SoulX-FlashHead: Oracle-guided Generation of Infinite Real-time Streaming Talking Heads

Tan Yu[*], Qian Qiao[*,†], Le Shen[*], Ke Zhou, Jincheng Hu, Dian Sheng,
Bo Hu, Haoming Qin, Jun Gao, Changhai Zhou, Shunshun Yin, Siyuan Liu[†]
AIGC Team, Soul AI Lab, China

Project Page: https://soul-ailab.github.io/soulx-flashhead/

ABSTRACT

Achieving a balance between high-fidelity visual quality and low-latency streaming remains a formidable challenge in audio-driven portrait generation. Existing large-scale models often suffer from prohibitive computational costs, while lightweight alternatives typically compromise on holistic facial representations and temporal stability. In this paper, we propose SoulX-FlashHead, a unified 1.3B-parameter framework designed for real-time, infinite-length, and high-fidelity streaming video generation. To address the instability of audio features in streaming scenarios, we introduce Streaming-Aware Spatiotemporal Pre-training equipped with a Temporal Audio Context Cache mechanism, which ensures robust feature extraction from short audio fragments. Furthermore, to mitigate the error accumulation and identity drift inherent in long-sequence autoregressive generation, we propose Oracle-Guided Bidirectional Distillation, leveraging ground-truth motion priors to provide precise physical guidance. We also present VividHead, a large-scale, high-quality dataset containing 782 hours of strictly aligned footage to support robust training. Extensive experiments demonstrate that SoulX-FlashHead achieves state-of-the-art performance on HDTF and VFHQ benchmarks. Notably, our Lite variant achieves an inference speed of 96 FPS on a single NVIDIA RTX 4090, facilitating ultra-fast interaction without sacrificing visual coherence.

1 Introduction

In the realm of real-time audio-driven portrait generation, large-scale Diffusion Transformer models [Gao et al., 2025, Zhong et al., 2025, Guo et al., 2024, Yang et al., 2025] have demonstrated exceptional performance in high-fidelity streaming generation. However, their deployment poses significant challenges. Approaches such as LiveAvatar [Huang et al., 2025] and SoulX-FlashTalk [Shen et al., 2025] are hindered by heavy computational overhead and complex pipeline parallelism. This renders low-latency streaming interaction on consumer-grade GPUs largely infeasible. Conversely, traditional quantization and pruning techniques often degrade video fidelity and struggle to capture complex facial micro-expressions or precise lip synchronization.

As shown in Tab. 1, we compare existing methodologies along four dimensions including streaming capability, real-time inference, infinite-length generation, and holistic representation. We define Holistic Representation as the modeling of pixel-level latent spaces utilizing image or video VAEs. This approach ensures the preservation of complete facial textures and visual coherence. In contrast, approaches like Ditto [Li et al., 2025] and SadTalker [Zhang et al., 2023] rely on motion VAE–based representations. These are inherently more abstract and lack a unified description of the facial structure. In contrast, approaches like Ditto [Li et al., 2025] and SadTalker [Zhang et al., 2023] rely on motion VAE–based representations which are inherently abstract and lack a unified facial representation. A clear trade-off exists between efficiency and representational capacity. Lightweight models enable real-time streaming but lack holistic details while high-fidelity models like Hallo3 [Cui et al., 2025] and AniPortrait [Wei et al., 2024] typically fail to support real-time or streaming inference. A unified framework satisfying all requirements within a moderate parameter scale remains elusive.

To bridge this gap, we introduce SoulX-FlashHead. This is a 1.3B-parameter model designed for real-time streaming video generation. Unlike previous works that compromise between speed and quality, SoulX-FlashHead achieves

[*]Equal contribution. {qiaoqian,yutan}@soulapp.cn, Core Contributors: Tan Yu, Qian Qiao, Le Shen, Ke Zhou
[†]Corresponding authors. Project Leader: liusiyuan@soulapp.cn

a balance by employing a two-stage training scheme comprising Streaming-Aware Spatiotemporal Pre-training and Oracle-Guided Bidirectional Distillation. This design specifically addresses two fundamental challenges in the field.

Data Noise and Audio Feature Instability. Lightweight models often struggle to learn precise conditional mappings from noisy datasets. Streaming interaction necessitates processing extremely short audio fragments which are typically around 1.32 seconds. Traditional self-supervised audio models like Wav2Vec [Baevski et al., 2020] are prone to feature space distribution shifts or collapse when handling such short sequences. This leads to signal distortion. We address this via Streaming-Aware Spatiotemporal Pre-training. At the data level, we constructed a rigorous cleaning pipeline to refine VividHead which is a high-quality dataset containing over 10,000 hours of highly aligned data. Furthermore, we introduced a Temporal Audio Context Cache mechanism during pre-training. This explicitly caches historical audio to compensate for context deficiency in short-slice inputs and ensures robust audio representations and precise audio-visual alignment.

Error Accumulation in Long-Sequence Generation. Real-time streaming requires autoregressive prediction where minor deviations in 1.3B models amplify rapidly and cause facial distortion or identity drift. While distribution distillation (DMD) [Yin et al., 2024] is commonly used to reduce error accumulation, existing methods ignore the misalignment between pre-training and distillation phases. Specifically, the teacher often uses ground truth motion frames while the student relies on predictions. This results in inaccurate guidance. We propose Oracle-Guided Bidirectional Distillation to overcome this. By utilizing Ground Truth motion frames as "Oracle" conditional anchors, we provide clear physical priors. This strong constraint mechanism suppresses error diffusion in long sequences and fully exploits the teacher model to improve fidelity at low inference steps.

SoulX-FlashHead is provided in two flexible deployment versions. SoulX-FlashHead-Lite targets ultra-fast interaction scenarios and achieves real-time inference on a single RTX 4090 GPU. SoulX-FlashHead-Pro pursues superior detail and enables real-time generation on dual RTX 5090 GPUs.

Table 1: Comparison of different audio-driven portrait generation methods. Our model uniquely achieves a balance of streaming capability, real-time performance, infinite generation length, and holistic representation with a lightweight 1.3B parameter size.

Method	Stream	Real-Time	Inf-len	Holistic Representation	**Size**
SadTalker [Zhang et al., 2023]	✓	✓	✓	✗	0.2B
Aniportrait [Wei et al., 2024]	✗	✗	✓	✓	1.7B
EchoMimicV3 [Chen et al., 2025a]	✓	✗	✗	✓	1.3B
Ditto [Li et al., 2024]	✓	✓	✓	✗	0.2B
Hallo3 [Cui et al., 2025]	✓	✗	✗	✓	5B
Sonic [Ji et al., 2025]	✗	✗	✓	✓	1.5B
SoulX-FlashHead	✓	✓	✓	✓	**1.3B**

2 Data

To enable high-fidelity and vivid portrait video generation, we constructed VividHead [3], a large-scale and high-quality dataset. Recognizing that data quality dictates the upper bound of downstream model performance, we developed the comprehensive data processing pipeline shown in Fig. 1 to transform unstructured raw web videos into clean and semantically rich training data. This process comprises two core phases where Data Preprocessing handles acquisition and standardization while Data Filtering and Annotation performs strict screening and fine-grained labeling.

VividHead consists of 330,000 high-quality short clips ranging from 3 to 60 seconds with a total duration of 782 hours. Each sample features 512×512 resolution image sequences, strictly time-aligned speech audio, and rich metadata encompassing language, ethnicity, and age. We restricted the collection to samples containing a single visible speaker with an active head region.

[3] https://huggingface.co/datasets/Soul-AILab/VividHead

Tab. 2 compares VividHead with existing mainstream datasets. Distinct from collections limited to laboratory settings such as MEAD or lower resolutions like HDTF, VividHead balances in-the-wild diversity with high visual quality standards across 15 languages and diverse demographics.

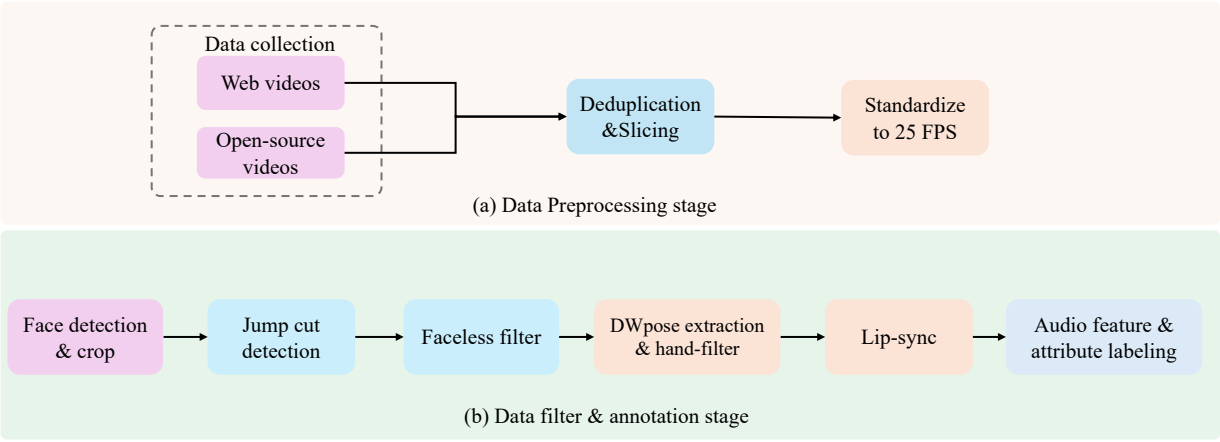

Figure 1: Overview of our comprehensive data filtering pipeline. We obtain 782 hours of high-quality audio–video data from 10k hours

Table 2: Comparison of TalkVivid with existing talking head datasets. TalkVivid distinguishes itself through a combination of large scale, high resolution, and diverse attribute coverage in wild scenarios.

Dataset	Speakers	Face Crop	Clips	Hours	Resolution	Language	Age	Ethnicty	Source
MEAD [Wang et al., 2020]	60	✓	281.4K	39	384p	English	20-35	-	Lab
HDTF [Zhang et al., 2021]	362	✓	10K	15.8	512p	-	-	-	wild
AVspeech [Ephrat et al., 2018]	150k	✗	2.5M	4700	720p, 1080p	-	-	-	wild
Hallo3 [Cui et al., 2025]	-	✓	101.5K	70	720p	-	-	-	wild
OpenHumanVid [Zheng et al., 2024]	-	✗	13.4M	16.7k	720p	-	-	-	wild
TalkVid [Chen et al., 2025b]	7729	✗	281.4K	1244	1080p, 2160p	15 langs	0-60+	3	wild
SpeakerVid [Zhang et al., 2025]	83k	✗	5.2M	8.7K	1080p	-	-	-	wild
VividHead	**60k**	✓	**330K**	**782**	**512p**	**15 langs**	**0-60+**	**3**	**wild**

2.1 Data Processing Pipeline

2.1.1 Data Preprocessing stage

We constructed the initial data pool by aggregating public datasets [Ephrat et al., 2018, Zhang et al., 2021, Cui et al., 2025, Chen et al., 2025b, Zhang et al., 2025] and extensive web resources. To address the redundancy inherent in large-scale multi-source collection, we applied a verification mechanism based on source video IDs and MD5 hash values to eliminate duplicate content and guarantee sample uniqueness.

For video segmentation, we employed distinct strategies where public datasets with existing timestamps were cropped precisely while unlabelled web data underwent adaptive scene detection via PySceneDetect [Castellano, 2024]. This approach divided long videos into coherent clips ranging from 5 to 50 seconds to preserve semantic context. All clips were subsequently normalized to a unified frame rate of 25 fps using FFMPEG to ensure temporal consistency for downstream modeling.

2.1.2 Data filter and annotation stage

In the data filtering phase, we first perform face detection and mask extraction on the initial frame of each clip, and valid clips are finally cropped to a resolution of 512×512 pixels centered on the detected face. Subsequently, optical flow analysis [Dosovitskiy et al., 2015] is employed to identify and remove sequences containing abrupt scene transitions.

To ensure consistent facial visibility, we inspect raw footage at a sampling rate of one frame per second and exclude videos where the proportion of frames lacking a detectable face exceeds a predefined threshold. We further utilize DWpose [Yang et al., 2023] to extract body keypoints and strictly filter out clips exhibiting hand-over-face occlusion to prevent generation artifacts. Finally, we employ the SyncNet [Chung and Zisserman, 2016] model to calculate LSE-C and LSE-D confidence scores to discard samples with poor audio-visual alignment.

High-quality clips passing these rigorous filters proceed to the automated data annotation and feature extraction pipeline. High-precision detectors extract face masks at the visual level to decouple the foreground from complex backgrounds. For audio, we separate streams and employ pre-trained Wav2Vec models to extract streaming audio embeddings as robust driving features. We also annotate clips with multi-dimensional attributes including gender, age, ethnicity, and language to enhance the capabilities of the model in controllable generation tasks.

3 SoulX-FlashHead

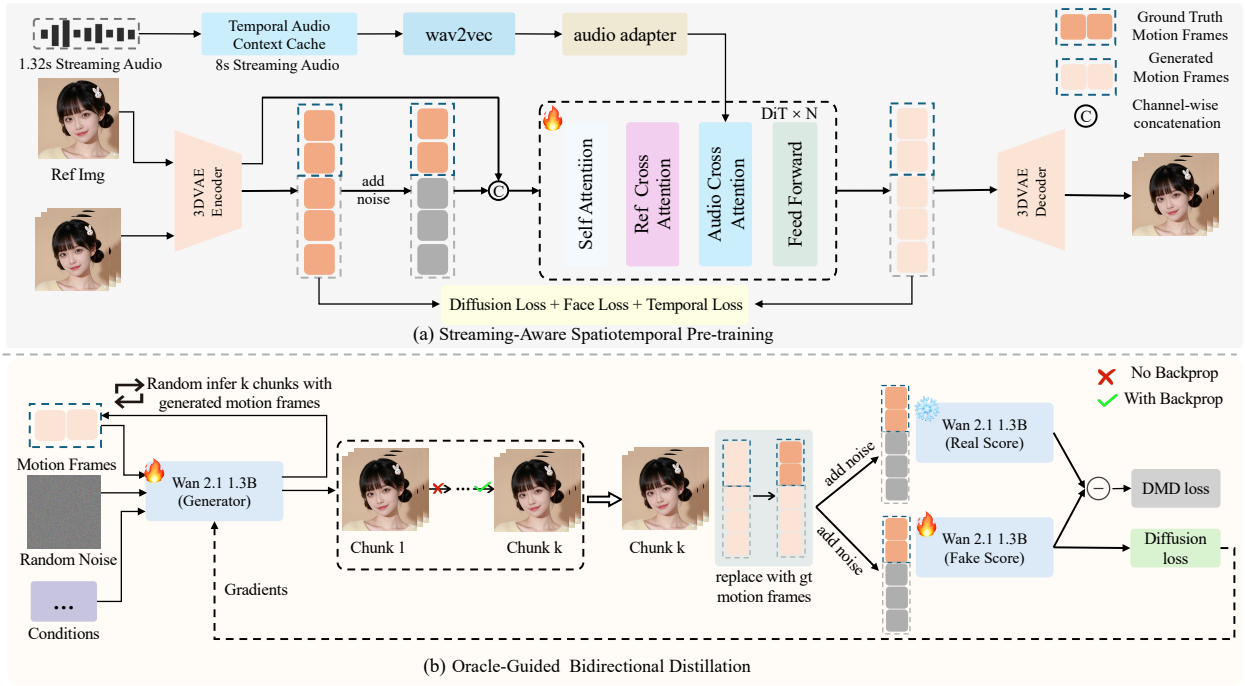

Figure 2: Framework Overview of SoulX-FlashHead. (a) Stage 1: Streaming-Aware Spatiotemporal Pre-training. We employ a Temporal Audio Context Cache to stabilize feature extraction from short streaming audio and utilize channel-wise concatenation for robust reference image injection. (b) Stage 2: Oracle-Guided Bidirectional Distillation. To mitigate error accumulation, the Student generates autoregressively conditioned on its own historical predictions, while the Teacher utilizes Ground Truth motion frames as an "Oracle" guide. The model is optimized via a Stochastic Truncation Strategy using DMD and latent regression losses.

We present SoulX-FlashHead, a framework designed to generate audio-synchronized videos that faithfully preserve reference image content given audio and image inputs. As illustrated in Figure 2, the architecture builds upon a 1.3B-parameter DiT backbone and employs a two-stage training strategy comprising Streaming-Aware Spatiotemporal Pre-training and Oracle-Guided Bidirectional Distillation. This training paradigm is supported by a comprehensive real-time inference acceleration pipeline.

3.1 Model Architecture

3D-VAE. To strictly address the trade-off between visual fidelity and inference latency, we employ distinct VAE architectures for our Pro and Lite variants. The Pro model utilizes the WAN 2.1 VAE [Wan et al., 2025] which encodes video frames into compact latent representations with a spatio-temporal downsampling factor of $4 \times 8 \times 8$. This configuration prioritizes high-fidelity reconstruction and fine-grained detail preservation suitable for quality-critical scenarios. Conversely, the Lite model incorporates the LTX-VAE [HaCohen et al., 2024] to optimize for real-time

performance. LTX-VAE achieves aggressive compression by downsampling inputs by a factor of $(32, 32, 8)$, yielding a pixel-to-token ratio of $8192 : 1$—approximately $32\times$ higher than the WAN scheme. While LTX-VAE integrates a diffusion decoding step to mitigate detail loss inherent to such high compression, it fundamentally serves as the engine for speed-balanced inference in latency-sensitive applications.

Diffusion Transformer (DiT). DiT [Peebles and Xie, 2023] replaces the U-Net [Ronneberger et al., 2015] in latent diffusion models (LDMs) [Rombach et al., 2022] with a Transformer, enhancing capacity and scalability across space and time. In video generation, it is often combined with a causal 3D VAE [Kingma and Welling, 2013, Zheng et al., 2024] for spatio-temporal compression, while conditional inputs are incorporated via adaptive normalization or cross-attention. The objective is to learn a vector field by minimizing the mean squared error (MSE) between predicted and ground-truth velocities:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t,\, p_t(\mathbf{x})} \|\mathbf{v}_t - \mathbf{u}_t\|^2, \tag{1}$$

where $\mathbf{x}_t = t \cdot \mathbf{x}_1 + (1 - t) \cdot \mathbf{x}_0$. We use Wan2.1 T2V [Wan et al., 2025] (1.3B) as our baseline, balancing performance and efficiency.

Audio Condition injection. To facilitate Streaming-Aware Spatiotemporal Pre-training and address streaming input instability, we designed a Temporal Audio Context Cache mechanism using a fixed-length window. We define the audio context window length as $T_{max} = 8$s. For any given raw audio stream $\mathcal{A}_{raw}$, we construct a standardized input queue $\mathcal{Q}_{audio}$ as follows:

$$\mathcal{Q}_{audio} = \begin{cases} \text{Concat}(\mathbf{0}_{\text{pad}}, \mathcal{A}_{raw}), & \text{if } |\mathcal{A}_{raw}| < T_{max} \\ \mathcal{A}_{raw}[-T_{max}:], & \text{if } |\mathcal{A}_{raw}| \geq T_{max} \end{cases} \tag{2}$$

Here, $\mathbf{0}_{\text{pad}}$ denotes silence audio padding, and $\mathcal{A}_{raw}[-T_{max}:]$ represents the trailing 8s segment of the audio stream. We then encode $\mathcal{Q}_{audio}$ using Wav2Vec. To integrate multi-level semantic cues, we aggregate multi-layer Wav2Vec features, yielding the full audio feature sequence $E_{audio} \in \mathbb{R}^{S \times Layers \times D}$, where S corresponds to the feature frame count for the 8s duration. During training and inference, we ensure precise synchronization with the current video frame by extracting the last N time-aligned audio feature frames from E_{audio} as the driving condition:

$$Z_{cond} = E_{audio}[-N:] \tag{3}$$

Finally, Z_{cond} is injected into selected DiT blocks via pixel-wise cross-attention to drive facial motion generation in an end-to-end manner.

Image Conditioning and Anti-Drift. Image conditioning is critical for maintaining visual consistency in long-video generation. To reinforce identity preservation, we forgo complex attention injection mechanisms in favor of channel-wise concatenation of the reference image latent with the input noise. This direct injection strategy provides a robust spatial structural prior, ensuring that high-frequency details from the reference image maintain precise pixel-level alignment throughout generation. Consequently, this approach effectively mitigates identity drift in long sequences.

Context History and First-Frame Adaptation. To ensure smooth transitions between video chunks, we utilize motion frames as context. However, streaming inference presents a unique "cold start" challenge where the first chunk lacks historical video context. We address this via a Dynamic Motion Frame Sampling strategy during pre-training. Specifically, we sample motion frames from the beginning of the Ground Truth video: with a probability of 0.9, we use the first n frames, and with a probability of 0.1, we use only the single first frame. This distribution effectively simulates the initial inference phase where the motion context consists solely of the reference image, enabling the model to adapt seamlessly to the first chunk's generation.

3.2 Model Training

To achieve high-fidelity real-time streaming generation with a 1.3B-parameter architecture, we employ a two-stage training strategy comprising Streaming-Aware Spatiotemporal Pre-training and Oracle-Guided Bidirectional Distillation. This approach addresses the instability of audio features in streaming environments and the accumulation of errors during long-sequence autoregressive generation.

Stage 1: Streaming-Aware Spatiotemporal Pre-training

Streaming talking head generation faces an inherent conflict between the need for high-fidelity lip synchronization and the instability of real-time audio feature extraction. In live inference scenarios, the model must process extremely short audio fragments of approximately 1.32 seconds. This fragmentation often induces distribution shifts or feature collapse within traditional self-supervised frameworks and results in signal distortion. To mitigate these issues at the data level, we established a rigorous processing pipeline that refines over 10,000 hours of raw footage into TalkVivid which is a high-quality dataset containing 1,000 hours of strictly aligned data. This foundation ensures the model learns from a robust and clean data distribution.

To guarantee stable feature extraction under streaming constraints, we introduce the Temporal Audio Context Cache mechanism as formulated in Equation 2. We maintain a fixed-length audio window of 8 seconds where raw audio inputs are padded with silence or truncated to ensure consistent feature extraction and context integrity. Furthermore, we adopt a probabilistic motion conditioning strategy to adapt the model to both continuous streaming and initial generation phases. During training, we utilize n Ground Truth frames as motion context with a probability of 0.9 to capture temporal dependencies. Conversely, we use a single frame with a probability of 0.1 to simulate the cold start scenario where only the reference image is available.

Stage 2: Oracle-Guided Bidirectional Distillation

Real-time streaming inference necessitates autoregressive generation where the model predicts future frames based on its own history. This process often suffers from error accumulation that leads to severe identity drift. To enable real-time performance, we initially adopt the Distribution Matching Distillation (DMD) framework to compress sampling steps and eliminate the need for classifier-free guidance. DMD optimizes the Kullback–Leibler divergence to minimize the distributional discrepancy between the original teacher and the distilled student at each timestep t. The gradient update is formulated as:

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = -\mathbb{E}_{t,\mathbf{z}} \left[\left(s_{\text{real}}(\psi(G_\theta(z), t), t) - s_{\text{fake}}(\psi(G_\theta(z), t), t) \right) \frac{\partial G_\theta(\mathbf{z})}{\partial \theta} \right], \tag{4}$$

Here, s_{real} represents the frozen teacher score network modeling the true distribution, while s_{fake} tracks the evolving student distribution generated by G_θ. All components are initialized from the Stage-1 Pretrained model.

Standard DMD overlooks the temporal dependency in streaming where the student relies on imperfect historical context. To address this, we introduce Oracle-Guided Bidirectional Distillation. Inspired by the error correction mechanism in Self-Forcing++, we explicitly simulate the autoregressive inference process during training. The student generator synthesizes K consecutive video chunks where each chunk is conditioned on its previously generated motion frames $\mathbf{m}_{pred}$ rather than ground truth.

We implement a Stochastic Truncation Strategy to balance computational efficiency with training stability. Instead of unrolling the full computation graph for K chunks, we randomly sample a truncation length k from a uniform distribution and generate only the first k chunks. During backpropagation, gradients are retained only for a randomly sampled denoising step t' of the k-th chunk, while strictly detaching all preceding steps from the computational graph.

Crucially, we leverage an Oracle supervision signal where the teacher model remains conditioned on the ground truth motion history $\mathbf{m}_{gt}$. This contrasts with the student which is conditioned on its accumulated history $\mathbf{m}_{pred}$. We further enforce trajectory alignment by imposing a latent space regression loss. The final objective aggregates the distribution matching loss and the regression penalty:

$$\mathcal{L}_{\text{total}} = \mathbb{E}_{k,t'} \left[\underbrace{D_{\text{KL}} \left(\underbrace{p_\phi(\mathbf{x}_0|\mathbf{m}_{\text{gt}})}_{s_{\text{real}}} \,\middle\|\, \underbrace{p_\theta(\mathbf{x}_0|\mathbf{m}_{\text{pred}})}_{s_{\text{fake}}} \right)}_{\mathcal{L}_{\text{DMD}}} + \lambda \underbrace{\|\mathbf{z}^{(k)}_{\text{student}} - \mathbf{z}^{(k)}_{\text{gt}}\|_2^2}_{\mathcal{L}_{\text{reg}}} \right] \tag{5}$$

The $\mathcal{L}$DMD term utilizes the Oracle distribution p_ϕ to guide the drifting student distribution p_θ back to the optimal manifold. Simultaneously, $\mathcal{L}_{\text{reg}}$ minimizes the Euclidean distance between the student's latent output and the ground truth to ensure physical trajectory alignment.

3.3 Real-time Inference Acceleration

To enable low-latency inference on consumer-grade hardware (e.g., NVIDIA RTX 4090 and RTX 5090), we implement a full-stack acceleration pipeline tailored for the 1.3B-parameter model.

Hybrid Sequence Parallelism. The primary computational cost lies in the DiT attention layers. We employ Hybrid Sequence Parallelism via xDiT to distribute the attention workload. By combining Ulysses and Ring Attention mechanisms, we achieve significant speedups in single-step inference for multi-GPU setups compared to standard implementations.

Kernel Optimization. We adopt FlashAttention-2 to optimize attention operations at the kernel level. This implementation maximizes memory bandwidth utilization on NVIDIA Ada Lovelace and Blackwell architectures. By minimizing memory access overhead and optimizing IO complexity, FlashAttention-2 [Dao, 2024] further reduces attention latency.

Table 3: **Quantitative comparison on the HDTF dataset.** Models marked with * support streaming, while those marked with △ do not. **For non-streaming methods, FPS is not a meaningful metric.**

Method	FID↓	FVD↓	Sync-C↑	Sync-D↓	FPS↑
SadTalker* [Zhang et al., 2023]	21.58	207.67	4.60	9.21	2.17
Aniportrait△ [Wei et al., 2024]	19.83	242.29	1.89	11.91	-
EchoMimic* [Chen et al., 2025a]	9.00	155.71	3.56	10.22	0.81
Ditto* [Li et al., 2024]	12.35	199.13	3.57	10.49	<u>45.04</u>
Hallo3* [Cui et al., 2025]	15.95	160.94	3.18	10.72	0.16
Sonic△ [Ji et al., 2025]	13.53	113.31	5.17	8.69	-
SoulX-FlashHead (Lite)*	11.37	126.52	4.21	9.49	**96**
SoulX-FlashHead (Pro)*	<u>9.97</u>	<u>111.38</u>	<u>5.73</u>	<u>8.77</u>	10.81
SoulX-FlashHead (Lite)△	10.78	115.94	5.12	8.80	-
SoulX-FlashHead (Pro)△	**8.31**	**103.14**	**6.04**	**8.46**	-

Runtime Optimization. Finally, we utilize `torch.compile` to unify the inference pipeline. This eliminates Python runtime overhead and enables graph-level fusion, ensuring optimal memory usage and execution efficiency on the target hardware.

4 Experiments

Implementation Details. We build our model upon the Wan2.1 T2V (1.3B) architecture, optimizing it to satisfy real-time constraints. Stage 1 pretrains at 2×10^{-4} higher learning rate, incorporating a warm-up strategy and optimized using the AdamW optimizer. The model is trained using 32 NVIDIA H20 GPUs. In the first stage, the global batch size is set to 256 and the model is trained for $100,000$ steps. For the subsequent distillation stage, We adhere to the Self-Forcing training paradigm, setting learning rates to 2×10^{-6} for the Generator and 4×10^{-7} for the Fake Score Network with a 1:5 update ratio. To simulate error accumulation in long-horizon generation, the Generator synthesizes up to $K = 5$ consecutive chunks during distillation. To accommodate variable aspect ratios in real-world data, we employ a bucketing strategy across both SFT and distillation stages. All experiments utilize a cluster of 32 NVIDIA H20 GPUs with a per-GPU batch size of 1.

Evaluation Metrics. We sampled 75 videos from the HDTF and VFHQ datasets for evaluation and compared our method against state-of-the-art approaches including SadTalker [Zhang et al., 2023], Aniportrait [Wei et al., 2024], EchoMimic [Chen et al., 2025a], Ditto [Li et al., 2024], and Hallo3 [Cui et al., 2025]. Our assessment relies on multiple metrics to provide a comprehensive analysis. We use the Fréchet Inception Distance [Heusel et al., 2017] to measure distributional discrepancies between generated and real frames and the Fréchet Video Distance [Unterthiner et al., 2019] to capture temporal consistency. To assess audio-visual synchronization accuracy and smoothness, we incorporate Sync-C [Chung and Zisserman, 2016] for lip motion alignment and Sync-D for the temporal stability of lip dynamics. Finally, we report inference efficiency in frames per second measured on a single NVIDIA H20 GPU.

4.1 Performance of SoulX-FlashHead

Main Results. We performed a comprehensive quantitative comparison with state-of-the-art (SOTA) methods on two benchmark datasets: HDTF [Zhang et al., 2021] and VFHQ [Xie et al., 2022]. The quantitative results are summarized in Tab. 3 and Tab. 4. On the HDTF dataset, the non-streaming Pro model achieves an FID of 8.31 and outperforms EchoMimic. In the streaming setting, the Pro model yields an FID of 9.97 which surpasses diffusion baselines including Hallo3 and AniPortrait. Regarding video smoothness measured by FVD, the Pro model scores 103.14 in non-streaming and 111.38 in streaming modes. Both results improve upon Sonic which scores 113.31, indicating that our method generates videos with high temporal coherence. Experiments on the VFHQ dataset demonstrate significant advantages in lip synchronization. The Pro model achieves Sync-C scores of 5.60 for non-streaming and 5.53 for streaming. These scores exceed both SadTalker at 4.49 and Sonic at 4.64. These results validate the effectiveness of the Oracle-Guided Bidirectional Distillation strategy in aligning lip movements with audio signals and maintaining synchronization precision in complex natural scenarios. Regarding inference efficiency and streaming strategy analysis, the Lite model

Table 4: Quantitative comparison on the VFHQ dataset.

Method	FID↓	FVD↓	Sync-C↑	Sync-D↓	FPS↑
SadTalker* [Zhang et al., 2023]	29.80	191.81	4.49	8.78	1.60
Aniportrait△ [Wei et al., 2024]	36.58	352.94	1.62	11.73	-
EchoMimic* [Chen et al., 2025a]	24.69	193.45	2.93	10.30	0.79
Ditto* [Li et al., 2024]	27.67	254.05	3.31	10.26	<u>41.24</u>
Hallo3* [Cui et al., 2025]	23.45	171.00	4.19	9.60	0.11
Sonic△ [Ji et al., 2025]	24.03	142.88	4.64	8.48	-
SoulX-FlashHead (Lite)*	16.95	167.90	4.70	8.66	**96**
SoulX-FlashHead (Pro)*	<u>14.05</u>	<u>140.27</u>	<u>5.53</u>	<u>8.01</u>	10.81
SoulX-FlashHead (Lite)△	15.18	156.20	5.33	8.12	-
SoulX-FlashHead (Pro)△	**13.67**	**133.69**	**5.60**	**7.81**	-

achieves an inference speed of 96 FPS. Compared to the lightweight method Ditto, the Lite model maintains real-time performance while delivering superior visual quality. Against computationally intensive models such as Hallo3 and EchoMimic, our approach offers a substantial speed advantage. A comparison between streaming and non-streaming data reveals minimal performance degradation. For instance, on HDTF, the streaming Lite model retains a Sync-C score of 4.21 and a low FVD. This confirms that the Temporal Audio Context Cache effectively mitigates audio feature collapse during real-time streaming while the motion frame sampling strategy successfully alleviates cold-start issues.

In conclusion, the Pro model achieves state-of-the-art visual and synchronization quality, while the Lite model strikes an optimal balance between high fidelity and real-time responsiveness, outperforming existing methods in comprehensive capability.

Long Video Generation. We further evaluated the model on the challenging 60-second long video generation task[4] task at 25 fps against state-of-the-art methods. at 25 fps against state-of-the-art methods. Fig. 3 demonstrates that our model maintains high-fidelity generation capabilities throughout the entire duration. The green indicators highlight severe error accumulation and artifacts in Hallo3 whereas our method ensures structural stability. In terms of lip synchronization marked by yellow boxes, methods relying on abstract motion representations like Ditto and SadTalker exhibit noticeable desynchronization while our approach maintains precise audio-visual alignment. Furthermore, the red regions reveal that SadTalker fails to preserve holistic consistency where structural connections between the headgear and the subject break due to the absence of unified pixel-level modeling. Our method conversely preserves robust holistic integrity.

5 Conclusion

This work introduces SoulX-FlashHead which is a unified framework that effectively reconciles the conflict between high-fidelity video generation and low-latency streaming interaction. By leveraging a 1.3B-parameter DiT backbone, we implemented a robust two-stage training strategy. The Streaming-Aware Spatiotemporal Pre-training ensures stable audio-visual alignment under unstable streaming conditions while the Oracle-Guided Bidirectional Distillation significantly mitigates identity drift and error accumulation in long-sequence autoregressive generation. Extensive experiments demonstrate that our approach achieves state-of-the-art performance on benchmark datasets and offers flexible deployment options ranging from the ultra-fast 96 FPS Lite version to the high-fidelity Pro version.

Despite these advancements, the current framework exhibits limitations primarily stemming from its parameter scale. The 1.3B model possesses a constrained capacity for understanding complex physical dynamics compared to larger-scale foundational models. Consequently, our generation is primarily optimized for facial and head regions. The model currently struggles to synthesize large-amplitude body movements or intricate hand gestures with the same level of precision found in facial features. Future work will explore scaling the architecture to enhance holistic motion modeling while maintaining real-time efficiency.

[4]The original videos for these comparisons can be found on our project page.

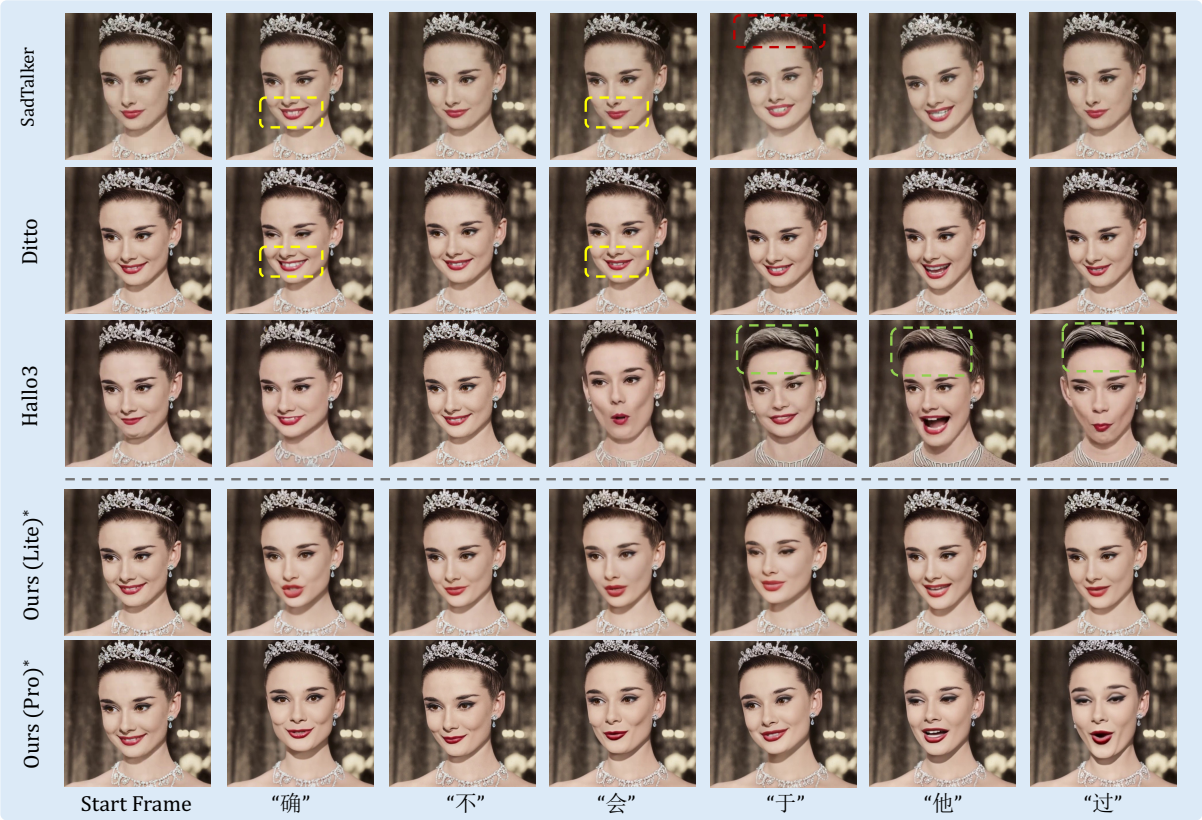

Figure 3: **Qualitative comparison on 60-second video generation at 25 fps.** Yellow dashed regions illustrate lip-synchronization mismatches in motion-based methods like Ditto and SadTalker while green indicators point to severe error accumulation and identity drift in Hallo3. Red boxes reveal holistic inconsistencies where elements like headgear separate from the subject due to the lack of unified pixel latent space modeling. In contrast, SoulX-FlashHead maintains robust lip synchronization, structural integrity, and holistic consistency throughout the sequence.

6 Ethics Statement

This research aims to advance digital human synthesis for beneficial applications. We confirm that all datasets utilized in this study are derived from publicly accessible academic repositories. The visual demonstrations presented in this report are fully synthetic and do not contain the Personally Identifiable Information (PII) of private individuals.

We acknowledge the dual-use nature of high-fidelity video generation technology and the potential risks associated with its misuse, such as the creation of deepfakes or the spread of misinformation. We firmly condemn any malicious application of this technology and advocate for the principles of Responsible AI. To mitigate these risks, we support the development of robust forgery detection algorithms and the implementation of invisible watermarking mechanisms to ensure content transparency and traceability. We remain committed to adhering to ethical guidelines and ensuring that our contributions promote the safe and positive evolution of the field.

References

Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, et al. Wan-s2v: Audio-driven cinematic video generation. *arXiv preprint arXiv:2508.18621*, 2025.

Zhizhou Zhong, Yicheng Ji, Zhe Kong, Yiying Liu, Jiarui Wang, Jiasun Feng, Lupeng Liu, Xiangyi Wang, Yanjia Li, Yuqing She, et al. Anytalker: Scaling multi-person talking video generation with interactivity refinement. *arXiv preprint arXiv:2511.23475*, 2025.

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

Shaoshu Yang, Zhe Kong, Feng Gao, Meng Cheng, Xiangyu Liu, Yong Zhang, Zhuoliang Kang, Wenhan Luo, Xunliang Cai, Ran He, et al. Infinitetalk: Audio-driven video generation for sparse-frame video dubbing. *arXiv preprint arXiv:2508.14033*, 2025.

Yubo Huang, Hailong Guo, Fangtai Wu, Shifeng Zhang, Shijie Huang, Qijun Gan, Lin Liu, Sirui Zhao, Enhong Chen, Jiaming Liu, et al. Live avatar: Streaming real-time audio-driven avatar generation with infinite length. *arXiv preprint arXiv:2512.04677*, 2025.

Le Shen, Qiao Qian, Tan Yu, Ke Zhou, Tianhang Yu, Yu Zhan, Zhenjie Wang, Ming Tao, Shunshun Yin, and Siyuan Liu. Soulx-livetalk: Real-time infinite streaming of audio-driven avatars via self-correcting bidirectional distillation. *arXiv e-prints*, pages arXiv–2512, 2025.

Tianqi Li, Ruobing Zheng, Minghui Yang, Jingdong Chen, and Ming Yang. Ditto: Motion-space diffusion for controllable realtime talking head synthesis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9704–9713, 2025.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8652–8661, 2023.

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, 2025.

Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024.

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

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6613–6623, 2024.

Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2403–2410, 2025a.

Tianqi Li, Ruobing Zheng, Minghui Yang, Jingdong Chen, and Ming Yang. Ditto: Motion-space diffusion for controllable realtime talking head synthesis. *arXiv preprint arXiv:2411.19509*, 2024.

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.

Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020.

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.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

Shunian Chen, Hejin Huang, Yexin Liu, Zihan Ye, Pengcheng Chen, Chenghao Zhu, Michael Guan, Rongsheng Wang, Junying Chen, Guanbin Li, Ser-Nam Lim, Harry Yang, and Benyou Wang. Talkvid: A large-scale diversified dataset for audio-driven talking head synthesis, 2025b. URL https://arxiv.org/abs/2508.13618.

Youliang Zhang, Zhaoyang Li, Duomin Wang, Jiahe Zhang, Deyu Zhou, Zixin Yin, Xili Dai, Gang Yu, and Xiu Li. Speakervid-5m: A large-scale high-quality dataset for audio-visual dyadic interactive human generation. *arXiv preprint arXiv:2507.09862*, 2025.

Brandon Castellano. PySceneDetect: Video Cut Detection and Analysis Tool, 2024. URL https://github.com/Breakthrough/PySceneDetect.

Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023.

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, 2016.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

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.

Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 657–666, 2022.