RI-AVATAR: STRUCTURE-AWARE REAL-TIME STREAMING FOR PERSONALIZED AVATARS

Anonymous authors
Paper under double-blind review

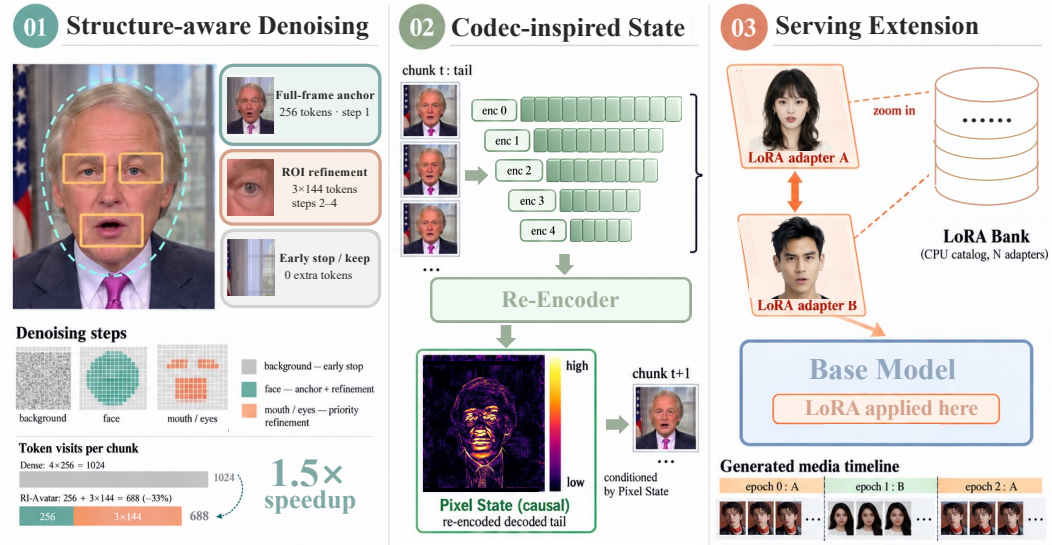


Figure 1: **Overview of RI-Avatar. Structured region-aware denoising (SRAD)** uses one full-frame anchor step followed by three face-ROI refinement steps, reducing per-chunk token visits from 4×256 to $256 + 3 \times 144$. **Codec-inspired Pixel State** re-encodes the decoded tail of each chunk as causal state for the next chunk, preserving continuity across chunk boundaries. **Stateful serving extension** maintains CPU-resident identity bundles and bounded active/standby GPU slots, with epoch-tagged handoff coordinated on the media timeline.

ABSTRACT

Speech-driven avatars require coherent real-time generation under a tight latency budget while maintaining temporal continuity across streaming chunks. Existing few-step diffusion avatars still spend uniform denoising computation over regions whose motion importance is highly non-uniform, while imperfect state propagation can introduce visible artifacts at chunk boundaries. We present RI-Avatar, a real-time streaming framework centered on two complementary ideas. First, structured region-aware denoising (SRAD) anchors each chunk with a full-frame step and concentrates subsequent refinement on speech-sensitive facial regions. Second, a codec-inspired Pixel State re-encodes the decoded RGB tail of each chunk as causal state for the next, preserving identity, pose and expression across chunk boundaries. On the unified HDTF/VFHQ evaluation, the SRAD route achieves a 1.45× resident speedup over the official native FlashHead Lite baseline while providing a substantially better quality–efficiency trade-off than matched token-merging and pruning controls. With the final RGB Pixel State runtime optimizations enabled, a separate representative 22.4-s continuous-speech stream reduces resident generation time from 8.149 s to 4.370 s, reaching 1.865× speedup on a single A800. As a deployment extension, we additionally implement a stateful identity handoff path with CPU-resident identity bundles, bounded active/standby GPU slots, and epoch-tagged media-timeline cutover; an A→B→A serving smoke verifies bundle and timestamp ordering during continuous generation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

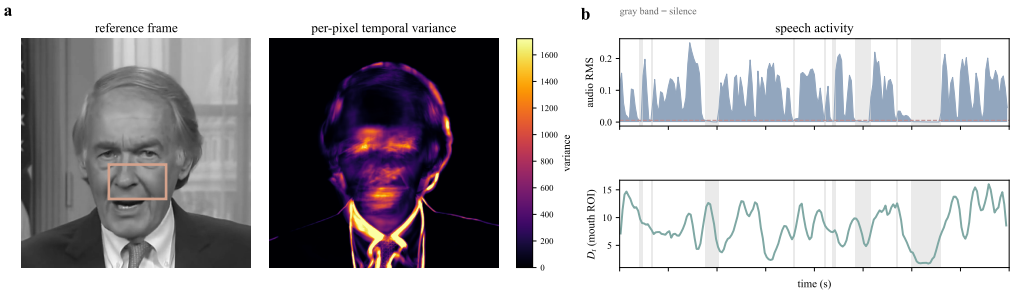


Figure 2: **Speech-driven change is spatially structured.** (a) Per-pixel temporal variance over an 8-s HDTF clip (200 frames): change concentrates on the mouth, jaw and collar region, while the background remains near zero. (b) Mean absolute frame difference within the mouth ROI (bottom) tracks speech activity (top; gray bands mark silence): D_t rises during speech and drops near zero in silence. This motivates spending denoising steps on the audio-sensitive face region rather than uniformly.

1 Introduction

Speech-driven digital humans are becoming an interface for conversation, live interaction and remote communication. Recent diffusion talking-head systems (Yu et al., 2026; Sun et al., 2026; Song et al., 2026; Huang et al., 2025; Cui et al., 2026; Li et al., 2025b) demonstrate high visual quality and real-time streaming generation. SoulX-FlashHead Lite, for example, reports 96 FPS on a single RTX 4090 (Yu et al., 2026). However, their operating assumption is usually a single identity whose reference and temporal state are already resident on the GPU. A practical interactive system is expected to do all three at once on a single commodity GPU: stream in real time, preserve temporal coherence across chunks, and switch identities instantly while the current stream continues playing. In a live support channel or a virtual meeting, a user may hand the conversation to a different avatar at any moment, and any loading pause is directly visible to viewers.

This setting creates two coupled bottlenecks. First, every chunk repeatedly evaluates a video diffusion model even though the perceptual importance of its spatial regions is highly non-uniform: mouth and jaw motion is strongly driven by speech (Thomas & Jordan, 2004; Park et al., 2022), while background and slowly changing appearance need less frequent refinement. Figure 2 makes this concrete: per-pixel temporal variance over an 8-s talking-face clip concentrates on the mouth, jaw and collar region while the background stays near zero, and frame-to-frame change inside the mouth region tracks speech activity over time. Second, loading a new identity involves more than replacing a parameter file. Reference latents, personalization parameters, audio context and rolling motion state are kept mutually consistent at a media timestamp.

These two bottlenecks are usually attacked by separate subsystems, with a faster generator addressing one and a faster adapter loader addressing the other. We observe that they in fact share one resource and one clock: both draw from the per-chunk compute budget of a single GPU, and both answer to the same media timeline. Region-wise denoising frees computation inside the current stream, and identity preparation can consume exactly that freed computation without missing playback deadlines. The bottleneck is therefore not simply model size or PCIe bandwidth, but the absence of a joint schedule.

Based on this insight, we introduce RI-Avatar, whose core streaming path combines two mechanisms. (i) Structured region-aware denoising assigns refinement computation according to speech-driven spatial importance while retaining a full-frame anchor for global structure. (ii) A decoded Pixel State from the previous chunk provides causal visual history for the next chunk, turning streaming generation into a differential update process. We further integrate this streaming path with a stateful serving extension that keeps personalized identity bundles in a CPU-resident catalog and coordinates their activation with chunk boundaries on the media timeline.

Our experiments evaluate spatial allocation, final-runtime optimization, and the handoff implementation at distinct scopes. On the unified 150-case HDTF/VFHQ benchmark with the legacy latent-tail state shared across all routes, RI-Avatar achieves a 1.5× resident speedup over the official native baseline while retaining a substantially better quality–efficiency trade-off than token-merging and pruning controls. A matched pilot additionally compares quality and throughput against seven external talking-face systems. Separately, with the final RGB Pixel State runtime optimizations enabled, a representative 22.4-s continuous-speech stream reaches 1.8× speedup. The serving test verifies an A-to-B-to-A session with monotonic epochs, PTS and frame indices while preserving the base-model checksum.

Our contributions are:

- We propose structured region-aware denoising (SRAD) for few-step streaming talking heads: one full-frame anchor step preserves global structure, while subsequent steps concentrate computation on speech-sensitive facial regions, substantially reducing spatial token visits.
- We introduce a codec-inspired Pixel State that re-encodes the decoded RGB tail of each chunk as causal visual state for the next chunk, improving continuity across chunk boundaries without an additional learned temporal module.
- We integrate SRAD and Pixel State into a real-time personalized-avatar runtime, and further implement a stateful serving extension with CPU-resident identity bundles, bounded active/standby GPU slots, and epoch-tagged media-timeline handoff. Our serving smoke test validates functional bundle and timeline ordering, while the main empirical evaluation focuses on generation efficiency and visual quality.

2 Related Work

Streaming talking-head avatars and temporal state. Diffusion-based talking-head systems have reached real-time, long-horizon generation. SoulX-FlashHead (Yu et al., 2026) provides our backbone and reports 96 FPS for its Lite variant on a single RTX 4090; StreamAvatar (Sun et al., 2026) distills a human-video diffusion model into an autoregressive real-time generator; Live Avatar (Huang et al., 2025), AvatarForcing (Cui et al., 2026) and Ditto (Li et al., 2025b) stream audio-driven avatars at infinite length or in few steps; EchoMimic (Chen et al., 2025; Meng et al., 2025b) animates portraits from audio and editable landmarks; InteractiveAvatar (Song et al., 2026) switches caches to react to user intent within a stream; and StreamDiffusionV2 (Feng et al., 2026) studies SLO-aware batching and pipelining for streaming video diffusion. A second axis is how temporal context is carried across chunks: neural video compression has shifted from predictive residue coding toward conditional coding with learned contexts (Li et al., 2021), where decoded frames serve as the state from which the next frame is produced. RI-Avatar borrows this intuition as a deterministic Pixel State reuse, and addresses the complementary problem of allocating denoising computation across semantic regions of the current stream and preparing a different identity without interrupting playback.

Efficient and region-adaptive diffusion. Training-free acceleration methods largely treat the visual field as a homogeneous token set: token merging (Bolya et al., 2023; Bolya & Hoffman, 2023), attention-driven token pruning (Wang et al., 2024), token sparsification (Chang et al., 2025) and timestep-embedding-guided feature caching (Liu et al., 2025). Region-adaptive variants go further: RAS reuses predicted noise outside selected regions (Liu et al., 2026); ProfilingDiT profiles foreground- and background-focused blocks for feature reuse (Ma et al., 2025); SDiT schedules updates by semantic segmentation and regional complexity (Lin et al., 2026); and LightningCP combines caching-based parallel denoising with decoupled foreground attention for U-Net-based talking heads (Long et al., 2026). Together they establish that spatially uniform computation is unnecessary. Existing region-adaptive acceleration methods have demonstrated the benefit of spatially non-uniform computation in image and general video diffusion. However, their released implementations are not directly compatible with our few-step streaming DiT backbone; even talking-head-specific LightningCP is implemented under substantially different architectures and inference protocols. We therefore use backbone-matched token reduction methods as controlled efficiency baselines. RI-

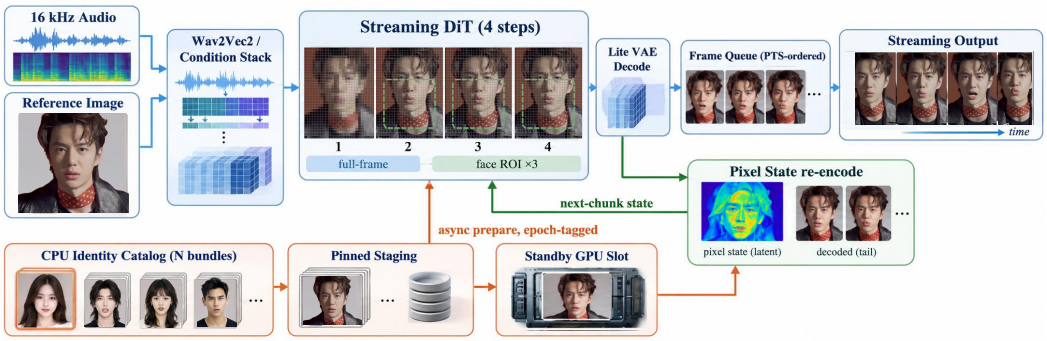


Figure 3: System overview of RI-Avatar. The blue path is the PTS-ordered streaming generator; the green loop re-encodes decoded frames as the next chunk's Pixel State; the orange path asynchronously stages an epoch-tagged identity bundle in a standby GPU slot.

Avatar applies this principle to audio-conditioned, few-step streaming, choosing a fixed-shape route by a causal controller before each chunk is denoised.

Personalized generation and serving. LoRA (Hu et al., 2022) is the standard vehicle for compact identity specialization. LLM serving systems page and batch thousands of adapters (Sheng et al., 2024; Chen et al., 2024) and provide the scheduling substrate (Kwon et al., 2023; Zheng et al., 2024) on which diffusion serving builds. Katz (Li et al., 2025a) serves diffusion models with many adapters , overlapping bounded asynchronous LoRA loading with the target request's early denoising steps. RI-Avatar instead targets a different regime: switching identity inside a live stream without interrupting playback. Orthogonally, portrait-animation work disentangles identity from motion at generation time (Xu et al., 2025; Shi et al., 2025; Luo et al., 2025).

3 Method

Figures 1 and 3 summarize the design. RI-Avatar treats a stream as a sequence of chunks on one media timeline (Section 3.1), spends the chunk budget where speech makes it matter (Section 3.2), carries a causal decoded state across chunks (Section 3.3), and prepares and commits identity switches on the same timeline (Section 3.4).

3.1 Streaming formulation

Let a stream be divided into chunks of M frames. The first L frames overlap with the previous chunk and $M - L$ frames are committed to playback. For chunk t, the system receives a reference image r, audio embeddings a_t, fresh noise z_t, and a Pixel State s_{t-1} obtained from the previous decoded tail. The output is

$$y_t = \mathcal{D}_{\mathrm{VAE}}\left(\mathcal{G}(r, a_t, s_{t-1}, z_t)\right), \qquad s_t = \mathcal{E}_{\mathrm{VAE}}(y_t[-L:]). \tag{1}$$

Only the $M - L$ new frames are emitted. The overlap remains internal state, preventing the visual reset that otherwise creates chunk clusters and flicker.

3.2 Structured region-aware denoising

A uniform schedule spends equal computation on every region, although only part of a talking face carries the speech signal: mouth and jaw follow the phonemes, while background and hair vary slowly. RI-Avatar therefore anchors each chunk with one full-frame step that preserves global structure and concentrates the remaining steps on an audio-sensitive face region (Figure 4).

The reference image is encoded once and paired with a face detector and a causal face lock. The detector provides a raw face box; a fixed-shape $n_b \times n_b$ latent bucket defines the protected region, while expanded feather masks cover eyes, mouth and boundary pixels. A context halo, when enabled, enlarges the model's crop without enlarging the region written back. Wav2Vec2 (Baevski

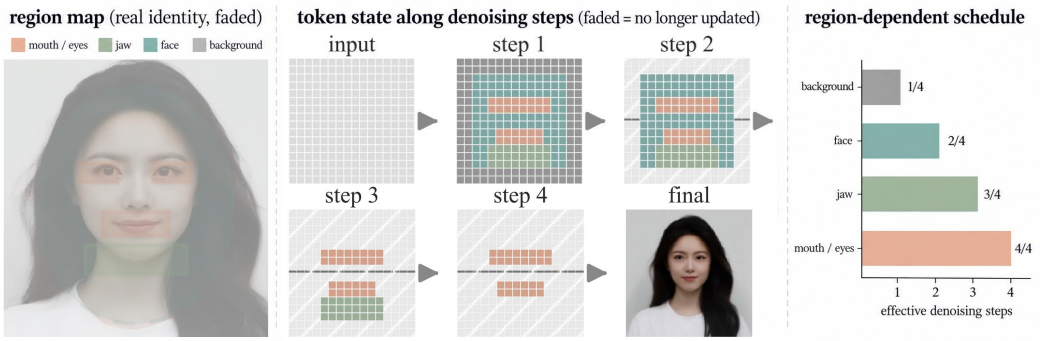


Figure 4: **Structured region-aware denoising (SRAD).** Left: region map on a faded real identity, where the background, face, jaw, and mouth/eyes carry different speech-driven importance. Middle: token state along the four denoising steps; faded regions are no longer updated. Step 1 refreshes the full grid, step 2 the face ROI, step 3 the jaw and mouth/eyes, and step 4 the mouth/eyes only. Right: resulting effective steps per region, where important regions receive more computation while unimportant regions stop earlier.

et al., 2020) produces frame-level embeddings from 16-kHz PCM audio. The audio-gated extension reduces refinement during silence; its separate routing experiment is reported below.

The first diffusion step processes the full $G \times G$ latent grid to establish global structure. Subsequent steps refine only the dynamic face ROI, including mouth and inner-face regions. The write-back uses feathered masks and temporal consistency. If detection or face lock fails, the system falls back to FlashHead Lite (Dense). This fixed-shape design preserves efficient kernels and avoids expanding the ROI to fit every detected face box.

Compute allocation. Let N be the full spatial token count and N_c the token count of the refinement crop, including any context halo. With K denoising steps, the spatial token visits decrease from KN to $N + (K-1)N_c$; the concrete bucket, grid and step configuration used in our experiments is given in Section 4.1.

3.3 Pixel State and Runtime Path

A streaming generator also decides what temporal context the next chunk sees: resetting this context produces chunk-boundary flicker, while regenerating it wastes budget. Pixel State is the codec-inspired part of RI-Avatar. The previous L complete RGB frames are re-encoded with the Lite VAE (Rombach et al., 2022) and fed into the next chunk as global state. This causal loop keeps identity, pose and expression aligned across chunks. We compile only the VAE encoder path used for state extraction; posterior sampling remains eager to preserve the tested behavior. VAE decode, resident warmup, preallocated frame collection, asynchronous device-to-host transfer and fixed-rate queue playback complete the deployment path.

3.4 Stateful Avatar Handoff

Serving the switch is the systems half of the problem: the current identity keeps rendering while the next one becomes ready, and no frame may mix the two. For a personalized identity, the serving bundle contains a content-addressed LoRA (Hu et al., 2022) adapter, reference image and derived reference state, identity metadata and a compatibility key. The SGLang (Zheng et al., 2024) worker owns a CPU-resident catalog and bounded active/standby GPU slots. When a request switches from A to B, B is transferred and prepared while A continues to render. A shadow pre-roll uses B's reference and the current audio history to build B-specific temporal state. The scheduler then atomically commits the complete bundle, temporal state, generation epoch and switch-event sequence at a media timestamp. The CyberVerse plugin forwards session, chunk, switch and displayed-watermark events, while the Go media fence rejects stale epochs and out-of-order PTS. The checked-in A-to-B-to-A smoke exercises this implementation on one visible session.

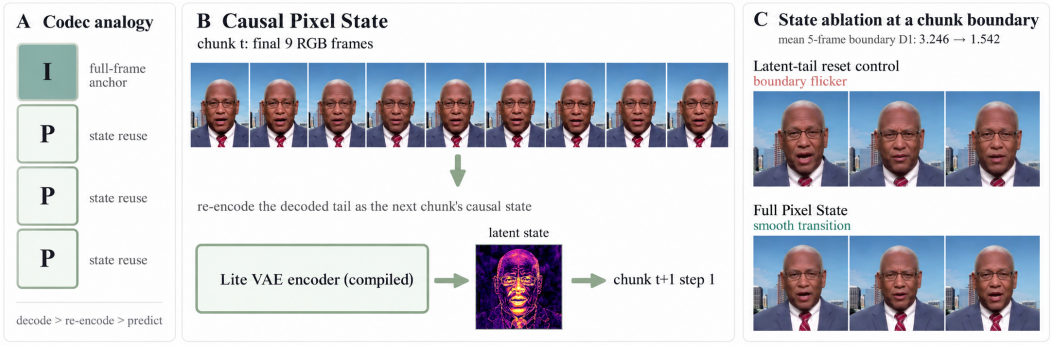


Figure 5: **Codec-inspired Pixel State.** The complete nine-frame decoded RGB tail of chunk t is re-encoded by the compiled Lite VAE encoder and conditions the first denoising step of chunk $t+1$. The real HDTF frames compare a latent-tail reset control with Full Pixel State at the selected chunk boundary: mean five-frame boundary D1 decreases from 3.246 to 1.542. Both outputs use the same identity and four-step backbone.

Table 1: Unified HDTF/VFHQ evaluation with legacy latent-tail state on all routes. Lower is better for FID, FVD and Sync-D; higher is better for Sync-C, coverage and FPS. FPS is derived from the pooled 448-frame resident times of Table 2 and is therefore identical for both datasets; speedup ratios against the official native baseline are reported in Table 2. **Bold**: best per column; shaded row: ours.

Dataset	Route	FID	FVD	Sync-C	Sync-D	Cov.	FPS
HDTF	FlashHead Lite (Dense)	**28.47**	**183.30**	**6.232**	**8.689**	100%	107.02
HDTF	ToMe 50%	36.35	556.00	5.518	9.109	100%	113.91
HDTF	Pruning 50%	46.63	1257.66	3.298	10.409	96.0%	122.57
HDTF	RI-Avatar (Ours)	30.04	289.17	6.122	8.702	100%	**155.12**
VFHQ	FlashHead Lite (Dense)	**41.34**	**271.92**	**4.543**	9.133	98.7%	107.02
VFHQ	ToMe 50%	43.57	473.64	4.088	9.256	97.3%	113.91
VFHQ	Pruning 50%	53.18	979.32	2.927	9.774	88.0%	122.57
VFHQ	RI-Avatar (Ours)	43.62	361.16	4.482	**9.067**	98.7%	**155.12**

4 Experiments

4.1 Evaluation protocol

We use the official FlashHead Lite 1.3B backbone (Yu et al., 2026) with four diffusion steps. HDTF (Zhang et al., 2021) and VFHQ (Xie et al., 2022) each contain 75 cases in the evaluated manifests, 150 cases in total. All four routes use the same Image2 references, audio and evaluator. We report FID (Heusel et al., 2017), FVD (Unterthiner et al., 2018), Sync-C and Sync-D (Chung & Zisserman, 2016) and SyncNet coverage; given the documented content bias of FVD (Ge et al., 2024), we report the complete metric vector rather than any single scalar. Published results under original references are not directly comparable to this protocol. Measurement and configuration details are deferred to Appendix A.

4.2 Comparison with Efficient Token-Reduction Solutions

We compare RI-Avatar against efficient token-reduction solutions on the same backbone: ToMe (Bolya et al., 2023; Bolya & Hoffman, 2023) (token merging) and Pruning (magnitude-based token pooling, cf. (Wang et al., 2024)), both at 50% retention in the selected blocks. FlashHead Lite (Dense) (*Dense* for short) is the uncompressed quality oracle and the efficiency denominator; RI-Avatar (Ours) (*Ours* for short) is our spatial route. Table 1 reports quality and Table 2 reports steady-state speed.

Table 2: Steady-state resident generation time under the legacy single-stream protocol (one 448-frame video, 16 chunks, legacy latent-tail state), with the official native CyberVerse FlashHead Lite as the sole speedup denominator. Quality for these routes is reported in Table 1 under the 150-case batch protocol. Shaded row: ours.

Route	Time (s)	Speedup
FlashHead Lite (Dense, official native)	4.186	1.000x
ToMe 50%	3.933	1.064x
Pruning 50%	3.655	1.145x
RI-Avatar (Ours)	2.888	**1.449x**

Table 3: Quantitative comparison on the HDTF pilot. * streaming; $^\triangle$ non-streaming (FPS not meaningful). FPS† is quoted from Yu et al. (2026) (single H20); FPS§ is our cold-process A800 measurement including model loading; internal-route FPS is derived from the legacy-protocol resident times of Table 2 (448-frame workload, pooled over datasets). EchoMimicV3 FPS is omitted because the value reported by Yu et al. (2026) covers EchoMimic v1. **Bold**: best; underlined: second.

Method	FID↓	FVD↓	Sync-C↑	Sync-D↓	FPS↑
SadTalker* (Zhang et al., 2023)	39.32	724.17	**6.93**	**8.00**	2.17†
AniPortrait$^\triangle$ (Wei et al., 2024)	54.58	1288.14	0.59	13.27	—
EchoMimicV3* (Meng et al., 2025a)	60.75	1335.31	4.62	9.88	—
Ditto* (Li et al., 2025b)	48.97	960.03	4.07	10.57	45.04†
Hallo3* (Cui et al., 2024)	58.04	1277.11	6.44	8.71	0.16†
JoyVASA$^\triangle$ (Cao et al., 2024)	**19.35**	861.67	6.62	8.18	1.76§
Hallo2$^\triangle$ (Cui et al., 2025)	27.94	**311.10**	6.74	8.00	0.18§
FlashHead Lite (Dense)* (Yu et al., 2026)	35.88	435.23	5.41	9.21	107.02
ToMe 50%* (Bolya et al., 2023; Bolya & Hoffman, 2023)	47.66	923.46	5.06	9.42	113.91
Pruning 50%*	46.98	1272.84	4.08	10.08	122.57
RI-Avatar (Ours)*	37.90	502.21	5.82	8.82	**155.12**

All three accelerated routes run faster than the official native Dense Lite, but by very different margins: ToMe 50% and Pruning 50% reach only 1.064x and 1.145x, while Ours reaches 1.449x (Table 2). Ours has lower FVD, higher Sync-C and lower Sync-D than both token controls on both datasets, with coverage equal to or higher than theirs. The exception to scalar dominance over ToMe is VFHQ FID: Ours scores 43.6193 versus 43.5704. Dense remains the stronger FVD reference, at 183.30 versus 289.17 on HDTF and 271.92 versus 361.16 on VFHQ. Thus the evidence supports a quality–efficiency trade-off, rather than quality equivalence to Dense.

4.3 Comparison with State-of-the-Art Methods

We further place RI-Avatar against seven external talking-face systems on a matched pilot: three HDTF and three VFHQ cases share identical Image2 references, audio and the unified evaluator across all methods, and synchronization is scored with the official SyncNet LSE (Chung & Zisserman, 2016). Tables 3 and 4 report quality and throughput side by side. Among the external systems, SadTalker (Zhang et al., 2023) and Hallo2 (Cui et al., 2025) reach the best synchronization scores. RI-Avatar (Ours) is the fastest system in the table and, despite spending most denoising steps on a face crop, improves Sync-C and Sync-D over the FlashHead Lite (Dense) backbone on both datasets, consistent with the observation that speech structures the perceptual value of denoising computation. In short, RI-Avatar retains quality comparable to the dense backbone at significantly higher throughput. External FPS values are quoted from the SoulX-FlashHead technical report (Yu et al., 2026), measured on a single H20 under its own protocol; given the hardware and protocol differences, they serve only as indicative throughput references, and we claim speedups solely against our own A800 measurements.

Token reduction alone gives limited acceleration in this four-step setting: measured against the official native Dense Lite, ToMe 50% and Pruning 50% reach only 1.064x and 1.145x because

Table 4: Quantitative comparison on the VFHQ pilot. Protocol, symbols and FPS provenance as in Table 3. **Bold**: best; underlined: second.

Method	FID↓	FVD↓	Sync-C↑	Sync-D↓	FPS↑
SadTalker* (Zhang et al., 2023)	102.52	1644.43	4.94	**8.11**	1.60†
AniPortrait△ (Wei et al., 2024)	137.56	1496.08	0.75	12.45	—
EchoMimicV3* (Meng et al., 2025a)	130.38	1496.85	3.21	9.84	—
Ditto* (Li et al., 2025b)	142.68	1592.05	3.69	9.77	41.24†
Hallo3* (Cui et al., 2024)	142.88	1560.90	4.69	9.47	0.11†
JoyVASA△ (Cao et al., 2024)	76.02	**1174.41**	4.57	9.04	1.76§
Hallo2△ (Cui et al., 2025)	**75.30**	1176.34	**5.00**	8.96	0.18§
FlashHead Lite (Dense)*	106.50	1209.46	4.38	8.89	107.02
ToMe 50%*	118.09	1518.72	4.17	8.67	113.91
Pruning 50%*	123.33	1783.43	3.57	9.13	122.57
RI-Avatar (Ours)*	113.28	1542.84	4.55	8.62	**155.12**

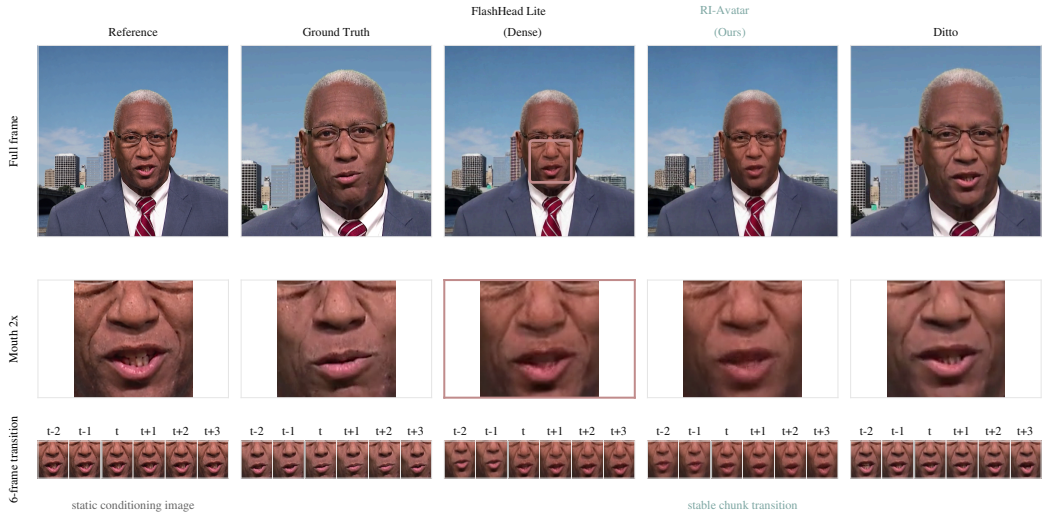


Figure 6: **Qualitative comparison on a matched HDTF pilot case.** All panels use real frames from the same identity, audio segment, and video time under the Image2-reference protocol. The selected 24-frame chunk boundary is the measured point at which FlashHead Lite (Dense) has the largest normalized mouth-region discontinuity relative to RI-Avatar. RI-Avatar preserves a smoother six-frame mouth trajectory, while Ditto provides a real-time external comparison. Red marks the measured Dense boundary artifact; teal marks the stable RI-Avatar transition.

merge and prune overheads consume most of the nominal compute savings, and both lose substantial quality (Tables 1 and 2).

4.4 Why the route works

We first separate the routing ablation from the full-dataset comparison: the ablation tests whether temporal and spatial allocation can reduce resident latency on an interrupted-speech input. It uses one RTX4090, the same 8.4-s interrupted-speech input at 20 FPS, and three repeated runs per route; all 12 runs use legacy latent-tail state and complete with zero fallback chunks. Table 5 reports the latency of temporal and spatial routing under that configuration.



Figure 7: **Qualitative results across identities.** Ground truth, FlashHead Lite (Dense) and RI-Avatar (Ours) frames at the same video time on ten held-out HDTF identities. RI-Avatar preserves identity, pose and lip shape close to the dense backbone.

Table 5: Routing ablation on interrupted-speech samples, three repeated runs per route. Mean resident time includes generation and state updates; these are legacy-state measurements.

Route	Wall (s)	Speedup
FlashHead Lite (Dense)	1.548	1.000×
Temporal	1.164	1.330×
Audio–Spatial	1.263	1.226×
RI-Avatar (Ours)	1.043	1.484×

The measured gain is concentrated in DiT (Peebles & Xie, 2023) computation: the full-frame first step preserves global layout, while later steps visit a face crop. Pixel State retains decoded history across chunks, while the audio path retains the conditioning signal that controls mouth motion. Previous-frame conditioning is inherited from the FlashHead streaming design (Yu et al., 2026); our runtime preserves its RGB-tail state semantics while optimizing its encoding and scheduling alongside spatial refinement. Separately, with the final RGB Pixel State runtime optimizations enabled, the representative 22.4-s continuous-speech setting on one A800 reduces resident generation time from 8.149 s to 4.370 s (1.865×; measurement setting and stage breakdown in Appendix A).

5 Limitations and Discussion

The unified 150-case benchmark and the final RGB Pixel State runtime measurement use different evaluation configurations, so we report the 1.5× benchmark speedup and the 1.8× final-runtime result separately. Image2 preprocessing and affine alignment alter the reference distribution and may not preserve all appearance or pose characteristics, limiting direct comparison with results reported under original-reference settings. The current handoff evaluation covers one visible session. Broader evaluation of concurrency, tail latency, CPU-offload working set, and full-dataset final-runtime behavior remains future work.

6 Conclusion

RI-Avatar organizes streaming generation around spatial importance, causal state, and identity handoff for real-time personalized avatars on a single commodity GPU. On the unified 150-case benchmark, its spatial route achieves a 1.5× resident speedup over the official native baseline while providing a favorable quality–efficiency trade-off relative to the tested token-merging and pruning controls. With the final RGB Pixel State runtime optimizations enabled, a separate representative 22.4-s continuous-speech stream reaches 1.8× acceleration. The SGLang implementation further demonstrates an ordered A-to-B-to-A personalized handoff on a single media timeline. Extending final-runtime and serving evaluation to larger-scale concurrent workloads is an important direction for future work.

AI USE STATEMENT

Generative AI tools were used for limited assistance with manuscript editing, reference search and formatting, and visual refinement of qualitative figures. All scientific content, including the method design, system implementation, experimental design, measurements, and analysis, was produced by the authors. All AI-assisted material was manually reviewed and verified. Figures were constructed from the authors' own system outputs, references were checked against the original publications or official records, and AI-assisted text was validated against the underlying experiments. The authors take full responsibility for the final manuscript and all reported results.

ETHICS STATEMENT

Our experiments use the HDTF (Zhang et al., 2021) and VFHQ (Xie et al., 2022) datasets, publicly released and research-licensed collections of talking-face videos. A small number of personalized adapters were trained on publicly available, publicly licensed photographs of public figures, used solely for internal evaluation of the identity-switching mechanism; no such material is redistributed, and all generated outputs shown are synthetic. This research targets beneficial applications such as accessible communication and remote presence. We acknowledge the dual-use nature of high-fidelity talking-face generation, including the potential for deepfakes and misinformation; we condemn any malicious application, and support forgery detection and provenance watermarking as safeguards. We will release our evaluation protocol and metadata under an appropriate research license, together with procedures for reporting provenance or licensing concerns.

REPRODUCIBILITY STATEMENT

The system formulation, method design, and evaluation protocol are specified in the main paper and appendix. Section 3 defines the streaming formulation, structured region-aware denoising, the Pixel State mechanism, and the stateful handoff path. Section 4.1 reports the datasets, metrics, and measurement protocol, and Appendix A reports the concrete chunk, bucket, grid and step configuration, the representative-efficiency measurement setting and stage breakdown, and the serving configuration used in our experiments.

We will release the evaluation manifests, the measurement scripts used to reproduce the main tables and figures, and the serving configuration required to reproduce the reported handoff smoke.

We will additionally release the runtime code and scripts used to reproduce the main results. The backbone weights and the HDTF/VFHQ data are available from their original public releases; we do not redistribute them.

REFERENCES

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020. URL https://arxiv.org/abs/2006.11477.

Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. *arXiv preprint arXiv:2303.17604*, 2023. URL https://arxiv.org/abs/2303.17604.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.09461.

Xuyang Cao, Guoxin Wang, Sheng Shi, Jun Zhao, Yang Yao, Jintao Fei, Minyu Gao, and Pei Xie. JoyVASA: Portrait and animal image animation with diffusion-based audio-driven facial dynamics and head motion generation. *arXiv preprint arXiv:2411.09209*, 2024. URL https://arxiv.org/abs/2411.09209.

Shuning Chang, Pichao Wang, Jiasheng Tang, Fan Wang, and Yi Yang. SparseDiT: Token sparsification for efficient diffusion transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2412.06028.

Lequn Chen, Zihao Ye, Yongji Wu, Danyang Zhuo, Luis Ceze, and Arvind Krishnamurthy. Punica: Multi-tenant LoRA serving. In *Proceedings of Machine Learning and Systems (MLSys)*, 2024. URL https://arxiv.org/abs/2310.18547.

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, pp. 2403–2410, 2025. doi: 10.1609/aaai.v39i3.32241. URL https://ojs.aaai.org/index.php/AAAI/article/view/32241.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016. URL https://www.robots.ox.ac.uk/~vgg/publications/2016/Chung16a/.

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. *arXiv preprint arXiv:2412.00733*, 2024. URL https://arxiv.org/abs/2412.00733.

Jiahao Cui, Hui Li, Yao Yao, Hao Zhu, Hanlin Shang, Kaihui Cheng, Hang Zhou, Siyu Zhu, and Jingdong Wang. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.07718.

Liyuan Cui, Wentao Hu, Wenyuan Zhang, Zesong Yang, Fan Shi, and Xiaoqiang Liu. AvatarForcing: One-step streaming talking avatars via local-future sliding-window denoising. *arXiv preprint arXiv:2603.14331*, 2026. URL https://arxiv.org/abs/2603.14331.

Tianrui Feng, Zhi Li, Shuo Yang, Haocheng Xi, Muyang Li, Xiuyu Li, Lvmin Zhang, Keting Yang, Kelly Peng, Song Han, Maneesh Agrawala, Kurt Keutzer, Akio Kodaira, and Chenfeng Xu. StreamDiffusionV2: A streaming system for dynamic and interactive video generation. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 8, pp. 1656–1669, 2026. URL https://proceedings.mlsys.org/paper_files/paper/2026/hash/698cfaf72a208aef2e78bcac55b74328-Abstract-Conference.html.

Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the content bias in Frechet video distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7277–7288, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ge_On_the_Content_Bias_in_Frechet_Video_Distance_CVPR_2024_paper.html.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2106.09685.

Yubo Huang, Hailong Guo, Fangtai Wu, Weiqiang Wang, Shifeng Zhang, Shijie Huang, Qijun Gan, Lin Liu, Sirui Zhao, Enhong Chen, Jiaming Liu, and Steven Hoi. Live avatar: Streaming real-time audio-driven avatar generation with infinite length. *arXiv preprint arXiv:2512.04677*, 2025. URL https://arxiv.org/abs/2512.04677.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th ACM Symposium on Operating Systems Principles (SOSP)*, 2023. doi: 10.1145/3600006.3613165. URL https://arxiv.org/abs/2309.06180.

Jiahao Li, Bin Li, and Yan Lu. Deep contextual video compression. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. URL https://arxiv.org/abs/2109.15047.

Suyi Li, Lingyun Yang, Xiaoxiao Jiang, Hanfeng Lu, Dakai An, Zhipeng Di, Weiyi Lu, Jiawei Chen, Kan Liu, Yinghao Yu, Tao Lan, Guodong Yang, Lin Qu, Liping Zhang, and Wei Wang. Katz: Efficient workflow serving for diffusion models with many adapters. In *2025 USENIX Annual Technical Conference (USENIX ATC 25)*, pp. 1037–1052. USENIX Association, 2025a. URL https://www.usenix.org/conference/atc25/presentation/li-suyi-katz.

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 ACM International Conference on Multimedia (ACM MM)*, 2025b. doi: 10.1145/3746027.3755075. URL https://arxiv.org/abs/2411.19509.

Bowen Lin, Fanjiang Ye, Yihua Liu, Zhenghui Guo, Boyuan Zhang, Weijian Zheng, Yufan Xu, Tiancheng Xing, Yuke Wang, and Chengming Zhang. SDiT: Semantic region-adaptive for diffusion transformers. *arXiv preprint arXiv:2601.12283*, 2026. URL https://arxiv.org/abs/2601.12283.

Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2411.19108.

Ziming Liu, Yifan Yang, Chengruidong Zhang, Yiqi Zhang, Lili Qiu, Yang You, and Yuqing Yang. Region-adaptive sampling for diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2346–2356, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liu_Region-Adaptive_Sampling_for_Diffusion_Transformers_CVPR_2026_paper.html.

Jianzhi Long, Wenhao Sun, Rong-Cheng Tu, and Dacheng Tao. Lightning fast caching-based parallel denoising prediction for accelerating talking head generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7503–7511, 2026. doi: 10.1609/aaai.v40i9.37690. URL https://ojs.aaai.org/index.php/AAAI/article/view/37690.

Xiangyang Luo, Ye Zhu, Yunfei Liu, Lijian Lin, Cong Wan, Zijian Cai, Yu Li, and Shao-Lun Huang. CanonSwap: High-fidelity and consistent video face swapping via canonical space modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10064–10074, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Luo_CanonSwap_High-Fidelity_and_Consistent_Video_Face_Swapping_via_Canonical_Space_ICCV_2025_paper.html.

Xuran Ma, Yexin Liu, Yaofu Liu, Xianfeng Wu, Mingzhe Zheng, Zihao Wang, Ser-Nam Lim, and Harry Yang. Model reveals what to cache: Profiling-based feature reuse for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17150–17159, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Ma_Model_Reveals_What_to_Cache_Profiling-Based_Feature_Reuse_for_Video_ICCV_2025_paper.html.

Rang Meng, Yan Wang, Weipeng Wu, Ruobing Zheng, Yuming Li, and Chenguang Ma. EchoMimicV3: 1.3B parameters are all you need for unified multi-modal and multi-task human animation. *arXiv preprint arXiv:2507.03905*, 2025a. URL https://arxiv.org/abs/2507.03905.

Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. EchoMimicV2: Towards striking, simplified, and semi-body human animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5489–5498, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Meng_EchoMimicV2_Towards_Striking_Simplified_and_Semi-Body_Human_Animation_CVPR_2025_paper.html.

Se Jin Park, Minsu Kim, Joanna Hong, Jeongsoo Choi, and Yong Man Ro. SyncTalkFace: Talking face generation with precise lip-syncing via audio-lip memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2062–2070, 2022. doi: 10.1609/aaai.v36i2.20102. URL https://ojs.aaai.org/index.php/AAAI/article/view/20102.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

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 (CVPR)*, pp. 10684–10695, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Ying Sheng, Shiyi Cao, Dacheng Li, Coleman Hooper, Nicholas Lee, Shuo Yang, Christopher Chou, Banghua Zhu, Lianmin Zheng, Kurt Keutzer, Joseph E. Gonzalez, and Ion Stoica. SLoRA: Scalable serving of thousands of LoRA adapters. In *Proceedings of Machine Learning and Systems (MLSys)*, 2024. URL https://arxiv.org/abs/2311.03285.

Yuxiang Shi, Zhe Li, Yanwen Wang, Hao Zhu, Xun Cao, and Ligang Liu. DeX-Portrait: Disentangled and expressive portrait animation via explicit and latent motion representations, 2025. URL https://arxiv.org/abs/2512.15524. arXiv preprint.

Quanyue Song, Yishan He, Yanfei Zhang, Shihao Cheng, Zhixiang He, Zhizhi Guo, Chi Zhang, Xuelong Li, and Caigui Jiang. InteractiveAvatar: Real-time streaming video generation for consistent and intent-aware avatars. *arXiv preprint arXiv:2606.22905*, 2026. URL https://arxiv.org/abs/2606.22905.

Zhiyao Sun, Ziqiao Peng, Yifeng Ma, Yi Chen, Zhengguang Zhou, Zixiang Zhou, Guozhen Zhang, Youliang Zhang, Yuan Zhou, Qinglin Lu, and Yong-Jin Liu. StreamAvatar: Streaming diffusion models for real-time interactive human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10887–10897, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Sun_StreamAvatar_Streaming_Diffusion_Models_for_Real-Time_Interactive_Human_Avatars_CVPR_2026_paper.html.

Sharon M. Thomas and Timothy R. Jordan. Contributions of oral and extraoral facial movement to visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 30(5):873–888, 2004. doi: 10.1037/0096-1523.30.5.873. URL https://pubmed.ncbi.nlm.nih.gov/15462626/.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges, 2018. URL https://arxiv.org/abs/1812.01717.

Hongjie Wang, Difan Liu, Yan Kang, Yijun Li, Zhe Lin, Niraj K. Jha, and Yuchen Liu. Attention-driven training-free efficiency enhancement of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2405.05252.

Huawei Wei, Zejun Yang, and Zhisheng Wang. AniPortrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024. URL https://arxiv.org/abs/2403.17694.

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 Workshops (CVPRW)*, pp. 656–665, 2022. URL https://arxiv.org/abs/2205.03409.

Zunnan Xu, Zhentao Yu, Zixiang Zhou, Jun Zhou, Xiaoyu Jin, Fa-ting Hong, Xiaozhong Ji, Junwei Zhu, Chengfei Cai, Shiyu Tang, Qin Lin, Xiu Li, and Qinglin Lu. HunyuanPortrait: Implicit condition control for enhanced portrait animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15909–15919, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Xu_HunyuanPortrait_Implicit_Condition_Control_for_Enhanced_Portrait_Animation_CVPR_2025_paper.html.

Tan Yu, Qian Qiao, Le Shen, Ke Zhou, Jincheng Hu, Dian Sheng, Bo Hu, Haoming Qin, Jun Gao, Changhai Zhou, Shunshun Yin, and Siyuan Liu. SoulX-FlashHead: Oracle-guided generation of infinite real-time streaming talking heads. *arXiv preprint arXiv:2602.07449*, 2026. URL https://arxiv.org/abs/2602.07449.

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 (CVPR)*, 2023. URL https://arxiv.org/abs/2211.12194.

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 (CVPR)*, pp. 3661–3670, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_Flow-Guided_One-Shot_Talking_Face_Generation_With_a_High-Resolution_Audio-Visual_Dataset_CVPR_2021_paper.html.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. SGLang: Efficient execution of structured language model programs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2312.07104.

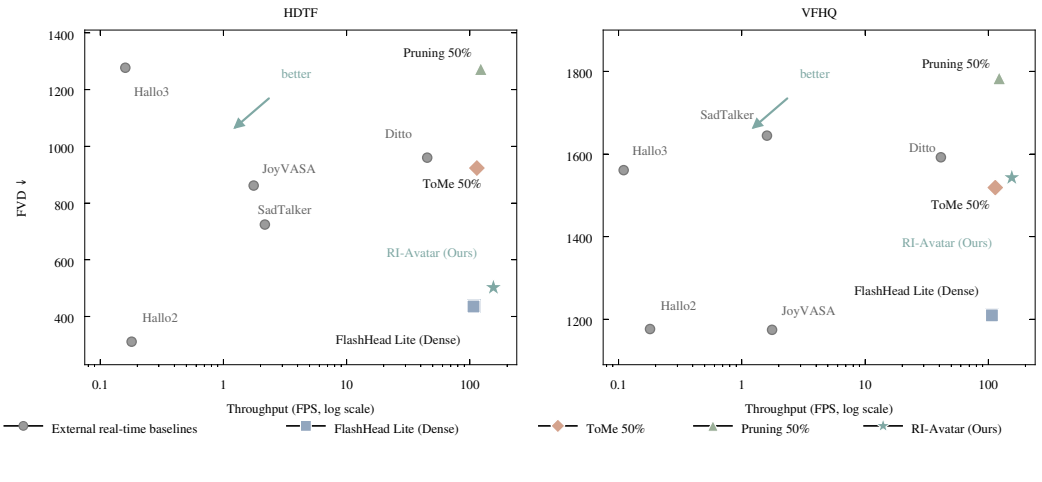


Figure 8: **Quality–throughput trade-off on HDTF and VFHQ.** External real-time baselines are shown in gray; internal FlashHead routes use the common legacy single-stream FPS protocol. RI-Avatar reaches the highest throughput in both panels while retaining substantially lower FVD than the matched ToMe and Pruning controls. EchoMimicV3 is omitted because no comparable FPS measurement is available.

A Evaluation and configuration details

Benchmark protocol. All evaluated videos contain 200 frames at 512x512 and 25 FPS with audio. A frozen reference-only affine inverse transform returns generated videos to the benchmark coordinates. Each stream uses chunks of $M = 33$ frames with an $L = 9$-frame overlap ($M - L = 24$ new frames committed per chunk); Pixel State re-encodes the previous $L = 9$ decoded frames; the face bucket is $n_b \times n_b = 12 \times 12$ within the $G \times G = 16 \times 16$ latent grid with $K = 4$ denoising steps, i.e. $256 + 3 \times 144 = 688$ rather than $4 \times 256 = 1024$ token visits per chunk.

Representative-efficiency breakdown. The measured breakdown of the final route is Audio 0.274 s, Spatial DiT 2.057 s, VAE decode 1.107 s, Pixel State encode 0.322 s and other request overhead 0.603 s. These measurements include generation, state updates and frame return, but exclude model initialization and MP4 muxing.

Serving configuration. The serving branch packages the FlashHead Lite runtime with a pinned RI-Avatar fork and eight runtime patches over SGLang 0.5.8, and registers 11 immutable LoRA/reference bundles covering 10 people. The A-to-B-to-A smoke uses two sample reference identities A and B: each chunk contains 28 frames, PTS values are 0, 1400 and 2800, and the generation epoch advances from 0 to 1 to 2; the base-model checksum is unchanged before and after the switch. These observations support the correctness of bundle identity, epoch and media-timeline ordering in the engineering smoke. The validated configuration is one visible stream with active/standby GPU slots on a single GPU.