Native Audio-Visual Alignment for Generation

Longbin Ji* **Guan Wang*** **Xuan Wei** **Chenye Yang**
Zhenyu Zhang† **Shuohuan Wang** **Yu Sun** **Jingzhou He**

ERNIE Team, Baidu Inc.

{jilongbin, wangguan15, zhangzhenyu07}@baidu.com

*Equal contribution. †Corresponding author.

Project page: ernie-research.github.io/NAVA.

Abstract

Joint audio-video generation aims to synthesize temporally synchronized and semantically coherent visual-acoustic content. However, existing open-source methods mainly rely on either dual-tower designs with posterior alignment or fully unified tri-modal designs that mix textual context, audio, and video in one shared space. The former weakens fine-grained audio-video co-evolution, while the latter couples semantic conditioning with low-level synchronization. To address these limitations, we propose **NAVA**, a *Native Audio-Visual Alignment* framework for joint audio-video generation. NAVA is built upon *context-conditioned native audio-visual alignment*: it first establishes audio-video correspondence in a dedicated interaction space, and then uses external context to condition the joint denoising process. Specifically, NAVA is instantiated with an *Align-then-Fuse MMDiT* architecture, which transitions from modality-aware audio-video alignment to modality-shared joint denoising. Furthermore, we introduce *Timbre-in-Context Conditioning* to associate reference timbre cues with corresponding speech spans to achieve controllable speech timbre. Experiments on Verse-Bench and Seed-TTS, together with a user study, demonstrate that NAVA achieves superior video quality, precise audio-visual synchronization, competitive audio quality, and stronger reference-timbre controllability using only 6.3B parameters.

1 Introduction

Audio-visual generation has made rapid progress in recent years. Compared with cascaded pipelines that synthesize one modality after another, joint audio-video generation models temporal and semantic correspondences within a unified generation process, thereby reducing error propagation and improving cross-modal coherence. Although commercial systems such as Seedance [1], Kling [2], and Veo [3] have demonstrated the potential of joint audio-video synthesis, their architectures and training recipes remain proprietary. Therefore, recent open-source efforts, including Ovi [4], LTX [5], and MoVA [6], have become crucial for reproducible research in audio-visual generation.

Despite this progress, most open-source methods still adopt a dual-tower architecture, where audio and video are generated in separate streams, and cross-modal interaction is introduced through additional alignment modules. As illustrated in Fig. 1(a), the paradigm conditions audio and video on textual context in separate feature spaces, and establishes audio-visual correspondence only through late-stage interaction. However, such posterior alignment weakens the joint evolution of audio and video during generation, making fine-grained synchronization and semantic consistency dependent on auxiliary cross-modal modules rather than a unified generative representation.

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

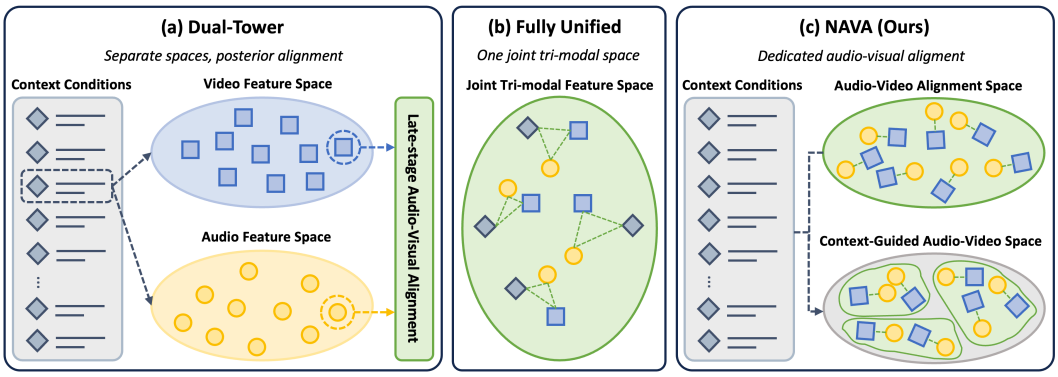

Figure 1: **Audio-visual generation paradigms.** (a) *Dual-Tower*: Separate streams with late alignment. (b) *Fully Unified*: A shared text-audio-video space. (c) *NAVA*: Native audio-video alignment with external context conditioning.

More recently, daVinci-MagiHuman [7] moves beyond dual-tower interaction by placing textual context, video, and audio tokens into a unified attention space for end-to-end tri-modal modeling. As shown in Fig. 1(b), while the design enables direct tri-modal interaction, it also couples high-level semantic control with low-level audio-visual synchronization. Consequently, semantic guidance, event correspondence, and temporal alignment are optimized in the same representation space, which may hinder the formation of a dedicated synchronization structure. This motivates us to separate audio-video correspondence from context conditioning in a dedicated synchronization space.

In this paper, we propose **NAVA**, a *Native Audio-Visual Alignment* framework with decoupled context conditioning. As shown in Fig. 1(c), NAVA first establishes audio-video correspondence in a dedicated alignment space, and then introduces context as external conditioning to guide the aligned representation. This formulation differs from both dual-tower methods, which align audio and video only after separate modeling, and fully unified tri-modal methods, which mix context, audio, and video in a shared space. By decoupling context conditioning from audio-visual synchronization, NAVA focuses its capacity on event-level correspondence, temporal consistency, and collaborative denoising, while remaining compatible with pretrained text-to-video backbones.

To realize this, NAVA employs an *Align-then-Fuse MMDiT* architecture. It first aligns heterogeneous audio and video representations with modality-aware layers, then applies shared fusion layers for compact collaborative denoising. Furthermore, we creatively introduce a *Timbre-in-Context Conditioning* mechanism, which treats timbre cues as contextual conditions for specific speech spans, enabling flexible content-timbre binding without auxiliary speaker-control branches.

In summary, the main contributions of this paper are as follows:

- We propose **NAVA**, a *Native Audio-Visual Alignment* framework that formulates joint audio-video generation as *context-conditioned native audio-visual alignment*, enabling precise event-level correspondence modeling with pretrained video generation backbones.
- We introduce an *Align-then-Fuse MMDiT* architecture for modality-aware audio-video alignment and efficient collaborative denoising, together with *Timbre-in-Context Conditioning* for flexible content-timbre binding across speech segments.
- Extensive experiments and user studies demonstrate that **NAVA** significantly outperforms representative dual-tower and fully unified baselines, achieving superior audio-visual synchronization, semantic consistency, visual quality, and timbre controllability.

2 Method

2.1 Formulation

Let h_a, h_v, and c denote audio tokens, video tokens, and context tokens, respectively. The context c mainly contains textual conditions and can be augmented with control signals such as reference tim-

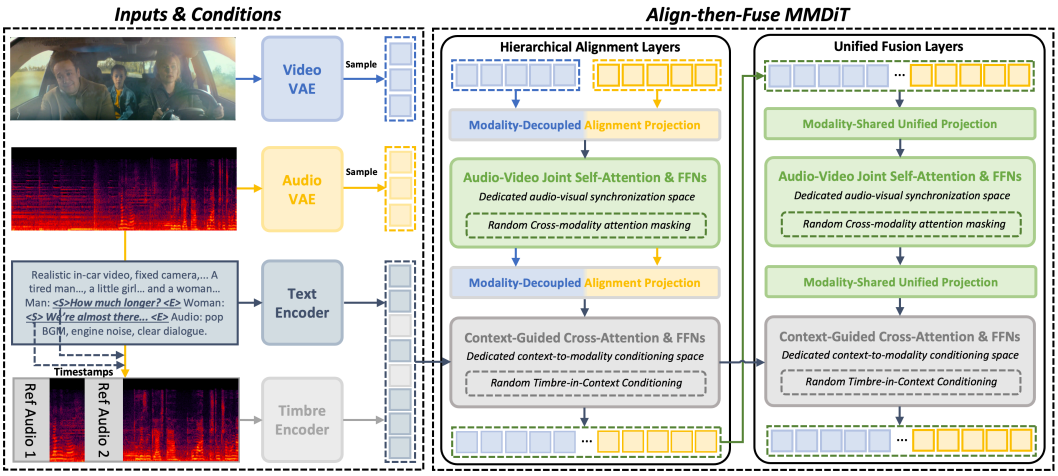

Figure 2: **Overview of NAVA.** *Hierarchical Alignment Layers* establish audio-video correspondence before *Unified Fusion Layers* perform collaborative denoising. Cross-attention injects text and reference timbre, with *Timbre-in-Context Conditioning* binding timbre cues to speech spans.

bre embeddings. We use this notation to abstract how different audio-visual generation paradigms organize audio, video, and context interactions during denoising.

Existing dual-tower methods [4; 5; 6] maintain separate audio and video generation streams and condition each modality independently:

$$\begin{aligned} h'_a &= \mathrm{CrossAttn}(h_a, c), \\ h'_v &= \mathrm{CrossAttn}(h_v, c). \end{aligned} \tag{1}$$

Audio-visual correspondence is then introduced through additional cross-modal interaction modules:

$$[\tilde{h}_a, \tilde{h}_v] = \mathrm{CrossModalAttn}(h'_a, h'_v). \tag{2}$$

This posterior alignment paradigm allows each modality to evolve largely in its own feature space before cross-modal correspondence is explicitly established, making fine-grained synchronization dependent on late-stage interaction.

Fully unified methods [7] instead place context, audio, and video tokens into a single attention space:

$$[\tilde{h}_a, \tilde{h}_v, \tilde{c}] = \mathrm{SelfAttn}([h_a, h_v, c]). \tag{3}$$

This design enables direct tri-modal interaction, but it also entangles high-level semantic conditioning with low-level audio-video synchronization within the same representation space.

In contrast, **NAVA** decouples audio-video synchronization from external context conditioning through *context-conditioned native audio-visual alignment*. Audio and video first interact in a dedicated synchronization space:

$$[h'_a, h'_v] = \mathrm{SelfAttn}([h_a, h_v]), \tag{4}$$

where self-attention is applied over the concatenated audio-video token sequence to form event-level correspondences without inserting context as peer tokens. Context is then injected as external conditioning:

$$[\tilde{h}_a, \tilde{h}_v] = \mathrm{CrossAttn}([h'_a, h'_v], c). \tag{5}$$

In this way, NAVA separates the roles of synchronization and conditioning: joint self-attention learns native audio-video correspondence, while cross-attention provides semantic and controllable guidance from external context.

2.2 Align-then-Fuse MMDiT

To instantiate context-conditioned native audio-visual alignment, NAVA adopts an *Align-then-Fuse MMDiT* architecture, as shown in Fig. 2. Video and audio are first encoded into latent tokens by

separate VAEs, while textual context and optional reference-timbre cues are encoded as conditioning tokens. The architecture follows a progressive design: early layers preserve modality-aware projections to stabilize heterogeneous audio-video interaction, while later layers share generation parameters to encourage compact collaborative denoising. This yields an align-then-fuse process, where audio and video first establish native correspondence and then evolve jointly in a shared generation space.

Hierarchical Alignment Layers. The early layers establish native audio-video correspondence before fully shared generation. Audio spectrogram latents and video latents differ substantially in spatial-temporal structure, token rate, and feature distribution. Directly sharing projections from the first layer can therefore force heterogeneous modalities into a common parameterization too early, suppressing modality-specific representations and destabilizing cross-modal interaction. We address this with *Modality-Decoupled Alignment Projection*, where audio and video tokens are first mapped by modality-specific projections and then placed into a shared audio-video interaction space for stable early-stage correspondence learning.

Within this space, *Audio-Video Joint Self-Attention & FFNs* perform repeated cross-modal interaction during denoising. Unlike posterior alignment modules that operate after separate generation streams, this joint interaction allows acoustic patterns and visual dynamics to co-evolve throughout the denoising process. As a result, event-level correspondences such as speech-lip motion, impact sounds, musical performance, and scene-dependent acoustic changes can be modeled within the generation trajectory itself. To handle token-rate mismatch, we rescale the rotary positional embedding of audio tokens by

$$\theta_{\text{rope}} = \frac{TR_v}{TR_a}, \tag{6}$$

where TR_v and TR_a denote the video and audio token rates, respectively. This rate-aware rescaling places audio and video tokens into a more comparable temporal coordinate system for joint attention.

Context is injected separately through *Context-Guided Cross-Attention & FFNs*. This preserves a dedicated audio-video synchronization space while allowing textual and timbre conditions to modulate the denoising trajectory. Compared with fully unified tri-modal attention, this design avoids inserting context tokens directly into the same self-attention space used for low-level audio-video synchronization.

Unified Fusion Layers. After audio-video correspondence has been established, NAVA transitions to *Unified Fusion Layers*. In these layers, audio and video tokens are processed with *Modality-Shared Unified Projection* and updated by shared transformer blocks. Since the preceding alignment layers have already reduced the representational gap between audio and video tokens, parameter sharing in later layers becomes more stable and efficient. This removes persistent stream separation and encourages compact collaborative denoising in a shared generation space. Context remains external through cross-attention, so semantic guidance and controllable conditions continue to modulate the joint denoising process without disrupting the learned synchronization structure.

2.3 Timbre-in-Context Conditioning

Textual context provides semantic guidance, while speech-driven audio-video generation further requires segment-level timbre control, i.e., specifying *who speaks which content*. We propose *Timbre-in-Context Conditioning*, which represents reference timbre cues as context tokens and binds them to their corresponding speech spans through the existing context-conditioning pathway.

Let $\mathcal{P}$ denote the textual prompt containing speech spans $\{\mathcal{S}_i\}_{i=1}^{N}$, and let $\mathcal{R}_i$ be the reference utterance specifying the desired timbre for $\mathcal{S}_i$. We extract a context-space timbre token as

$$\mathbf{s}_i = E_{\text{tim}}(\mathcal{R}_i), \tag{7}$$

where E_{tim} denotes the timbre encoder. Each speech span is then augmented as

$$\mathcal{S}_i \rightarrow \big[\langle \text{S} \rangle,\, \mathbf{s}_i,\, \text{Text}(\mathcal{S}_i),\, \langle \text{E} \rangle\big], \tag{8}$$

where $\langle \text{S} \rangle$ and $\langle \text{E} \rangle$ mark the boundaries of a timbre-conditioned speech span. Applying this replacement to all speech spans yields the final context sequence:

$$\mathbf{c} = \text{Augment}\left(\mathcal{P}; \{(\mathcal{S}_i, \mathbf{s}_i)\}_{i=1}^{N}\right). \tag{9}$$

During denoising, NAVA accesses this augmented context through context-guided cross-attention. Thus, timbre cues are associated with speech spans within the original prompt structure rather than injected as a global control signal. This is important for multi-speaker generation, where different utterances may require different speaker identities or timbre styles. Because timbre information is represented in the context pathway, the mechanism requires no auxiliary speaker-control branch or backbone modification. It naturally supports compositional control by assigning different timbre tokens to different speech spans, while keeping the audio-video denoising backbone unchanged.

2.4 Training and Inference

2.4.1 Progressive Multi-Task Training

NAVA is trained with a progressive multi-task strategy over T2AV, TI2AV, T2A, T2V, and TIA2AV tasks, covering audio-only, video-only, and paired audio-visual denoising trajectories. The training schedule consists of three stages. First, we train on audio-only and paired audio-visual data with a 3:1 sampling ratio to initialize the audio pathway and stabilize audio denoising while preserving the visual capability inherited from the pretrained video backbone [8]. We then shift the audio-only/audio-visual ratio to 1:2 and train on high-quality audio data together with the full audio-visual dataset to improve audio fidelity and audio-visual synchronization. Finally, we fine-tune on curated high-quality audio-visual data to improve instruction following and controllable generation, including multi-speaker dialogue, complex motion, and camera control.

2.4.2 Structured Dropout for Guidance

To support condition-factorized guidance, we construct paired conditional and partially unconditional denoising paths during training, enabling guidance signals to be estimated from controlled prediction differences. For audio-visual alignment, we apply *Random Cross-modality Attention Masking*, where cross-modal attention entries between audio and video tokens are randomly masked while intra-modal attention remains intact. This exposes the model to both coupled and partially decoupled audio-video denoising regimes, whose prediction contrast is later used for alignment guidance. For timbre control, we apply *Random Timbre-in-Context Conditioning* by dropping or replacing timbre tokens with null tokens for a subset of speech spans. This trains the model under timbre-conditioned and timbre-free contexts, providing the prediction contrast required for timbre guidance.

2.4.3 Condition-Factorized Classifier-Free Guidance

During inference, we build on the audio-visual guidance formulation of LTX [5] and extend it with reference-timbre guidance. Let $\mathbf{v}_\theta^{c,a,\tau}(z_t)$ denote the prediction at step t, where z_t is the noisy audio-video latent, and c, a, and τ denote textual context, audio-video interaction, and reference timbre conditioning, respectively. We define three guidance directions:

$$\begin{aligned}
\Delta_{\text{text}} &= \mathbf{v}_\theta^{c,a,\tau}(z_t) - \mathbf{v}_\theta^{\varnothing,a,\tau}(z_t), \\
\Delta_{\text{align}} &= \mathbf{v}_\theta^{c,a,\tau}(z_t) - \mathbf{v}_\theta^{c,\varnothing,\tau}(z_t), \\
\Delta_{\text{timbre}} &= \mathbf{v}_\theta^{c,a,\tau}(z_t) - \mathbf{v}_\theta^{c,a,\varnothing}(z_t).
\end{aligned} \tag{10}$$

The final guided prediction is

$$\hat{\mathbf{v}}_\theta(z_t) = \mathbf{v}_\theta^{c,a,\tau}(z_t) + s_{\text{text}}\Delta_{\text{text}} + s_{\text{align}}\Delta_{\text{align}} + s_{\text{timbre}}\Delta_{\text{timbre}}, \tag{11}$$

where s_{text}, s_{align}, and s_{timbre} control prompt adherence, audio-visual synchronization, and timbre preservation, respectively. This factorized formulation supports decoupled alignment guidance and fine-grained timbre control during inference.

3 Experiments

3.1 Experimental Setup

Implementation details. NAVA has 6.3B parameters with 30 MMDiT blocks, where the first 10 blocks are *Hierarchical Alignment Layers* and the remaining 20 are *Unified Fusion Layers*. We

Table 1: **General capabilities on Verse-Bench.** We compare representative audio-video generation models in terms of synchronization, video quality, and audio quality. NAVA achieves the strongest overall synchronization and video quality, with competitive audio quality and the fewest parameters.

Model	Params	Resolution	AV-Align			Video Quality ↑	Audio		
			Sync-C ↑	Sync-D ↓	IB ↑		WER ↓	PQ ↑	FD ↓
Ovi 1.1 [4]	10B	720p	7.484	7.979	0.199	0.636	0.102	5.843	0.942
MOVA [6]	A18B (32B)	720p	7.289	7.808	0.269	0.603	0.126	**7.233**	0.922
Davinci [7]	15B	540p	7.149	7.816	0.269	0.600	0.151	5.956	0.931
LTX 2.3 [5]	19B	512p	7.248	7.690	**0.337**	0.576	0.106	6.946	**0.829**
NAVA (ours)	**6.3B**	**720p**	**7.791**	**7.566**	0.313	**0.659**	**0.099**	6.861	0.833

Table 2: **Reference-timbre generation on Seed-TTS.** WER and speaker similarity are evaluated under an audio-only protocol for both speech and audio-video models.

Model Category	Model	WER ↓	Speaker Similarity ↑
Audio	CosyVoice [15]	4.29	60.9
	CosyVoice2 [16]	2.57	65.2
	Qwen2.5-Omni [17]	2.72	63.2
Audio-Video	DreamID-Omni [11]	31.76	35.7
	NAVA	**4.20**	**66.7**

initialize corresponding layers from Wan2.2-5B [8], use Wan2.2-VAE for video latents with a $4 \times 16 \times 16$ compression ratio, and use LTX2.3-VAE for multi-channel audio latents. The model is trained with AdamW at a learning rate of 5×10^{-5} on 128 NVIDIA H100 GPUs, with an effective batch size of 512 for 70K steps following the three-stage schedule in Sec. 2.4. We apply random cross-modality attention masking and timbre-condition dropout with probabilities of 20% each, and sample image conditions with probability 50%.

Benchmarks and baselines. Following MoVA [6] and daVinci-MagiHuman [7], we adopt Verse-Bench [9] for objective audio-visual evaluation, covering speech videos, sound effects, and musical instruments. We further evaluate timbre controllability on the Seed-TTS benchmark [10]. For Verse-Bench, we compare with Ovi-1.1 [4], MoVA [6], LTX-2.3 [5], and daVinci-MagiHuman [7], covering dual-tower and tri-modal unified paradigms. For Seed-TTS, we compare with DreamID-Omni [11]. Since DreamID-Omni requires paired reference audio and image inputs, we use a fixed reference image for all samples and provide the corresponding reference audio. For fair comparison, we evaluate the base version of each model without additional super-resolution, distillation, or post-processing modules. We also apply Gemini-3-Flash rewriting to all test prompts to match each model's expected inference format while preserving the original benchmark semantics.

Evaluation metrics. We evaluate the proposed method along four dimensions: audio–visual alignment, video quality, audio quality, and timbre controllability, covering both perceptual fidelity and cross-modal consistency. For audio–visual alignment, we report Sync-C and Sync-D from SyncNet [12], which measure the confidence and temporal offset of lip–audio synchronization, respectively. We further use the ImageBind score (IB-Score) [13] to assess cross-modal semantic consistency between the generated video and audio. For video quality, we report identity consistency and aesthetic score. For audio quality, we employ Audiobox-Aesthetics [14], a no-reference audio assessment model trained to predict human perceptual judgments along multiple aesthetic axes. Specifically, we report Production Quality (PQ) to assess perceived audio fidelity, and Fréchet Distance (FD) to measure the distributional gap between generated and reference audio in the learned audio feature space. In addition, we report word error rate (WER) to measure speech intelligibility and content accuracy. For timbre controllability, we compute Seed-TTS timbre similarity between the generated speech and the reference utterance. Higher values are better for Sync-C, IB-Score, video quality, PQ, and timbre similarity, whereas lower values are better for Sync-D, FD, and WER.

Table 3: **Ablation of Align-then-Fuse MMDiT.** Comparison of *Hierarchical Alignment Layers* (HAL), *Unified Fusion Layers* (UFL), and their combination.

HAL Layers	UFL Layers	Model Params	Alignment		Video Quality ↑	Audio Quality	
			Sync-C ↑	IB ↑		PQ ↑	WER ↓
	○	5B	7.643	33.22	67.53	5.296	0.182
○	○	6.3B	**7.684**	**34.34**	**67.67**	**5.377**	0.177
○		7.7B	7.030	30.91	66.62	5.347	**0.167**

Table 4: **General-audio evaluation on Verse-Bench Subsets 1 and 2.** Higher CLAP and PQ and lower WER and FAD-VGGish are better.

Method	WER ↓	PQ ↑	CLAP ↑	FAD-VGGish ↓
Ovi-1.1	0.102	5.843	0.202	7.902
MoVA	0.126	**7.233**	0.219	7.113
daVinci-MagiHuman	0.151	5.956	0.207	7.529
LTX-2.3	0.106	6.946	**0.223**	**5.887**
NAVA	**0.099**	6.861	0.221	6.241

3.2 Main Results

3.2.1 Quantitative Evaluation

Table 1 reports quantitative results on Verse-Bench. NAVA achieves the best overall trade-off across audio–visual alignment, video quality, audio quality, and model efficiency. With only 6.3B parameters, NAVA obtains the highest Sync-C score of 7.791 and the lowest Sync-D score of 7.566, demonstrating superior temporal synchronization between generated speech and visual motion. It also achieves the best video quality score of 0.659, suggesting that the proposed Align-then-Fuse design preserves strong visual generation capability while enabling synchronized audio generation.

For semantic audio–visual consistency, NAVA obtains an IB-Score of 0.313, outperforming Ovi-1.1 and remaining competitive with MoVA and Davinci, although LTX 2.3 achieves the highest IB-Score. For audio quality, NAVA achieves the lowest WER of 0.099, indicating improved speech intelligibility and content accuracy. Its PQ and FD scores, 6.861 and 0.833, are also competitive among baselines, showing that NAVA maintains high perceived audio fidelity and a close distributional match to reference audio. These results indicate that NAVA substantially improves audio–visual synchronization and video quality without sacrificing audio quality, despite using the fewest parameters among the compared audio-video models.

Table 2 evaluates reference-timbre speech generation on the EN subset of the Seed-TTS benchmark. Audio-only speech models such as CosyVoice [15], CosyVoice2 [16], and Qwen2.5-Omni [17] provide strong references for pure speech generation. Despite operating as an audio-video generation model with synchronized visual generation, NAVA achieves the highest speaker similarity of 66.7 and a competitive WER of 4.20. Within the audio-video model category, NAVA substantially outperforms DreamID-Omni, reducing WER from 31.76 to 4.20 and improving speaker similarity from 35.7 to 66.7. These results demonstrate the effectiveness of *Timbre-in-Context Conditioning*, which binds reference timbre cues to corresponding speech spans through the context pathway. Overall, NAVA provides a strong balance across synchronization, semantic consistency, video quality, audio quality, and timbre controllability.

General-audio quality and semantics. To complement speech-oriented WER, we evaluate general audio on Verse-Bench Subsets 1 and 2, which include speech, environmental sounds, musical instruments, transportation, and everyday sound events. We report CLAP [18] for audio–text semantic correspondence and FAD-VGGish [19] for distribution-level audio realism in the VGGish embedding space, alongside PQ and WER. CLAP applies to both speech and non-speech audio, whereas WER measures transcription accuracy; the two metrics therefore capture complementary aspects of generated audio. FAD-VGGish and the FD values in Table 1 use different feature representations and should not be compared numerically. Table 4 shows that NAVA achieves the lowest WER and remains competitive in perceptual quality, semantic correspondence, and distribution-level realism,

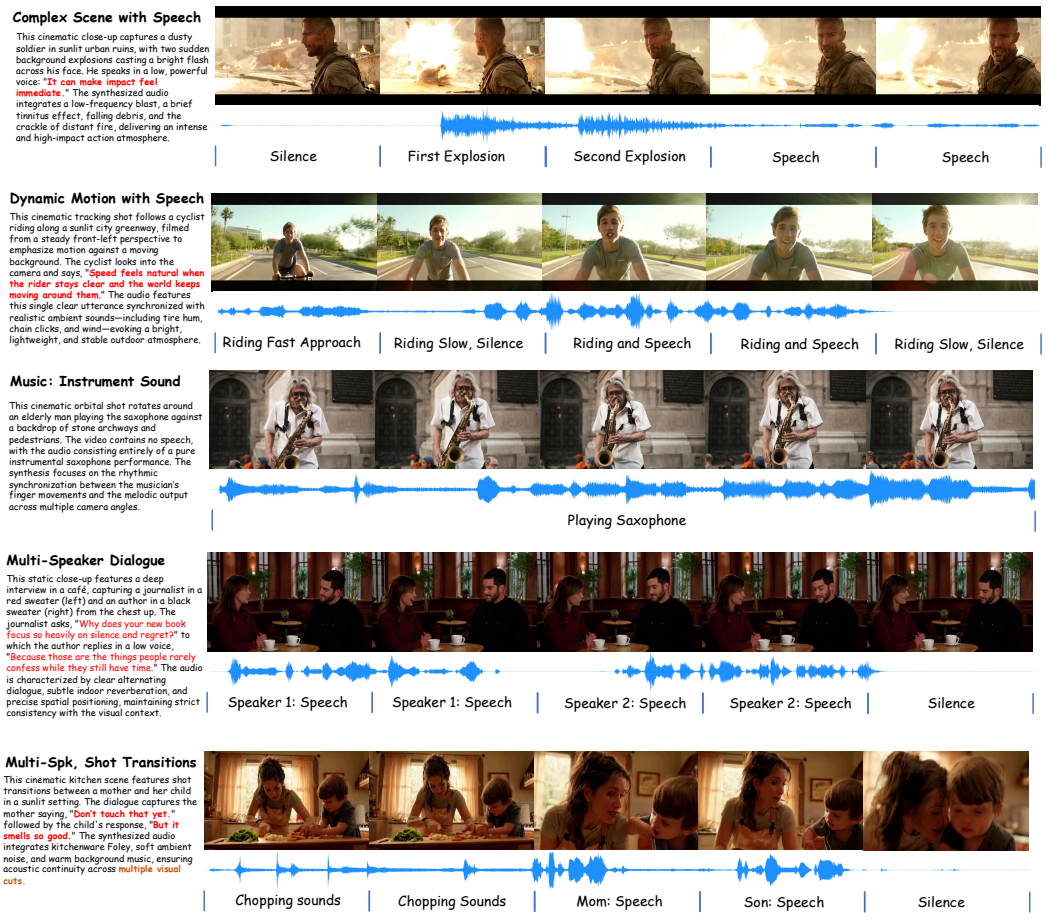

Figure 3: **Qualitative visualization of NAVA.** Generated frames, audio waveforms, and event annotations illustrate audio-visual synchronization across complex speech scenes, dynamic motion, musical performance, multi-speaker dialogue, and shot transitions.

without being best on every general-audio metric. These metrics broaden the evaluation beyond speech, but do not establish frame-level synchronization of arbitrary non-speech events: SyncNet targets lip–speech alignment, and ImageBind measures scene-level correspondence rather than precise event timing.

Comparison with additional audio-video generators. NAVA outperforms UniAVGen [20] and JAVISDiT++ [21] across the reported Verse-Bench metrics. The detailed comparison and metric definitions are provided in Section 5.1, Table 7.

3.2.2 Qualitative Evaluation

Visualization. Fig. 3 visualizes representative NAVA generations across challenging scenarios, including speech in complex acoustic scenes, speech during dynamic motion, musical performance, multi-speaker dialogue, and shot transitions. The sampled frames, waveforms, and event annotations show that NAVA can synthesize temporally synchronized speech, sound effects, and instrumental audio under complex visual contexts. The examples also demonstrate controllable speaker assignment and coherent generation across multi-speaker and multi-shot settings.

User Study. We conduct a GSB human evaluation on 250 cases spanning text-to-audio-video (T2AV) and text-image-to-audio-video (TI2AV) generation. T2AV uses synthetic prompts covering single- and dual-speaker speech, camera control, ambient sounds, instruments, and complex acoustic events; TI2AV uses Verse-Bench samples. MoVA is excluded from T2AV because its re-

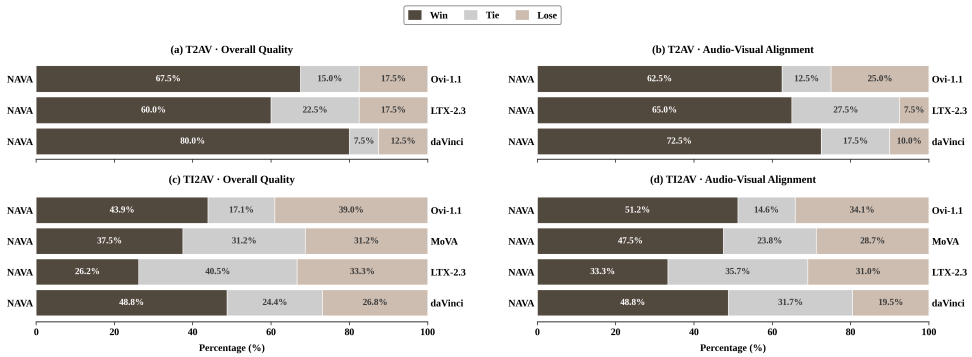

Figure 4: **Human evaluation on T2AV and TI2AV.** Bars report NAVA's win/tie/lose percentages against each baseline for overall quality and audio-visual alignment.

leased model does not support text-only generation. Participants compare paired outputs for overall audio-visual quality and alignment accuracy, assigning Win, Tie, or Lose.

Figure 4 shows that NAVA is preferred over all compared baselines on T2AV, with quality win rates of 60.0–80.0% and alignment win rates of 62.5–72.5%. On TI2AV, NAVA is favored over Ovi-1.1, MoVA, and daVinci in both dimensions. LTX-2.3 retains an advantage in overall quality, while NAVA remains competitive in alignment. These results highlight favorable perceptual quality and audio-visual synchronization across both generation settings.

Joint multi-speaker audio-video evaluation. A joint multi-speaker human evaluation against DreamID-Omni shows favorable preferences for NAVA in speaker-to-character binding, timbre similarity, lip-sync, and visual identity consistency. The four-dimensional results and discussion are provided in Section 5.1, Table 8.

3.3 Ablation Studies

Controlled comparison of interaction architectures. To isolate architectural effects from model scale and training recipe, we train three variants with the same Wan2.2-5B initialization, training data, token budget, optimization settings, and inference protocol, with parameter counts closely matched at approximately 6.3B. The dual-stream variant maintains separate audio and video pathways connected by dedicated cross-modal modules. The tri-modal unified variant places audio, video, and text tokens in a shared self-attention space from the beginning. NAVA instead jointly processes the audio and video denoising targets while introducing text and timbre as external context. Table 5 shows that NAVA obtains the strongest synchronization, audio–visual semantic correspondence, and video quality, with the lowest WER and perceptual audio quality comparable to the dual-stream variant. This supports separating interaction between the two denoising targets from external semantic conditioning, rather than attributing the gains to additional model parameters. The corresponding computational costs are reported in Appendix 5.5: NAVA reduces forward FLOPs and GPU-hours by approximately 7.2% relative to dual-stream and 12.1% relative to tri-modal unified processing.

Table 5: **Controlled architecture comparison at approximately 6.3B parameters.** Initialization, data, and training schedule are matched; scores are specific to this setting.

Architecture	Params	Sync-C ↑	IB ↑	Video Quality ↑	PQ ↑	WER ↓
Dual-stream	6.30B	7.183	0.324	66.43	**5.381**	0.180
Tri-modal unified	6.29B	4.217	0.285	67.10	5.281	0.251
NAVA	6.29B	**7.684**	**0.343**	**67.67**	5.377	**0.177**

Ablation on Align-then-Fuse MMDiT. Table 3 studies the roles of *Hierarchical Alignment Layers* (HAL) and *Unified Fusion Layers* (UFL). The UFL-only variant removes early modality-aware

Table 6: **Ablation of condition-factorized CFG.** Alignment CFG is evaluated on Verse-Bench, while timbre CFG is evaluated on the Seed-TTS benchmark.

	Sync-C ↑	Sync-D ↓	IB ↑	Video Quality ↑	PQ ↑	WER_{VB} ↓	WER_{Seed} ↓	ASV ↑
Alignment CFG on Verse-Bench								
NAVA w/o Align CFG	6.170	8.755	0.355	**0.667**	6.658	0.126	–	–
NAVA w/ Align CFG	**7.791**	**7.566**	**0.402**	0.659	**6.860**	**0.099**	–	–
Timbre CFG on Seed-TTS								
NAVA w/o Timbre CFG	–	–	–	–	–	–	**3.78**	65.5
NAVA w/ Timbre CFG	–	–	–	–	–	–	4.20	**66.7**

alignment and directly shares generation parameters across audio and video, leading to weaker Sync-C and IB scores. This suggests that fully shared generation without an explicit alignment stage is insufficient for establishing fine-grained audio-video correspondence. The HAL-only variant achieves stronger audio-related metrics, including improved speech intelligibility and audio fidelity, but degrades IB and video quality, indicating that persistent modality-aware alignment can preserve unimodal denoising ability at the cost of compact high-level fusion. Combining HAL and UFL achieves the best overall trade-off, obtaining the strongest synchronization, cross-modal consistency, and video quality. These results support the proposed align-then-fuse design: HAL first aligns heterogeneous audio-video representations, while UFL promotes shared high-level generation and collaborative denoising after correspondence has been established.

Sensitivity to HAL/UFL allocation. Under a matched reduced training schedule, 10 HAL + 20 UFL performs best among the tested 5/25, 10/20, and 20/10 allocations. The full comparison and analysis are provided in Section 5.1, Table 9.

Ablation on Condition-Factorized CFG. Table 6 evaluates the proposed condition-factorized classifier-free guidance. Alignment CFG substantially improves audio-visual correspondence, increasing Sync-C from 6.170 to 7.791, reducing Sync-D from 8.755 to 7.566, and improving IB from 0.355 to 0.402. These gains come with only minor changes in video quality, while WER decreases from 0.126 to 0.099 and PQ improves from 6.658 to 6.860, showing that alignment guidance strengthens synchronization without compromising unimodal generation quality. For timbre control, Timbre CFG improves reference-timbre consistency, increasing ASV from 65.5 to 66.7, with a mild WER trade-off from 3.78 to 4.20. Overall, these ablations show that different guidance directions modulate distinct generation attributes: Alignment CFG improves audio-video correspondence, while Timbre CFG improves timbre consistency. This validates factorized guidance for separately adjusting synchronization and timbre controllability at inference time, without retraining or modifying the generation backbone.

4 Conclusion

We presented **NAVA**, a Native Audio-Visual Alignment framework for joint audio-video generation. NAVA decouples audio-visual synchronization from context conditioning by establishing audio-video correspondence in a dedicated alignment space and using context as external guidance. We instantiate this formulation with an *Align-then-Fuse MMDiT* architecture, which bridges modality-aware alignment and unified audio-video denoising, and introduce *Timbre-in-Context Conditioning* for segment-level reference-timbre control. Experiments on Verse-Bench and Seed-TTS demonstrate that NAVA achieves strong audio-visual synchronization, visual quality, semantic consistency, and timbre controllability. These results indicate that native audio-visual alignment with decoupled context conditioning is a promising direction for scalable and controllable audio-video generation.

NAVA remains limited on long-tail and highly compositional audio events, including rare animal sounds, music, singing, and complex mixtures of scene sounds. Addressing these limitations requires broader, carefully curated audio-visual data for rare events and compositionally rich scenarios. Future work includes deeper audio-visual coupling and earlier fusion through joint audio-visual tokenizers or unified representations to improve synchronization, semantic consistency, and generalization.

References

[1] Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

[2] Kuaishou Technology. Kling 3.0. `https://kling.ai`, 2026.

[3] Google DeepMind. Veo 3.1, 2025. URL `https://deepmind.google/models/veo`.

[4] Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation, 2025. URL `https://arxiv.org/abs/2510.01284`.

[5] Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

[6] OpenMOSS Team, Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, et al. Mova: Towards scalable and synchronized video-audio generation. *arXiv preprint arXiv:2602.08794*, 2026.

[7] Ethan Chern, Hansi Teng, Hanwen Sun, Hao Wang, Hong Pan, Hongyu Jia, Jiadi Su, Jin Li, Junjie Yu, Lijie Liu, et al. Speed by simplicity: A single-stream architecture for fast audio-video generative foundation model. *arXiv preprint arXiv:2603.21986*, 2026.

[8] 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.

[9] Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang Yu. Universe-1: Unified audio-video generation via stitching of experts. *arXiv preprint arXiv:2509.06155*, 2025.

[10] Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

[11] Xu Guo, Fulong Ye, Qichao Sun, Liyang Chen, Bingchuan Li, Pengze Zhang, Jiawei Liu, Songtao Zhao, Qian He, and Xiangwang Hou. Dreamid-omni: Unified framework for controllable human-centric audio-video generation. *arXiv preprint arXiv:2602.12160*, 2026.

[12] 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.

[13] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023.

[14] Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

[15] Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024.

[16] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

[17] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

[18] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2023.

[19] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2019.

[20] Guozhen Zhang, Zixiang Zhou, Teng Hu, Ziqiao Peng, Youliang Zhang, Yi Chen, Yuan Zhou, Qinglin Lu, and Limin Wang. Uniavgen: Unified audio and video generation with asymmetric cross-modal interactions. *arXiv preprint arXiv:2511.03334*, 2025.

[21] Kai Liu, Yanhao Zheng, Kai Wang, Shengqiong Wu, Rongjunchen Zhang, Jiebo Luo, Dimitrios Hatzinakos, Ziwei Liu, Hao Fei, and Tat-Seng Chua. Javisdit++: Unified modeling and optimization for joint audio-video generation. *arXiv preprint arXiv:2602.19163*, 2026.

[22] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in neural information processing systems*, 34:24206–24221, 2021.

[23] Yuchen Hu, Yu Gu, Chenxing Li, Rilin Chen, and Dong Yu. Video-to-audio generation with fine-grained temporal semantics. *arXiv preprint arXiv:2409.14709*, 2024.

[24] Ilpo Viertola, Vladimir Iashin, and Esa Rahtu. Temporally aligned audio for video with autoregression. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[25] Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation network with rectified flow matching. *Advances in neural information processing systems*, 37:128118–128138, 2024.

[26] Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28901–28911, 2025.

[27] Jun Wang, Xijuan Zeng, Chunyu Qiang, Ruilong Chen, Shiyao Wang, Le Wang, Wangjing Zhou, Pengfei Cai, Jiahui Zhao, Nan Li, et al. Kling-foley: Multimodal diffusion transformer for high-quality video-to-audio generation. *arXiv preprint arXiv:2506.19774*, 2025.

[28] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020.

[29] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354, 2024.

[30] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10219–10228, 2023.

[31] Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Fei, et al. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. *arXiv preprint arXiv:2503.23377*, 2025.

[32] Jun Wang, Chunyu Qiang, Yuxin Guo, Yiran Wang, Xijuan Zeng, Chen Zhang, and Pengfei Wan. Klear: Unified multi-task audio-video joint generation. *arXiv preprint arXiv:2601.04151*, 2026.

5 Appendix

5.1 More Supplementary Experiments

We provide additional baseline comparisons, joint multi-speaker audio-video evaluation, and a sensitivity study of the HAL/UFL allocation to complement the experiments in the main paper.

5.1.1 Comparison with Additional Audio-Video Generators

We further evaluate UniAVGen [20] and JAVISDiT++ [21] under the same Verse-Bench protocol as NAVA. Table 7 reports visual aesthetic score (AES), audio–visual semantic correspondence (IB), perceptual audio quality (PQ), CLAP-based Fréchet distance (CLAP-FD), synchronization confidence, and speech intelligibility. AES is an individual visual-quality measure rather than the composite Video Quality score in Table 1; CLAP-FD is also distinct from FAD-VGGish. NAVA outperforms

Table 7: **Additional contemporary baselines on Verse-Bench.** Parameter counts are rounded to one decimal place.

Model	Params	AES ↑	IB ↑	PQ ↑	CLAP-FD ↓	Sync-C ↑	WER ↓
JAVISDiT++	2.1B	43.79	0.209	5.598	0.943	0.953	1.061
UniAVGen	7.1B	40.75	0.237	6.543	0.872	4.935	0.131
NAVA	6.3B	**45.59**	**0.313**	**6.861**	**0.832**	**7.791**	**0.099**

both baselines across the reported metrics and uses fewer parameters than UniAVGen. These comparisons extend baseline coverage, but differences in initialization, training data, and model scale prevent causal attribution to architecture alone; the controlled comparison in Table 5 addresses this confound.

5.1.2 Joint Multi-Speaker Audio-Video Evaluation

We additionally conduct a pairwise human evaluation of multi-speaker audio-video generation against DreamID-Omni [11]. Participants judge whether each voice is bound to the correct visual character, whether the generated voice matches the assigned reference timbre, whether speech is synchronized with mouth motion, and whether each speaker's visual identity remains consistent. Table 8

Table 8: **Joint multi-speaker audio-video preferences against DreamID-Omni.** Good and Bad denote preference for NAVA and DreamID-Omni, respectively. GSB is the Good fraction minus the Bad fraction.

Dimension	Good (%)	Same (%)	Bad (%)	GSB
Speaker-to-character binding	34.0	51.5	14.5	+0.195
Timbre similarity	69.0	26.0	5.0	+0.640
Lip-sync	24.0	63.0	13.0	+0.110
Visual identity consistency	64.5	32.0	3.5	+0.610

shows that NAVA receives more favorable judgments than DreamID-Omni in all four dimensions, with the largest margins in reference-timbre similarity and visual identity consistency. The higher tie rates for speaker-to-character binding and lip-sync indicate that the methods are frequently judged comparable on those dimensions. These results complement the audio-focused Seed-TTS evaluation with joint audio-video evidence for speaker-specific voice control, character correspondence, and synchronization. They do not independently isolate span-level conditioning: a direct comparison with global timbre injection or removal of speech-span boundaries remains necessary to attribute the gains specifically to that mechanism.

5.1.3 Sensitivity to HAL/UFL Allocation

We further compare 5 HAL + 25 UFL, 10 HAL + 20 UFL, and 20 HAL + 10 UFL under a common reduced training schedule. All three variants use the same initialization, model scale, training data,

optimization settings, number of training steps, and inference protocol; their results are compared only within this matched setting. The main-model evaluation and endpoint ablations are reported separately. Table 9 favors 10 HAL + 20 UFL across all reported metrics. Too few modality-aware

Table 9: **Sensitivity to HAL/UFL allocation under a matched reduced training schedule.** The 10/20 split performs best among the three tested intermediate allocations.

HAL/UFL	Sync-C ↑	IB ↑	Video Quality ↑	PQ ↑	WER ↓
5/25	3.396	0.224	67.55	5.449	0.241
10/20	**5.418**	**0.280**	**67.92**	**5.531**	**0.178**
20/10	4.907	0.248	67.12	5.381	0.203

layers may provide insufficient depth to align heterogeneous representations, while too many reduce the depth available for shared fusion. HAL retains modality-specific attention projections early to accommodate heterogeneous audio and video latents, while UFL shares those projections after correspondence has been established; this staged sharing balances modality awareness and common generation capacity. The results support the selected split among the tested allocations, without establishing that it is globally optimal.

5.2 Data Pipeline

Large-scale collection and preprocessing. We construct a large-scale audio-visual training corpus from heterogeneous sources, including Koala-36M, TED-style speech videos, and raw movie/TV footage. The raw videos are first segmented at scale with a Hadoop-based pipeline. To improve data quality and reduce shortcut learning from overlaid text, we apply OCR-based filtering and subtitle removal using PaddleOCR. We further remove redundant or near-duplicate clips by extracting video embeddings with VideoCLIP and performing large-scale k-means clustering, followed by category-level merging and filtering of small clusters.

Modality-aware tagging and subset construction. We then annotate each clip with both visual and acoustic metadata. For visual content, we use VLM-based filtering and tagging to retain clips with clear visual quality and coherent events or transitions, and assign semantic tags such as movies, documentaries, TV series, live streams, speeches, news, and interviews. For audio content, we combine YAMNet-based audio classification with an omni-modal tagger to categorize clips into single-speaker speech, multi-speaker speech, ambient sound, music, and singing. These tags are used to construct different data subsets for pretraining and supervised fine-tuning.

Hierarchical audio-visual caption annotation. For caption annotation, we adopt a two-stage strategy. On the full-scale dataset, video and audio captions are generated separately using Qwen3-VL and Qwen3-Omni, and are then fused by either direct concatenation or rewriting by Gemini-3-Flash. For high-quality and multi-speaker subsets, we use Gemini-3-Pro to produce more accurate, structured, and temporally grounded audio-visual captions.

Multi-operator quality filtering. Finally, we apply a set of quality assessment operators to filter low-quality samples. These operators cover visual quality, including aesthetics, sharpness, brightness, and motion score; audio quality, including AudioBox Aesthetic scores; and audio-visual alignment, measured by SyncNet, SyncFormer, and ImageBind. The resulting corpus consists of diverse, deduplicated, high-quality, and richly annotated audio-visual clips, supporting scalable multi-stage audio-visual model training.

5.3 Data Statistics

Our raw collection contains approximately 20M audio clips and 100M video clips. After subtitle filtering, quality filtering, near-duplicate removal, and audio-visual alignment filtering, we obtain around 15M clips for large-scale training. Koala-36M contributes approximately 20% of the final training corpus. The average video duration is about 7 seconds. For the supervised fine-tuning stage, we further apply multi-operator collaborative filtering and retain 160K high-quality samples with accurate captions and strong audio-visual alignment.

5.4 Prompt Engineering

Audio-visual generation is conditioned on multiple intertwined factors, including scene layout, subject appearance, motion, camera behavior, lighting, style, speech, speaker characteristics, environmental sounds, music, and spatial acoustics. Unlike purely visual generation, audio-visual generation requires the prompt to specify not only what appears in the video, but also what is heard and how the audible events are temporally aligned with the visual dynamics. Therefore, we use structured dense captions rather than free-form short descriptions for both training and inference.

We design a unified prompt rewriting template that decomposes each video into global visual semantics, temporal dynamics, camera and composition, and audio events. The audio branch is compatible with non-speech scene sounds, single-speaker speech, multi-speaker dialogue, music, singing, and ambient audio. For speech videos, utterances are explicitly marked with `<S>` and `<E>` tokens, while speaker timbre, emotion, speaking rate, and sound-field position are also described. For non-speech videos, the template emphasizes action sounds, contact and friction sounds, object sounds, environmental ambience, and reverberation. This design yields captions that are visually detailed, temporally ordered, acoustically grounded, and consistent across heterogeneous audio-visual data.

Audio-Visual Prompt Template. The full template used for prompt rewriting is shown below.

Unified Audio-Visual Caption Prompt Template

```
Describe the video in a structured audio-visual caption. The caption should
cover visual content, temporal dynamics, camera behavior, and all audible
events. Use natural language, but preserve the temporal order of the video.

[Global visual description]
This is a [style/genre] video with a [overall mood/atmosphere] feeling.
The main subject(s) are [subject(s)] located in [scene/environment].
Describe the subject appearance, posture, facial expression, clothing,
materials, objects, and spatial arrangement. Also describe the background
elements, lighting condition, dominant color tone, and environmental mood.

[Temporal dynamics]
At the beginning of the video, [initial visual state]. Then, [action or event
development 1]. Next, [action or event development 2]. As the video continues,
describe [motion details], [object interactions], and [environmental feedback].
At the end of the video, [ending state], producing [final visual/audio-visual
effect].

[Camera and composition]
Describe the camera angle, shot scale, framing, and composition. Specify
whether the camera is fixed, handheld, tracking, panning, tilting, zooming, or
pushing in. Mention whether there are cuts or scene transitions. The caption
should highlight how the camera captures [key action], [subject details], and
[environmental dynamics].

[Audio description]
Describe all audible content in temporal order. The audio may contain
non-speech sounds, single-speaker speech, multi-speaker dialogue, music,
singing, or ambient sound. Always describe the relationship between sound and
visual events when possible.

If there is no speech:
    State that there is no spoken dialogue or voice-over.
    Describe the main sound effects, including [action sounds],
    [contact/friction sounds], [object sounds], [environmental sounds],
    [music or singing if present], [spatial echo/reverberation], and
    [overall auditory impression].

If there is single-speaker speech:
    Describe the speaker's action, expression, position, voice timbre,
    emotion, speaking rate, and sound-field position.
    Transcribe the utterance using:
    <S>[spoken content]<E>
    Also describe the background ambience, object sounds, reverberation,
    and whether music is present.

If there is multi-speaker dialogue:
    For each speaker turn, describe the speaker's action, expression,
    voice timbre, emotion, speaking rate, and sound-field position.
    Transcribe each utterance using:
    [Speaker A description]: <S>[utterance A]<E>
    [Speaker B description]: <S>[utterance B]<E>
    Continue this format for additional speakers or turns.
    Also describe turn-taking, speaker interaction, background ambience,
    object sounds, reverberation, and mixing quality.
```

Example Captions. Representative rewritten captions are shown below. These examples demonstrate that the proposed template can consistently capture visual details, temporal evolution, camera behavior, dialogue structure, speaker-specific acoustic properties, and background sound events.

Table 10: **Computational costs of the controlled approximately 6.3B architecture variants.** FLOPs are reported in TFLOP and training time in GPU-hours.

Architecture	Params	Forward	Training step ($3\times$)	GPU-hours
Dual-stream	6.30B	175.8	527.4	33,093
Tri-modal unified	6.29B	185.7	557.2	34,962
NAVA	6.29B	**163.2**	**489.6**	**30,720**

Example: Compact Dialogue Caption

This realistic animated video shows a two-person conversation in a warm coffee shop. The young man on the left wears a dark blue hoodie and looks curious, while the older man on the right wears round glasses, has a white beard, holds a coffee cup with both hands, and appears calm and gentle. They sit face-to-face at a wooden table, both framed from the chest up. The background contains a coffee menu board, warm pendant lights, plants, and soft window light.
At the beginning of the video, the young man on the left asks clearly: `<S>Can I use any timbre I want?<E>` The older man on the right nods gently and responds in a low and warm voice: `<S>Yes, in all kinds of complex scenes.<E>` The video then ends.
The camera remains fixed in a medium-close shot, with no cuts, zooms, pans, or tilts. The audio contains soft cafe ambience, subtle cup contact sounds, and indoor reverberation. The two voices alternate clearly with accurate left-right sound-field positioning, highlighting speaker identity, timbre contrast, and a natural conversational rhythm.

Infrastructure and Training Cost. We train NAVA with a distributed infrastructure designed for large-scale, long-context audio-visual modeling. To reduce memory overhead, we adopt Fully Sharded Data Parallel (FSDP), which shards model parameters, gradients, and optimizer states across devices. This allows training with long multimodal sequences while maintaining a sufficiently large global batch size.

A key systems bottleneck is the online preparation of heterogeneous media samples. We therefore use an asynchronous server-based preprocessing pipeline for high-concurrency audio and video handling. Rather than performing media I/O, audio-video extraction, and VAE encoding inside the training workers, dedicated data servers process samples asynchronously and prefetch ready-to-use training examples. This reduces stalls from file access, media decoding, audio extraction, and feature preparation, improving GPU utilization during large-scale training.

To handle the heterogeneous sequence lengths of mixed-modality data, we use two batching strategies. First, samples with similar sequence lengths are grouped into buckets to reduce padding overhead. Second, each dataloader micro-batch contains samples from a single modality, which makes the per-step sequence length more predictable and reduces padding within each forward-backward pass. We then interleave modalities across consecutive micro-batches through gradient accumulation, so that each effective global batch retains diverse multimodal supervision. This improves efficiency of jointly training on audio, video and audio-video with substantially different token-length distributions.

Training is conducted in three stages. Stages 1 and 2 use 160 NVIDIA H100 GPUs, together with asynchronous data servers, for approximately three weeks, corresponding to about $160 \times 21 \times 24 = 80{,}640$ H100 GPU-hours. Stage 3 uses 160 NVIDIA H100 GPUs for one additional week, corresponding to about $160 \times 7 \times 24 = 26{,}880$ H100 GPU-hours. Overall, the full training pipeline requires approximately $107{,}520$ H100 GPU-hours, assuming continuous training.

5.5 Computational Cost of the Controlled Architecture Comparison

Table 10 complements the matched-capacity architecture comparison in Table 5. The three variants share initialization, data, token budget, and training schedule, but their attention organization leads to different computational costs. Training-step FLOPs are reported as approximately three times the single-forward cost, accounting for forward and backward computation. The GPU-hours below correspond only to these controlled runs, not to the full multi-stage training pipeline described above. NAVA lowers forward computation and GPU-hours by approximately 7.2% relative to the

dual-stream variant and 12.1% relative to the tri-modal unified variant, while improving synchronization, semantic correspondence, video quality, and WER. The comparison therefore supports a more favorable performance–compute trade-off without relying on a larger parameter count.

5.6 Related Work

Video-to-Audio Generation. Video-to-audio generation synthesizes acoustic content conditioned on a given video, and often serves as a cascaded component for audio-visual content creation. Early methods explore multimodal representation learning and cross-modal conditioning, using Transformer architectures or visual-textual encoders to fuse video and text cues [22; 23]. Recent systems improve temporal precision and generation efficiency through high-frame-rate visual features, rectified flow matching, and large-scale audio-visual training [24; 25]. More recent works such as MMAudio [26] and Kling-Foley [27] adopt diffusion or MMDiT-style architectures and leverage large-scale video-audio corpora such as VGGSound [28] and WavCaps [29]. Although these approaches can generate plausible audio for existing videos, they are inherently conditioned on a fixed visual trajectory and do not address native joint generation, where audio and video should co-evolve during synthesis.

Audio-Video Joint Generation. Unlike video-to-audio generation, audio-video joint generation synthesizes both modalities within a shared generation process and therefore requires tighter temporal and semantic coordination. Early attempts such as MM-Diffusion [30], Javis-DiT [31], and Universe-1 [9] explore cross-modal attention, expert composition, or multimodal diffusion for coordinated generation. Recent open-source systems, including UniAVGen [20], Ovi [4], LTX [5], and MoVA [6], mainly adopt dual-tower designs that maintain separate audio and video streams and introduce posterior fusion or alignment. Such designs can exploit pretrained unimodal priors, but delayed audio-video interaction may limit fine-grained synchronization and semantic consistency.

Unified modeling has been explored to strengthen cross-modal interaction. Apollo [32] applies joint attention over concatenated multimodal tokens, while daVinci-MagiHuman [7] places textual context, video, and audio tokens into a shared tri-modal space. These designs enable direct multimodal interaction, but fully mixing semantic context with generation modalities can entangle high-level conditioning with low-level audio-video synchronization. In contrast, NAVA establishes audio-video correspondence in a dedicated interaction space and injects context as external conditioning, separating native synchronization from semantic and controllable guidance.

Controllable Audio-Visual Generation. Controllable audio-visual generation requires not only synchronized audio and video, but also flexible conditioning on identity, reference audio, speaker style, or timbre. UniAVGen [20] and DreamID-Omni [11] incorporate reference tokens or identity/-timbre conditions to support controllable generation. However, many reference-conditioning mechanisms are applied as global controls or auxiliary branches, which can be less flexible for multi-speaker scenarios where different utterances require different timbres. NAVA instead represents reference timbre cues as context tokens tied to specific speech spans. This enables compositional content-timbre binding through the existing context-conditioning pathway, without introducing an additional speaker-control branch or modifying the denoising backbone.