level-index. The kernel size K_l increases from level 0 to L, effectively expanding the temporal context window from local to more global. The resulting activations $\mathbf{f}_l$ from all the levels are concatenated and provided as input to each speech-injection block in the base video diffusion model.

3.2 Speech Injection

For each speech-injection block, the multi-level speech encoding is combined using a weighted gating mechanism:

$$\mathbf{A} = \sum_{i}^{L} \big[\text{softmax}(\mathbf{W})\big]_i \mathbf{L}_i \tag{1}$$

where $\mathbf{L} \in \mathbb{R}^{L \times d}$ represents the outputs of the L encoder levels, $\mathbf{W} \in \mathbb{R}^{L}$ are learnable selection weights, and d is the feature dimension. Next, we employ a multi-headed FiLM layer to condition the video generation model with the speech conditioning $\mathbf{A}$. Unlike approaches Wang et al. [2025a] that use windowed cross-attention, this multi-head FiLM design provides more representational flexibility (refer to Tab. 3, 1). While each FiLM head learns shift and scale transformations, applying windowed cross-attention can be interpreted as only learning shifts, limiting its expressiveness. Complete process in the multi-head film layer is presented in Algorithm 1. Formally, each multi-head FiLM layer gets hidden states $\mathbf{H}_{in}$ of video tokens from the preceding layer, and speech conditioning $\mathbf{A}$.These hidden states of sequence length S are projected into queries by a linear layer $\mathbf{W}_q$. and Per-head FiLM parameters γ_k and β_k are learned through a linear transformation $\mathbf{W}_c^{(k)}$ on speech conditioning.

Algorithm 1 Multi-Head FiLM Layer

Input: $H_{\text{in}} \in \mathbb{R}^{S \times d_{\text{in}}}$, $\mathbf{A} \in \mathbb{R}^{L \times d}$, K
Output: $F_{\text{mod}} \in \mathbb{R}^{S \times d}$
$H_{\text{proj}} \leftarrow W_Q(H_{\text{in}})$ # $(S \times d)$
$H, \mathbf{A}' \leftarrow \text{reshape_to_heads}(H_{\text{proj}}, \mathbf{A})$
for $k = 1$ **to** K **do**
    $(\gamma_k, \beta_k) \leftarrow \mathbf{W}_c^{(k)}(\mathbf{A'}_k); \mathbf{F}_k \leftarrow \gamma_k \odot H_k + \beta_k$
end for
$\mathbf{F}_{\text{mod}} \leftarrow \text{Concat}(\mathbf{F}_1, \ldots, \mathbf{F}_K)$

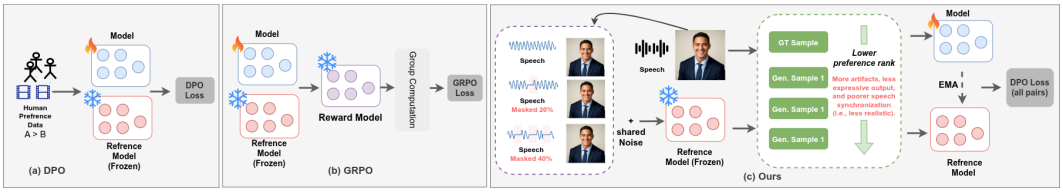


Figure 3: We present a novel reward-free reinforcement learning method based on a guided ranking strategy, which generates informative negative samples alongside real ones during training. Unlike GRPO (b), our approach does not require a reward model, and unlike DPO (a), it does not rely on human intervention to generate preference data.

3.3 No-Reward RL Fine-tuning

3.3.1 Motivation

While training our base framework with multi-level speech representation can improve animation quality, The objective function still relies on the flow-matching objective to learn speech-to-animation transfer. This objective provides no explicit guidance for distinguishing natural from unrealistic motion, as the trajectory-matching L2 loss operates pixel-wise and is agnostic to physical plausibility. To address this, recent generative frameworks adopt reinforcement learning-driven finetuning methods such as direct preference optimization (DPO) Wallace et al. [2024], Liu et al. [2025] or group relative policy optimization (GRPO) Xue et al. [2025], Park et al. [2025] for better physical plausibility or human preference alignment. These approaches utilize reward models or human annotations Cui et al. [2025b] (as visualized in Fig. 3) to perform preference optimization on the pretrained models.

However, large-scale human annotation becomes infeasible, and reward models rely on specific assumptions that introduce biases. More fundamentally, the core issues lie in the use of reward models: what kind of reward should be used? How can we formally define what constitutes a realistic generation or a physically plausible one through a scalar reward signal? These open questions turn this process into an ill-defined one. For example, a naïve implementation might use SyncNet Chung and Zisserman [2016a] as a reward signal to improve speech-lip synchronization. However, this introduces a strong bias toward lip matching while neglecting other perceptual qualities such as expressiveness, image quality, head motion, and eyebrow movement. Comprehensively judging every such aspect is non-trivial, and even aggregating multiple pretrained reward models fails to capture the full picture, a task that would ultimately require a true world model, which does not yet exist.

Considering these limitations, we instead introduce a novel approach that eliminates the need for both a reward model and human annotators. During the second-stage optimization, we teach the model realism by synthesizing auto-ranked negative samples that isolate the specific aspects the model needs to improve.

3.3.2 The Emergence of Realism

For each ground-truth (GT) sample, we synthesize N progressively speech-masked samples to form the ordered set $\mathcal{S} = \{s_0, s_1, \ldots, s_N\}$, where s_0 represents the original GT and s_i denotes a synthesized sample with a higher masking ratio than s_{i-1}. We adopt a random hard-masking strategy, using small windows scattered across the speech sequence. This produces outputs that are

progressively less expressive and less synchronized, with increasing visual artifacts arising from the distribution gap between masked and unmasked speech. Crucially, the degree of degradation scales with the masking ratio: the more aggressively the speech is masked, the more unrealistic, unsynchronized, and artifact-prone the output becomes. By keeping the initial noise consistent across all synthetic samples, we can automatically rank them based on the masking percentage. These masked samples also serve as negative examples relative to the GT sample s_0. This approach allows us to automatically synthesize N unrealistic samples and leverage the gap between the ground-truth (GT) sample and the synthetic samples to encourage the model to learn realistic features like image quality. Realism and physical plausibility emerge naturally from this discrepancy between the synthetic outputs and the GT data. This design isolates the effect of non-realism, giving the model a clear and focused signal to learn its significance.

We use the auto-ranked tuple of samples as guidance. Because they share the same initial noise, they are effectively identical samples with varying levels of realism degradation. We then apply DPO over all pairs (s_i, s_j) with $i < j$, giving the total loss:

$$\begin{aligned} \mathcal{L}_{\text{DPO}}^{\text{all}} &= -\mathbb{E}_{t\sim\mathcal{U}(0,1)} \sum_{i<j} \log \sigma\Big(- \beta\omega_t[\Delta_\theta - \Delta_{\text{ref}}]\Big), \\ \text{where } \Delta_u &= \|v_i - u(s_i^t, t)\|^2 - \|v_j - u(s_j^t, t)\|^2. \end{aligned} \tag{2}$$

Here, v_i, v_j are velocity targets for win/loss samples; v_θ, v_{ref} are model and reference predictions; β is temperature; and ω_t is a timestep weight. We also update the reference model using an exponentially moving average (EMA), gradually shifting the distributions of both models toward more realistic generations. This enables the model to learn realism in a self-supervised manner, relying only on real speech–video pairs, without human annotations or reward models (Figure 3).

4 Experiments

Datasets. Our training data consists of three datasets: Hallo3 Cui et al. [2025a], CelebV-HQ Zhu et al. [2022], and the Seamless Interaction dataset Agrawal et al. [2025]. Since the Seamless Interaction dataset consists of studio-quality footage captured in a controlled environment, its high audio clarity makes it valuable for learning accurate speech–lip correspondence. However, because it does not reflect in-the-wild conditions, we include it in the training set with a lower sampling probability. For evaluation following prior methods, we incorporate around 200 clips from the HDTF dataset Zhang et al. [2021], and for the in-the-wild setting, we test our model on a collection of in-the-wild samples from the Hallo3 dataset Cui et al. [2025a] comprising 50 randomly selected videos.

Baselines. We evaluate our model against ground-truth and recent state-of-the-art methods, including Hallo3 Cui et al. [2025a], which is based on the CogVideoX Yang et al. [2025] video generation backbone with cross-attention for speech conditioning; FantasyTalking Wang et al. [2025a], which leverages a larger Wan 14B Wan et al. [2025] model and additional facial encoding for speech-to-portrait animation; and Sonic Ji et al. [2025], built on Stable Video Diffusion Blattmann et al. [2023a] with audio injected via cross-attention mechanisms.

4.1 Perceptual Evaluation.

Qualitative Results. We visualize samples from the HDTF dataset in Figure 4, comparing our generations against competing methods and the Ground Truth (GT). Visually, we observe accurate phoneme-to-lip-shape matching, which demonstrates our model's superior lip-sync capability. Furthermore, as demonstrated in Figure 4, our model generates highly diverse and expressive head movements. This is particularly evident in Figure 4, where our method exhibits noticeably richer and more varied facial expressions, particularly in head pose and eyebrow movements, compared to the more limited range of other methods.

User Study. ReFree-S2V outperforms all baselines across all metrics, showing the most significant gains in Lip Synchronization and Video Realism. Notably, when compared to the ground truth (GT), our method achieves near-parity in lip synchrony while maintaining strong results in diversity and consistency. Furthermore, in terms of visual realism, our generated samples are preferred over the ground truth in approximately 24% of cases.

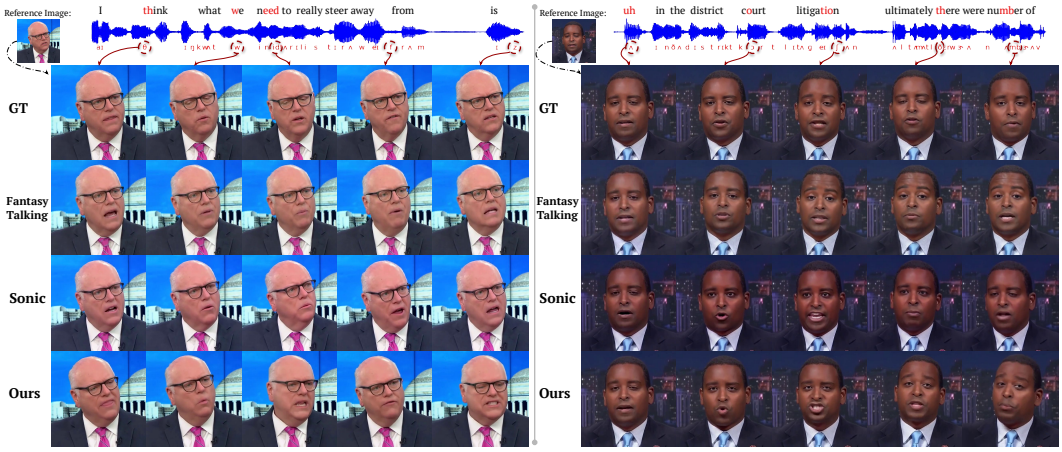


Figure 4: Qualitative results on HDTF dataset Zhang et al. [2021]. Note the accurate phoneme-to-lip synchronization on phonemes like /θ/ (think) and /f/ (from).

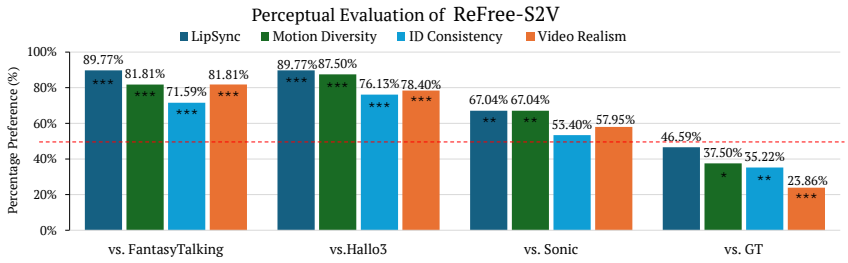


Figure 5: **Perceptual evaluation of ReFree-S2V.** We report pairwise preference percentages of our method against four baselines. The red line indicates the chance level (50%). Here, *: ($p < 0.05$), **: ($p < 0.01$), and ***: ($p < 0.001$).

4.2 Quantitative Evaluation

Evaluation Metrics. To assess motion–video alignment, we adopt the widely used Sync-C and Sync-D metrics Chung and Zisserman [2016b], which measure the confidence and temporal distance of audio–visual synchronization. These metrics reflect how well generated human motions align with speech, particularly in highly dynamic segments. For facial expression quality, we employ *mesh-FID*. Specifically, we extract 3D facial meshes using MediaPipeGoogle [2020] and compute the Fréchet Inception Distance (FID) Heusel et al. [2017] directly on the vertex representations without normalizing them into image space. This metric quantifies discrepancies between generated and ground-truth videos in terms of facial expressions, as well as head movement and pose.

To evaluate overall visual realism and alignment with the reference videos, we report both FID Heusel et al. [2017] and FVD Unterthiner et al. [2019]. Finally, to measure the degree of motion dynamics in the generated videos, we adopt the *Dynamic Degree* metric Huang et al. [2024]. This metric penalizes static videos with minimal motion, encouraging expressive results that include not only accurate lip movements but also coherent facial responses.

Setup. To ensure fair comparison across different methods, we resize all videos to the lowest common resolution, 480×480, and standardize the frame rate to 15 FPS. We evaluate on the standard HDTF dataset Zhang et al. [2021]. Since HDTF is generally considered an easier benchmark, we also follow prior work by evaluating on a challenging "in-the-wild" subset from the hallo3 dataset.

As shown in Table 1, our method establishes new state-of-the-art performance across most metrics. Notably, we achieve the lowest FID, the best Sync-C and Sync-D scores, and the largest improvement in dynamic degree, demonstrating superior lip synchronization, diversity, expressiveness, and overall fidelity to ground truth. The only exception is FVD, where we remain highly competitive. In

Table 1: Performance comparison with existing approaches on the HDTF dataset.

Method	Sync-C ↑	Sync-D ↓	FID ↓	FVD ↓	Mesh-FID ↓	Dyn. Deg. ↑
Sonic	7.406	7.444	13.242	72.142	0.0507	0.238
FantasyTalking	3.518	11.072	16.488	122.167	0.1094	0.021
Hallo3	5.949	9.181	14.656	**70.372**	0.0336	0.516
$\text{Ours}_{\text{w/o RL}}$	7.687	7.160	12.319	88.015	0.0290	0.534
Ours	**7.999**	**7.011**	**11.643**	86.182	**0.0099**	**0.559**

Table 2: Performance comparison with existing approaches on the in-the-wild samples from the Hallo3 dataset. *Hallo4 is conditioned on GT face mesh and skeleton.

Method	Sync-C ↑	Sync-D ↓	FID ↓	FVD ↓	Mesh-FID ↓	Dyn. Deg. ↑
Sonic	5.990	8.224	34.979	270.047	0.1340	0.22
FantasyTalking	3.001	11.294	37.877	277.888	0.2249	0.02
Hallo3	5.041	9.669	38.762	281.060	0.1204	0.36
OmniAvatar	6.063	8.636	71.222	268.140	0.4679	0.28
Hallo4*	5.435	9.279	39.238	**178.847**	**0.0265**	0.58
$\text{Ours}_{\text{w/o RL}}$	6.064	8.096	36.921	277.768	**0.1017**	**0.76**
Ours	**6.651**	**7.781**	**33.345**	**258.755**	0.1304	0.58

the unconstrained "in-the-wild" setting (Table 2), our method also establishes new state-of-the-art performance across all metrics. Notably, our RL-based variant achieves the best scores for FID, FVD, Sync-C, and Sync-D, while the non-RL variant demonstrates the highest dynamic degree and Mesh-FID. We attribute the RL version's slightly lower dynamic degree (compared to the non-RL version) to the training process: RL effectively dampens unrealistic, explosive motions, grounding the output in realism. This hypothesis is supported by the RL version's superior FID and FVD scores. In comparison to a method trained with DPO (Hallo4) with human annotation, our method yields a noticeable improvement while being more scalable and efficient. To summarize, our method sets a new benchmark across both settings, delivering significant improvements in realistic, expressive, and synchronized video generation.

4.3 Ablations & Analysis.

4.3.1 RL-based Training.

We fine-tune our supervised model on a dataset of approximately 20,000 samples randomly selected from the Zhu et al. [2022] dataset, each consisting of a tuple of (ground truth, generated sample 0 (0% masked), generated sample 1 (20% masked), generated sample 2 (40% masked)) data points, as described in Sec. 3.3. The model is trained for roughly 5,000 steps with an effective batch size of 8, using a beta value of 2,500 and a learning rate of 1×10^{-5} with AdamW as the optimizer. Additionally, we apply an exponential moving average (EMA) to the reference model, which allows the model to gradually shift toward a better distribution.

As shown in Tables 1 and 2, our RL-enhanced model achieves state-of-the-art performance across most metrics. Compared to the non-RL variant, we observe that reinforcement learning markedly improves speech synchrony, as indicated by higher Sync-C and Sync-D scores, as well as better FID and FVD values. We attribute these gains to the RL training, which effectively enhances both the



Figure 6: Effect of using multi-level speech representations on capturing long-range future dependencies. With multi-level guidance, the model successfully initiates the motion sequence well before the corresponding speech stimulus, whereas the model without multi-level guidance fails to anticipate this future motion.

realism and the audio-visual alignment in speech-to-video generation. We also provide an ablation on different values for the beta in Table 4

4.3.2 Ablations for Architecture Design

We evaluate the impact of capturing multiple views of the audio signal by employing a multi-level speech encoder that encodes different temporal neighborhood sizes at each level. For each speech-injection block, we subsequently use a level-selection layer to choose the most appropriate neighborhood scale. To assess the effectiveness of this design, we compare it against a baseline that uses a speech encoder without multi-level embedding. As shown in Table 3, removing the multilevel embedding and selection mechanism leads to a noticeable drop in performance. We also visualize this effect in Figure 6, where the multilevel embedding helps initiate an anticipatory head motion in response to speech stimuli far into the future, while the version without fails to capture such a relation.

Table 3: **Component-wise ablation.** Synchronization metrics shown across architectural variations. Bold indicates best.

Method Configuration	Sync-C ↑	Sync-D ↓
Ours (Full)	**1.721**	12.505
w/o Multi-headed FiLM (Cross-Attn)	1.653	**12.476**
w/o Multi-Level Embedding	1.405	12.740

We evaluate the effect of using a Multi-headed FiLM layer as opposed to cross-attention for speech-to-video frame conditioning (Table 3). In both settings, speech tokens attend to video frames; however, replacing Multi-headed FiLM with cross-attention results in a drop in performance. This result suggests that in a windowed conditioning setting, Multi-headed FiLM is more effective, as it provides both scale and shift modulation of the visual features. In contrast, cross-attention can be interpreted as primarily inducing a shift, lacking the explicit multiplicative scaling that Multi-headed FiLM offers.

Table 4: Ablation on the value of beta in No-Reward RL training.

Beta Value	Sync-C ↑	Sync-D ↓	FID ↓	FVD ↓	Mesh-FID ↓	Dyn. Deg. ↑
2.5k	**6.671**	7.952	42.207	362.084	**0.1922**	**0.84**
5k	6.665	**7.904**	43.962	307.885	0.2608	0.68
7.5k	6.527	8.155	**40.780**	**270.773**	0.3232	0.60

4.3.3 Effect of Speech Masking

We analyze how speech-masking effects the generation fidelity and demonstrate how this approach can be a proxy for getting samples with worse realism and speech-synchrony.

Table 5: Quantitative results under varying levels of audio speech masking.

Masking %	Sync-C ↑	Sync-D ↓	FID ↓	FVD ↓	Mesh-FID ↓	Dyn. Deg. ↑
0%	6.064	8.096	36.921	277.768	0.1017	0.762
20%	5.368	8.877	32.772	288.066	0.2232	0.724
40%	3.211	12.694	46.881	289.911	0.3444	0.800

Here, we refer masking as adding silent regions to the conditioned audio (i.e., setting the audio signal to zero). These silent regions already exist in the dataset. Therefore during the supervised training phase, the model learns to associate silence with a lack of lip, head, or expression movement, matching real-world expectations and the dataset distribution.

We exploit this behavior by introducing random regions of silence to the audio condition at varying levels. By applying this silent masking across different window sizes and random locations within the audio signal, we effectively decrease lip synchrony, head movement, and overall expressiveness. While this leads to reduced realism, all generated videos remain temporally consistent. The primary effect of masking is simply to dampen the influence of speech globally. Since the initial noise is shared and the first frame remains constant, the only variable that remains is audio condition. Consequently, a muted or silenced audio condition cannot yield a more expressive or realistic output.

This trend is further supported by Table 5, where masking speech results in lower overall realism, not just reduced speech synchrony. For instance, Mesh-FID measures meshes in a normalized image space, effectively quantifying expression and head motion. As shown, performance on it worsens significantly as a larger portion of the speech signal is masked.

5 Conclusion

We introduced ReFree-S2V, a speech-to-video framework that captures expressive speech representations through multi-level guidance. Our approach improves speech–lip synchrony and visual realism via a novel reward-free reinforcement learning scheme, eliminating the need for human annotators and avoiding the biases and under-representation of realism often present in learned reward models. Both quantitative and qualitative experiments show that ReFree-S2V sets a new state of the art across standard metrics, outperforming previous methods. Our user study further demonstrates that ReFree-S2V achieves near–ground-truth speech synchrony and delivers significant improvements in perceived realism.

Acknowledgements

This work was funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK 2853/1 “Neuroexplicit Models of Language, Vision, and Action” - project number 471607914.

References

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

OpenAI. Video generation models as world simulators, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Xu Bin, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 83048–83077, 2025. URL `https://proceedings.iclr.cc/paper_files/paper/2025/file/ce31378e9f41d8907e97dab172b6c559-Paper-Conference.pdf`.

Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, Ke Sun, Linrui Tian, Guangyuan Wang, Qi Wang, Zhongjian Wang, Jiayu Xiao, Sheng Xu, Bang Zhang, Peng Zhang, Xindi Zhang, Zhe Zhang, Jingren Zhou, and Lian Zhuo. Wan-s2v: Audio-driven cinematic video generation, 2025. URL `https://arxiv.org/abs/2508.18621`.

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 video diffusion transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21086–21095, 2025a.

Xiaozhong Ji, Xiaobin Hu, Zhihong Xu, Junwei Zhu, Chuming Lin, Qingdong He, Jiangning Zhang, Donghao Luo, Yi Chen, Qin Lin, et al. Sonic: Shifting focus to global audio perception in portrait animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 193–203, 2025.

Mengchao Wang, Qiang Wang, Fan Jiang, Yaqi Fan, Yunpeng Zhang, Yonggang Qi, Kun Zhao, and Mu Xu. Fantasytalking: Realistic talking portrait generation via coherent motion synthesis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9891–9900, 2025a.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Allen Z Ren, Justin Lidard, Lars L Ankile, Anthony Simeonov, Pulkit Agrawal, Anirudha Majumdar, Benjamin Burchfiel, Hongkai Dai, and Max Simchowitz. Diffusion policy policy optimization. *arXiv preprint arXiv:2409.00588*, 2024.

Weijia Wu, Chen Gao, Joya Chen, Kevin Qinghong Lin, Qingwei Meng, Yiming Zhang, Yuke Qiu, Hong Zhou, and Mike Zheng Shou. Reinforcement learning for large model: A survey. *arXiv preprint arXiv:2508.08189*, 2025.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Runtao Liu, Haoyu Wu, Ziqiang Zheng, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8009–8019, 2025.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Zhaoqing Wang, Xiaobo Xia, Zhuolin Bie, Jinlin Liu, Dongdong Yu, Jia-Wang Bian, and Changhu Wang. Taming camera-controlled video generation with verifiable geometry reward. *arXiv preprint arXiv:2512.02870*, 2025b.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023b.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6): 1–15, 2020.

Siddharth Gururani, Arun Mallya, Ting-Chun Wang, Rafael Valle, and Ming-Yu Liu. Space: Speech-driven portrait animation with controllable expression. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 20914–20923, 2023.

Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20395–20405, 2022.

K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. 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*, MM '20, page 484–492, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413532. URL `https://doi.org/10.1145/3394171.3413532`.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. adtalker: 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.

Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *Advances in Neural Information Processing Systems*, 37:660–684, 2024.

Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J Kim. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo. *arXiv preprint arXiv:2506.07464*, 2025.

Jiahao Cui, Yan Chen, Mingwang Xu, Hanlin Shang, Yuxuan Chen, Yun Zhan, Zilong Dong, Yao Yao, Jingdong Wang, and Siyu Zhu. Hallo4: High-fidelity dynamic portrait animation via direct preference optimization and temporal motion modulation. *arXiv preprint arXiv:2505.23525*, 2025b.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016a.

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pages 650–667. Springer, 2022.

Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, Praveen Chowdary, Joe Chuang, Antony D'Avirro, Jon Daly, Ning Dong, Mark Duppenthaler, Cynthia Gao, Jeff Girard, Martin Gleize, Sahir Gomez, Hongyu Gong, Srivathsan Govindarajan, Brandon Han, Sen He, Denise Hernandez, Yordan Hristov, Rongjie Huang, Hirofumi Inaguma, Somya Jain, Raj Janardhan, Qingyao Jia, Christopher Klaiber, Dejan Kovachev, Moneish Kumar, Hang Li, Yilei Li, Pavel Litvin, Wei Liu, Guangyao Ma, Jing Ma, Martin Ma, Xutai Ma, Lucas Mantovani, Sagar Miglani, Sreyas Mohan, Louis-Philippe Morency, Evonne Ng, Kam-Woh Ng, Tu Anh Nguyen, Amia Oberai, Benjamin Peloquin, Juan Pino, Jovan Popovic, Omid Poursaeed, Fabian Prada, Alice Rakotoarison, Alexander Richard, Christophe Ropers, Safiyyah Saleem, Vasu Sharma, Alex Shcherbyna, Jia Shen, Jie Shen, Anastasis Stathopoulos, Anna Sun, Paden Tomasello, Tuan Tran, Arina Turkatenko, Bo Wan, Chao Wang, Jeff Wang, Mary Williamson, Carleigh Wood, Tao Xiang, Yilin Yang, Zhiyuan Yao, Chen Zhang, Jiemin Zhang, Xinyue Zhang, Jason Zheng, Pavlo Zhyzheria, Jan Zikes, and Michael Zollhoefer. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. 2025. URL `https://ai.meta.com/research/publications/seamless-interaction-dyadic-audiovisual-motion-modeling-and-large-scale-dataset/`.

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

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016b.

Google. Mediapipe. `https://github.com/google/mediapipe`, 2020. Accessed: 2026-01-22.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

A Supplementary Material

A.1 Computational Efficiency

We compare the computational efficiency of our method against recent state-of-the-art approaches for audio-driven portrait animation in Table 6.

Table 6: Computational efficiency comparison.

Method	VRAM (GB) ↓	FPS ↑
Sonic	13.8	0.8
FantasyTalking	36.1	0.2
Hallo3	48.3	0.1
Ours	30.6	0.7

A.2 Additional Qualitative results



Figure 7: Qualitative results on in-the-wild AI-generated images. Note the accurate lip shapes for individual phonemes and expressive facial expressions for non-speech vocalizations like “giggles”.

A.3 Additional Quantitative results

We provide additional evaluation on the "in-the-wild" subset in Table 7. Note that FantasyTalking does not provide code for its evaluation metrics, so the metrics were re-implemented in our work. We observe the best performance in Subject Dynamics and Body Dynamics, indicating improved expressiveness. We also observe a slight drop in ID-Consistency, which we attribute to the use of DINO features in the metric computation. Since DINO features are not identity-aware, more dynamic and expressive motions can lead to larger feature variations across frames, resulting in lower consistency scores. while we see almost no difference in the aesthetic score across methods.

Table 7: Evaluation on the "in-the-wild" subset comparing FantasyTalking metrics.

Method	ID-Consistency ↑	Subj. Dynamics ↑	BG Dynamics ↑	Aesthetic ↑
Sonic	**0.952**	1.812	0.950	5.098
FantasyTalking	0.941	0.896	0.745	**5.169**
Hallo3	0.938	1.755	0.760	5.058
Ours	0.926	**3.426**	**2.698**	5.133

A.4 Social Risks and Potential Mitigation

Video generation frameworks carry social risks, such as misuse for deceptive content and privacy violations. These concerns underscore the importance of ethical guidelines, informed consent, and transparent disclosure of AI-generated content. Given the inclusion of personal identities, careful handling and robust watermarking are essential to distinguish AI-generated from real content.

A.5 Limitations

Although our method achieves improved realism, it is inherently limited by the underlying video generation model, resulting in high runtime and VRAM usage.

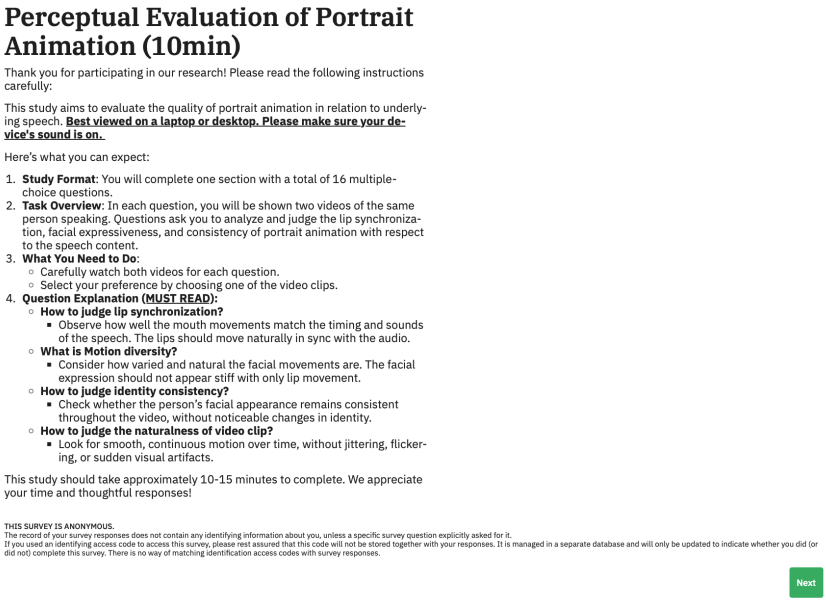
A promising direction for future work is to improve efficiency via faster inference strategies. Techniques such as few-step distillation and efficient sampling could significantly reduce the number of required steps, enabling faster generation while preserving visual quality.

A.6 Additional Implementation Details

Our full training pipeline was executed on a cluster of eight NVIDIA H100 GPUs. For the Supervised Fine-Tuning (SFT) phase, we trained the model for approximately 50,000 steps using the AdamW optimizer. The learning rate was set to 1×10^{-4} for the initial 30,000 steps and subsequently decayed to 1×10^{-5} for the remaining 20,000 steps.

To improve the fidelity of facial dynamics, we employed regional loss weighting to provide a stronger L_2 loss guidance signal. Specifically, we applied a weight of 100 to the lip region and 30 to the rest of the face, while the remainder of the video frames maintained a baseline weight of 1.

A.7 User study



Perceptual Evaluation of Portrait Animation (10min)

Thank you for participating in our research! Please read the following instructions carefully:

This study aims to evaluate the quality of portrait animation in relation to underlying speech. **Best viewed on a laptop or desktop. Please make sure your device's sound is on.**

Here's what you can expect:

1. **Study Format**: You will complete one section with a total of 16 multiple-choice questions.
2. **Task Overview**: In each question, you will be shown two videos of the same person speaking. Questions ask you to analyze and judge the lip synchronization, facial expressiveness, and consistency of portrait animation with respect to the speech content.
3. **What You Need to Do**:
   - Carefully watch both videos for each question.
   - Select your preference by choosing one of the video clips.
4. **Question Explanation (MUST READ):**
   - **How to judge lip synchronization?**
     - Observe how well the mouth movements match the timing and sounds of the speech. The lips should move naturally in sync with the audio.
   - **What is Motion diversity?**
     - Consider how varied and natural the facial movements are. The facial expression should not appear stiff with only lip movement.
   - **How to judge identity consistency?**
     - Check whether the person's facial appearance remains consistent throughout the video, without noticeable changes in identity.
   - **How to judge the naturalness of video clip?**
     - Look for smooth, continuous motion over time, without jittering, flickering, or sudden visual artifacts.

This study should take approximately 10-15 minutes to complete. We appreciate your time and thoughtful responses!

THIS SURVEY IS ANONYMOUS.
The record of your survey responses does not contain any identifying information about you, unless a specific survey question explicitly asked for it.
If you used an identifying access code to access this survey, please rest assured that this code will not be stored together with your responses. It is managed in a separate database and will only be updated to indicate whether you did (or did not) complete this survey. There is no way of matching identification access codes with survey responses.

Next

Figure 8: Screenshot of the user study

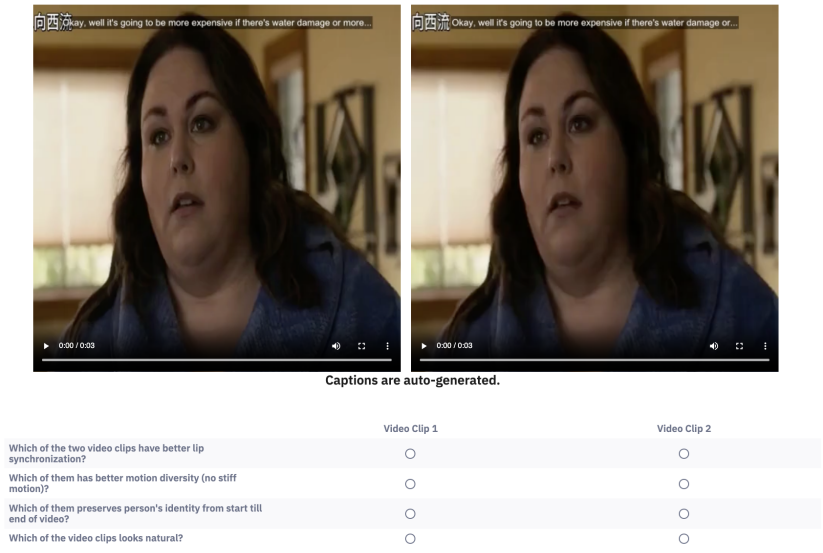


Figure 9: Screenshot of the user study

Input Sampling. Portrait identities and driving audio were drawn uniformly at random from the Hallo3 in-the-wild test set described in Section 4. For each sampled input we generated a video with our method and with each comparison method, and formed side-by-side pairs. This produced a pool of 48 comparison pairs spanning four conditions: ours vs. FantasyTalking, ours vs. Hallo3, ours vs. Sonic, and ours vs. ground truth (GT), balanced at 12 pairs per condition.

User Study Design. Each participant was shown a randomly drawn subset of 16 pairs from this pool, balanced at 4 pairs per condition, so that different participants saw different stimuli and the study collectively covers all 48 pairs rather than repeatedly testing one fixed set. For every pair, participants answered four forced-choice sub-questions, following a design similar to Sonic:

- "Which of the two video clips have better lip synchronization?"
- "Which of them has better motion diversity (no stiff motion)?"
- "Which of them preserves person's identity from start till end of video?"
- "Which of the video clips looks natural?"

Each participant therefore provided 16 x 4 = 64 responses. Left/right placement within each pair was randomized independently per pair, and pair order was randomized across conditions so that comparisons against the same source were not presented consecutively. Participants were blind to which video came from which system and were not told that one condition was ground truth. Participants could also play each clip more than once before responding.

Participants and compensation. 22 participants took part in our user study, mainly including university students and academic staff. Participants were volunteers and therefore were not compensated. No personally identifying data like age, gender, ethnicity and education levels was collected due to data-protection laws.

Results. We perform significance tests, shown in Figure 5 of the main draft. Our method is preferred over all three baselines on all four attributes, significantly so in 10 of 12 bars; the two exceptions being ID consistency (53.4%) and video realism (58.0%) against Sonic, where the preference is positive but not significant. Against ground truth, our method is below chance on all four attributes, significantly so for video realism (23.9%), ID consistency (35.2%), and motion diversity (37.5%), while lip synchronization (46.6%) is statistically indistinguishable from real video. The realism gap is expected as video generation models have yet to achieve the same fidelity as ground-truth data.

A.8 Additional License Information

In this work, we use several open-source datasets and pretrained models. While all resources are properly cited throughout the paper, we provide here a non-exhaustive summary of the corresponding licenses and usage terms.

The Hallo3 dataset Cui et al. [2025a] is released under the CC BY-NC-ND 4.0 license. HDTF Zhang et al. [2021] is distributed under the GNU GPL v3.0 license. The Wan model Wan et al. [2025] is released under the Apache 2.0 license.

The CelebV-HQ dataset Zhu et al. [2022] is available for non-commercial research purposes only.

The Seamless Interaction dataset Agrawal et al. [2025] is released under the CC BY-NC 4.0 license.

Finally, some initial frame images used in the paper and supplementary videos were obtained from free-to-use resources requiring author attribution. These images were sourced from `freepik.com` and `vecteezy.com`, including work by Adam Zubek-Nizol.

NeurIPS Paper Checklist

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 claim that our method improves realism through a novel reward-free RL approach. Section 4 validates the effectiveness of the method: Tables 1 and 2 present the overall performance results, while Table 3 provides an ablation study analyzing the contribution of each component.

   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: We discuss the limitations of our approach, including computational requirements and efficiency, in Section A.5 of the supplementary material.

   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: We do not have theoretical results, only empirical.

   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: We provide architectural details, and the implementation details include the hyperparameters required to train the model in Sections 3 and 4, as well as Appendix A.6. We also provide ablations for several key hyperparameters, and we plan to release the code.

   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: We plan to release the source code upon acceptance.

   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: We describe the implementation and training details in Section 3.1 and the beginning of Section 4.3.1 as well as Appendix A.6.

   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: [N/A]

   Justification: The compute constraints, unfortunately, prevent us from computing error bars.

   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: Comprehensive details regarding the hardware specifications (NVIDIA H100 cluster), optimization hyperparameters, training duration, and regional loss configurations are provided in Sections 3, 4.3.1, and Appendix A.6.

   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: The research conducted in this paper conforms with 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: We discuss possible positive and negative consequences of our work in the Introduction and in Appendix A.4.

    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: Our work builds upon publicly available models and datasets, including Wan Wan et al. [2025], which include documented safety protocols, licensing terms, and usage restrictions. We follow the original terms of use of these resources and only use datasets and assets that are publicly available for research purposes. In addition, we do not release new large-scale datasets containing sensitive personal information, and all visual examples included in the paper and supplementary material comply with the licensing and attribution requirements of the original assets.

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: We properly cite and credit all external assets used in this work, including datasets, pretrained models, codebases, and visual assets. We explicitly provide license information and terms of use for the main resources used in the project, including CC BY-NC-ND 4.0 (Hallo3), GNU GPL v3.0 (HDTF), Apache 2.0 (Wan), CC BY-NC 4.0 (Seamless Interaction), and the non-commercial research license of CelebV-HQ. We also provide attribution for supplementary visual assets obtained from Freepik and Vecteezy. Additional licensing details are included in the supplementary material.

    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: No new assets introduced.

    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: [Yes]

    Justification: The paper and supplementary material include the instructions provided to participants, along with details of the user study protocol and evaluation procedure. Participants were volunteers and did not receive monetary compensation. Screenshots and example interfaces are included where applicable in the Appendix.

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: [No]

    Justification: The crowdsourcing experiment for evaluation of the perceptual quality of the generated results does not mandate an IRB approval in the home institution.

    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: The method does not contain any LLM-specific component.

    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.