OpenMOSS

OmniVAE: An Audio-Video VAE with Cross-Modal Alignment for Joint Generation

Jun Zhan[1,2,3,*,‡] **Chen Yang**[1,2,3,*] **Yitian Gong**[1,3,*] **Donghua Yu**[1,3]
Kuangwei Chen[1,3] **Wenbo Zhang**[1,3] **Kexin Huang**[1,3] **Qi Luo**[1,3] **Zhe Xu**[1,3]
Ying Zhu[1,3] **Jin Wang**[1,3] **Tengyue Zhang**[2,4] **Qi Chen**[2,4]
Zhiyu Zhang[2,3] **Cheng Chang**[1,3] **Songlin Wang**[3] **Junqi Dai**[1] **Jiasheng Ye**[1]
Xiaogui Yang[3] **Tianyi Liang**[2,3] **Xiangyu Peng**[3] **Zhaoye Fei**[1,2,3]
Shimin Li[1,3] **Qinyuan Cheng**[1,3] **Xie Chen**[2,4] **Xinchi Chen**[1] **Xipeng Qiu**[1,2,3,†]

[1]Fudan University [2]Shanghai Innovation Institute
[3]MOSI Intelligence [4]Shanghai Jiao Tong University
[*]Equal Contribution [‡]Project Lead [†]Corresponding Author

Abstract

Recent generative models are moving beyond silent video or standalone audio synthesis toward the joint generation of synchronized audio and video. Despite this progress, jointly generating audio and video with fine-grained cross-modal correspondence remains challenging due to their fundamental structural differences. Most existing methods use audio and video VAEs trained separately. As a result, the two latent spaces lack cross-modal alignment, leaving the downstream generative model to learn cross-modal synchronization from scratch. We present **OmniVAE**, a jointly trained audio-video VAE that learns fine-grained semantic alignment between audio and video latent representations. Beyond reconstruction, OmniVAE uses a segment-level audio-video contrastive objective to capture temporal-semantic correspondence and align the two latent spaces. In parallel, it distills features from pretrained modality-specific semantic encoders into each modality, improving the downstream learnability of both latent spaces. Extensive experiments show that both objectives consistently improve the learnability of the latent spaces, translating into higher generation quality and more accurate cross-modal synchronization in downstream text-to-audio-video generation. These findings underscore the importance of learning unified representations as a foundation for omni-modal modeling.

Homepage: https://openmoss.ai/OmniVAE.github.io
Code: https://github.com/OpenMOSS/OmniVAE
Model: https://huggingface.co/OpenMOSS-Team/OmniVAE

1 Introduction

Unified audio-video generation aims to produce a video and its temporally synchronized, semantically coherent soundtrack jointly from a single text prompt, rather than generating the two modalities separately and

combining them afterward. However, existing systems typically rely on independently trained, modality-specific VAEs [1]. A tokenizer defines the representation space in which generation is learned, yet reconstruction alone does not ensure that this space is semantically structured or cross-modally aligned. The downstream generator must therefore learn both the data distribution and the correspondence between audio and video, making fine-grained synchronization particularly challenging.

Recent studies point to two complementary ways of improving this representation space. A growing body of work enriches modality-specific tokenizers with semantic structure: VA-VAE [2] and RAE [3] improve visual latent representations using pretrained vision encoders, while SpeechTokenizer [4] and MOSS-Audio-Tokenizer [5] pursue semantically meaningful audio representations. Meanwhile, video-to-audio methods such as MMAudio [6] and HunyuanVideo-Foley [7] demonstrate that explicit synchronization features [8] substantially improve temporal precision, highlighting the value of cross-modal alignment.

Motivated by these advances, we aim to make audio-video latent representations more learnable for downstream generative models by embedding modality-specific semantics and cross-modal correspondence directly into the VAE latent spaces. We therefore present **OmniVAE**, which learns aligned modality-specific latent spaces for audio and video while retaining the reconstruction capability of each modality. Its training is guided by two complementary objectives beyond reconstruction. A *modality-specific semantic distillation* objective injects semantic structure into each latent space using frozen pretrained encoders, while a *segment-level bidirectional InfoNCE* objective [9] aligns audio and video latents at fine temporal granularity. Both objectives are training-only and incur no additional inference cost. Experiments show that the resulting representations are more learnable both within individual modalities and jointly across audio and video, leading to higher generation quality and more accurate audio-video synchronization in downstream text-to-audio-video generation.

Our contributions are threefold. **(1)** We present **OmniVAE**, a jointly trained audio-video VAE that learns aligned modality-specific latent spaces with fine-grained semantic and temporal correspondence across modalities. To the best of our knowledge, OmniVAE is the first audio-video VAE designed explicitly for cross-modal latent alignment. **(2)** We introduce two complementary training objectives: modality-specific semantic distillation improves the learnability of each latent space, while segment-level audio-video contrastive learning aligns the two modalities at fine temporal granularity. Both objectives are training-only and incur no additional inference cost. **(3)** Through latent-space probing and downstream text-to-audio-video generation, we demonstrate that these objectives improve complementary aspects of the representation. Their combination consistently delivers the strongest overall generation quality and audio-video synchronization.

2 Related Work

2.1 Multimodal Unified Representation

We use *multimodal unified representation* to refer to a representation scheme in which different modalities are organized within a common semantic coordinate system, allowing semantically corresponding content to be directly related across modalities. Such unification does not require a shared encoder or identical latent distributions; it can be achieved through either a shared representation space or explicitly aligned modality-specific spaces.

Prior work explores such unification at different levels. Global contrastive methods organize heterogeneous modalities within a shared semantic space [11–13], while discrete representation learning seeks modality-agnostic codes that support cross-modal transfer [14–16]. These approaches capture high-level semantic correspondence but provide limited modeling of fine-grained temporal relationships. In contrast, SyncNet [17] and Synchformer [8] explicitly learn temporal correspondence between audio and video for synchronization.

However, these representations are primarily designed for discriminative tasks and are generally not optimized to reconstruct the original signals. They therefore cannot directly serve as latent spaces for generative modeling. OmniVAE bridges this gap by learning decodable audio and video representations that preserve modality-specific reconstruction while establishing fine-grained cross-modal alignment.

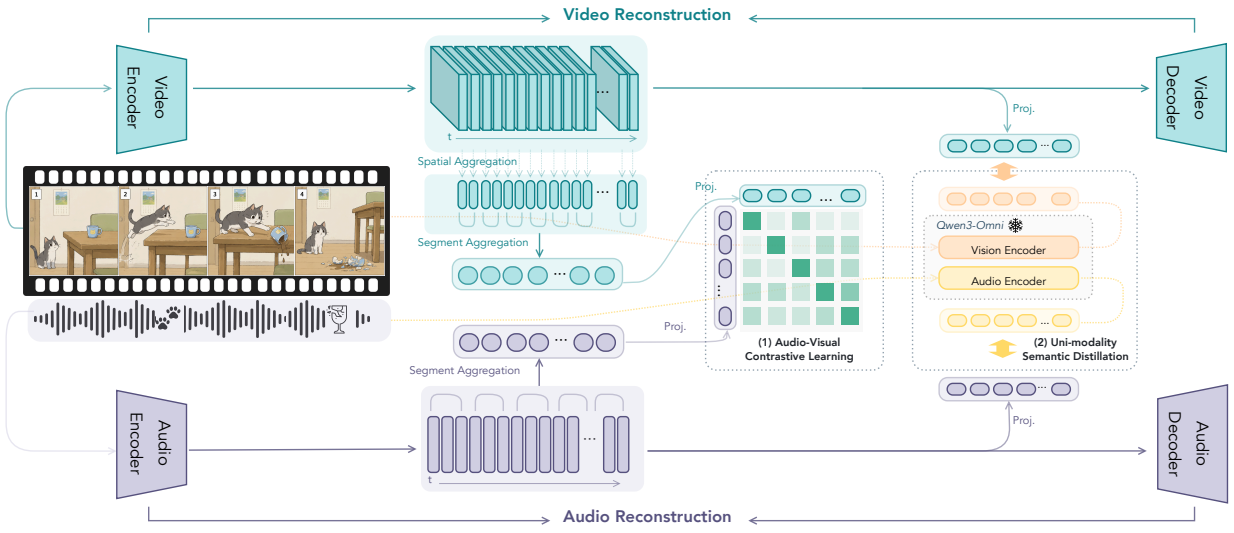

Figure 1 Overview of OmniVAE. OmniVAE keeps a separate video VAE (top) and audio VAE (bottom), each trained with its standard reconstruction objective (*Video / Audio Reconstruction*). Beyond reconstruction, we introduce two training-only objectives on top of the latents: **(1) Audio-Video Contrastive Learning** aligns the two modalities in a shared space at fine temporal granularity—a spatial aggregator collapses each video latent frame into a token sequence, a segment aggregator pools video and audio latents into matched per-segment features, and a bidirectional InfoNCE loss is applied on the resulting similarity matrix. **(2) Uni-modality Semantic Distillation** supervises each branch with features from the frozen Qwen3-Omni [10] vision and audio encoders, improving the learnability of each latent space on its own. The contrastive head and distillation projectors are used only during training; at inference, the two VAEs run independently with no added cost.

2.2 Audio and Video Tokenizers

Modern latent generative models rely on tokenizers that compress high-dimensional raw signals into compact latents on which generation is performed. On the video side, continuous VAEs under the latent-diffusion paradigm (e.g., Wan-VAE [18] and the Movie Gen VAE [19]) compress pixels into a continuous latent through spatio-temporal downsampling; on the audio side, neural codecs and continuous VAEs (e.g., DAC [20], the Stable Audio VAE [21], and Qwen-Audio-VAE [22]) play an analogous role. Qwen-Audio-VAE further scales this reconstruction-oriented paradigm to low-bitrate, high-throughput general-audio encoding. Such tokenizers are traditionally trained with reconstruction (plus adversarial) objectives alone, so their latent spaces lack explicit semantic structure. A recent line of work instead injects semantics *into the tokenizer itself*. On the visual side, recent work enriches generative tokenizers with semantic representations through feature alignment (VA-VAE [2] and REPA-E [23]), direct reuse of pretrained foundation encoders (RAE [3]), or joint reconstruction and semantic pretraining (UniTok [24] and VTP [25]). Collectively, these studies show that semantic structure, rather than reconstruction fidelity alone, is critical to the learnability of visual latent spaces. On the audio side, SpeechTokenizer [4] and MOSS-Audio-Tokenizer [5] similarly integrate semantic learning into audio tokenization, seeking representations that preserve acoustic fidelity while better supporting downstream generative modeling. Nevertheless, these semantic tokenizers remain modality-specific: they improve the representation of each modality independently but do not establish correspondence across modalities. To the best of our knowledge, OmniVAE is the first semantic tokenizer that both distills modality-specific semantics into audio and video latents and explicitly aligns the two latent spaces at the segment level.

2.3 Audio-Video Generation

Audio-video generation aims to jointly synthesize video and audio that are semantically consistent and temporally synchronized. MM-Diffusion [26] first proposed a coupled dual-U-Net diffusion framework

with random-shift attention to jointly denoise video frames and waveforms, and AV-DiT [27] extends this idea to Diffusion Transformer (DiT) backbones. Industrial systems such as Sora 2 [28], Veo 3 [29], and Seedance 2.0 [30] now natively generate video with synchronized audio at scale. On the open-source side, JavisDiT [31] introduces a hierarchical spatio-temporal synchronization prior inside a joint DiT, while Ovi [32], LTX-2 [33], and MOVA [34] all couple a video backbone and an audio backbone in twin-tower (dual-stream) designs through bidirectional cross-modal attention for synchronized audio-video generation. A closely related line injects synchronization features as *external conditioning* into the generator: MMAudio [6] feeds Synchformer-derived [8] frame-level visual features into a video-to-audio flow-matching transformer through AdaLN modulation and reports roughly a 50% relative improvement in synchronization, and HunyuanVideo-Foley [7] likewise relies on Synchformer features for gated temporal modulation—both underscoring the importance of explicit cross-modal alignment signals. In all of these systems, the video and audio latents are produced by independently pretrained VAEs, so cross-modal correspondence must either be learned from scratch by the generator or injected as external conditioning features. This motivates OmniVAE: we build cross-modal correspondence directly into the tokenizer by aligning the audio and video latent spaces, so that downstream joint generators operate on latents that already encode it.

3 OmniVAE

3.1 Overview

OmniVAE retains separate encoders and decoders for each modality. The video branch encodes $x^v \in \mathbb{R}^{(1+T)\times H\times W\times 3}$ into a latent $z^v \in \mathbb{R}^{C_v\times T_v\times (H/8)\times (W/8)}$ with $4\times$ temporal and $8\times$ spatial downsampling, where $T_v = 1 + T/4$ denotes the number of latent frames; the audio branch encodes a 48 kHz waveform $x^a \in \mathbb{R}^{1\times L}$ into a 1D latent $z^a \in \mathbb{R}^{C_a\times T_a}$, where $T_a = L/d_a$ for a temporal downsampling factor d_a. The two branches are architecturally independent and do not interact at inference, so they can be deployed jointly or used on their own. The video and audio branches instantiate the Wan2.2 VAE backbone [35] and a DAC-style continuous VAE [20].

On top of these two VAEs, we introduce two training-only objectives. During training we use paired samples (x^v, x^a) extracted from the same temporal window of a single source video, so that the two streams are naturally aligned in both semantics and time. A *segment-level audio-video contrastive* objective aligns z^v and z^a at fine temporal granularity in a shared embedding space, providing the downstream generator with a latent representation in which cross-modal correspondence is already established and reducing the burden of jointly modeling both modalities. In parallel, a *per-modality semantic distillation* objective supervises each branch with features from a frozen pretrained semantic encoder, making each branch's latent space easier to model on its own. Neither objective is active at inference time, so OmniVAE preserves the standard VAE interface and incurs no additional runtime cost.

3.2 Segment-Level Audio-Video Alignment

Our goal is to make fine-grained audio-video correspondence an intrinsic property of the VAE representations. Inspired by the segment-level contrastive formulation of Synchformer [8], we map the heterogeneous audio and video latents into temporally corresponding segment representations and contrast matched segments against negatives at multiple levels. The resulting contrastive objective encourages both VAE encoders to preserve fine-grained audio-video correspondence in their latent representations.

Segment Representation. Given a paired audio-video clip, two modality-specific aggregation paths map the heterogeneous latents to sequences of S temporally aligned features in a shared d-dimensional space. For video, the latent $z_i^v \in \mathbb{R}^{C_v\times T_v\times H_v\times W_v}$ is first spatially aggregated into a frame sequence in $\mathbb{R}^{(T_v-1)\times d_v}$ and then temporally pooled and projected into $u_i^v \in \mathbb{R}^{S\times d}$. We exclude the first latent frame because, in the causal video VAE, it covers a different temporal extent from the remaining frames. For audio, the latent $z_i^a \in \mathbb{R}^{C_a\times T_a}$ is transformed along the temporal dimension and pooled directly into $u_i^a \in \mathbb{R}^{S\times d}$. We implement the video spatial aggregator with a lightweight Transformer and the audio temporal aggregator with ConvNeXt-1D blocks. We write their segment features as $u_{i,s}^v, u_{i,s}^a \in \mathbb{R}^d$, where i indexes the clip and $s \in \{1, \dots, S\}$ its

temporal segment.

Hierarchical Negative Construction. Learning fine-grained audio-video correspondence requires a negative pool that combines hard temporal distinctions with broad semantic diversity. Our preprocessing pipeline divides each long source video into 8-second clips and further partitions each clip into $S = 48$ segments of $1/6$ second. Synchformer draws negatives only from the other segments within the anchor clip and from the segments of one additional clip [8]. In contrast, we retain the source-video identity of each clip and exploit the resulting source-video–clip–segment hierarchy together with the distributed batch to construct negatives at three levels. For an anchor segment (i, s), its paired segment in the other modality is the positive, while the negative set $\mathcal{N}_{i,s}$ contains: **(i) intra-clip negatives**, the other $S - 1$ segments in the same clip; **(ii) sibling-clip negatives**, segments from a temporally disjoint clip of the same source video; and **(iii) cross-video negatives**, segments sampled from unrelated source videos across the distributed batch. In the default setting, we use all $S - 1 = 47$ available intra-clip negatives and sample fixed numbers of 48 sibling-clip and 24 cross-video negatives, yielding an approximately 2:2:1 ratio across the three sources. The first two groups retain similar content and context while differing in temporal alignment, encouraging fine-grained temporal discrimination; the third expands semantic diversity. We use the same construction in both contrastive directions. The effects of clip duration and segment granularity are studied in Sec. 4.5.1.

Bidirectional Contrastive Objective. Let $\mathcal{P}_{i,s} = \{(i, s)\} \cup \mathcal{N}_{i,s}$ be the candidate set for anchor (i, s) and define $\mathrm{sim}_{i,s,j,t} = u^v_{i,s} \cdot u^a_{j,t}$. Because the features are normalized, this dot product is cosine similarity. For a batch of B paired clips and a learnable temperature τ, the video-to-audio InfoNCE loss [9] is

$$\mathcal{L}_{v\to a} = -\frac{1}{BS}\sum_{i=1}^{B}\sum_{s=1}^{S}\log\frac{\exp(\mathrm{sim}_{i,s,i,s}/\tau)}{\sum_{(j,t)\in\mathcal{P}_{i,s}}\exp(\mathrm{sim}_{i,s,j,t}/\tau)}. \tag{1}$$

$\mathcal{L}_{a\to v}$ is defined analogously by swapping the two modalities, and the final alignment loss is $\mathcal{L}_{\text{avclip}} = \frac{1}{2}(\mathcal{L}_{v\to a} + \mathcal{L}_{a\to v})$.

3.3 Modality-Specific Semantic Distillation

Semantic supervision from pretrained encoders can make VAE latents easier for downstream generative models to learn [23, 36, 37]. We therefore distill semantic features into the audio and video branches independently, improving the structure of each latent space without requiring cross-modal inputs.

Teacher Selection. Our data include both images and videos on the visual side, and both general audio and speech on the audio side. Qwen3-Omni [10] provides a visual encoder shared across images and videos and an audio encoder that covers both general audio and speech. We freeze these two encoders and use them as the video and audio teachers, respectively.

Feature Distillation. For each modality, a lightweight projector maps the VAE latent to the channel dimension and temporal resolution of its teacher features, using convolution for channel projection and average pooling for temporal alignment. For video, the first causal latent frame corresponds only to the first input frame, so we align it with the teacher's image features and align the remaining latent frames with its video features. The audio branch uses a 1D convolutional projector. Following iREPA [37], we spatially normalize the visual teacher features to emphasize local variation over the dominant global component. At each aligned position j, we apply a sigmoid-cosine loss between the projected latent $\tilde{z}_j$ and teacher feature f_j:

$$\mathcal{L}_{\cos} = -\frac{1}{|J|}\sum_{j\in J}\log\sigma\big(\cos(\tilde{z}_j, f_j)\big), \tag{2}$$

where J indexes spatio-temporal positions for video and temporal positions for audio. This yields the video and audio distillation losses $\mathcal{L}_{\text{vd}}$ and $\mathcal{L}_{\text{ad}}$, respectively.

3.4 Training Objective

Loss Formulation. We define separate objectives for the video and audio VAEs. Each contains its modality-specific reconstruction and distillation losses, together with the shared segment-level contrastive loss:

$$\begin{aligned} \mathcal{L}_v &= \lambda_{\text{vr}}\mathcal{L}_{\text{vr}} + \lambda_{\text{vd}}\mathcal{L}_{\text{vd}} + \lambda_{\text{avclip}}\mathcal{L}_{\text{avclip}}, \\ \mathcal{L}_a &= \lambda_{\text{ar}}\mathcal{L}_{\text{ar}} + \lambda_{\text{ad}}\mathcal{L}_{\text{ad}} + \lambda_{\text{avclip}}\mathcal{L}_{\text{avclip}}. \end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{vr}}$ and $\mathcal{L}_{\text{ar}}$ are the video and audio reconstruction losses, and $\mathcal{L}_{\text{vd}}$ and $\mathcal{L}_{\text{ad}}$ are their distillation losses. The reconstruction terms comprise L1, LPIPS [38], and KL for video, and waveform, multi-resolution mel, and KL losses for audio. The shared $\mathcal{L}_{\text{avclip}}$ appears in both objectives because it supervises both encoders, but is evaluated only once during joint optimization. Since these loss groups differ substantially in scale, we set the coefficients λ using loss-magnitude weighting to keep their contributions comparable during training.

Training Schedule. We use a three-stage schedule to accommodate the different convergence rates and memory requirements of reconstruction and semantic learning. **Stage 1** independently pretrains the two VAEs using only reconstruction objectives, allowing them to acquire basic reconstruction capability through substantially more training steps than required by the semantic objectives. Both branches use adversarial supervision, and the audio branch additionally uses discriminator feature matching. **Stage 2** introduces contrastive alignment and semantic distillation. Contrastive learning requires long clips and a large negative pool, making this stage particularly memory intensive. We therefore remove the discriminators and feature matching, and jointly train both VAEs with the contrastive head and distillation projectors. Removing the audio discriminator, however, can introduce mild electronic artifacts in reconstructed audio. In **Stage 3**, we consequently freeze the audio encoder and fine-tune only its decoder with the Stage 1 audio losses, restoring adversarial and feature-matching supervision. This refinement improves perceptual audio quality without altering the aligned latent space; the video branch remains unchanged.

3.5 Downstream Generation

To verify that OmniVAE's latent space benefits downstream generation, we freeze the trained OmniVAE as a tokenizer and build a joint text-to-audio-video (T2AV) model on top of it, trained with flow matching in the video and audio latent spaces. Following common practice, we proceed in two stages: we first pretrain strong single-modality priors—a visual branch (T2I/T2V) and an audio branch (T2A)—and then couple them through lightweight cross-attention bridges and train jointly with the per-modality flow-matching losses. The model supports both joint and single-modality decoding at inference.

4 Experiments

4.1 Experimental Setup

Datasets. Existing large-scale audio-visual datasets are dominated by human speech. To enrich the model's knowledge of the associations between non-speech sound events and their visual sources, we devise a top-down data collection pipeline. We first use an LLM to construct an audio-visual concept tree, which is subsequently verified by human annotators. For each leaf concept, the LLM then generates multiple search queries to retrieve videos whose visual and acoustic content matches the concept. This pipeline substantially broadens the coverage and diversity of sound events and audio-visual concepts in our corpus. Web-sourced videos, however, are noisy and often exhibit weak visual-audio correlation—e.g., background music or first-person voice-overs decoupled from the on-screen content. To obtain data with fine-grained audio-visual correspondence, we further apply a filtering pipeline that combines audio-visual semantic alignment scoring [12], DeSync-based synchronization filtering [8], and audio event detection [39]. Combined with AudioSet [40], the resulting corpus contains about 23M audio-visual clips, evenly split between speech and non-speech samples. Joint generation training further requires three types of captions—visual, audio, and

unified audio-visual—for which we use Qwen3.5-VL [41] for visual captions and two in-house fine-tuned variants of Qwen3-Omni [10] for the audio and unified audio-visual captions, respectively.

Model. For semantic distillation, we use features from layer 27 of the Qwen3-Omni [10] visual encoder and layer 18 of its audio encoder as supervision. In terms of model size, the two VAEs comprise approximately 1.08B parameters and are the only components retained at inference. The contrastive head and distillation projectors add approximately 49M training-only parameters and are discarded after training, introducing no inference overhead. Table 1 provides the detailed parameter breakdown.

Module	#Params
Retained at inference	
Video VAE (enc. + dec.)	704.7 M
Audio VAE (enc. + dec.)	371.6 M
Subtotal	*1,076.3 M*
Training-only	
Contrastive head	20.5 M
Distill. projectors	29.0 M
Subtotal	*49.5 M*
Total (training)	**1,125.8 M**

Table 1 Parameter counts. Training-only modules are discarded at inference.

Training and Baselines. We organize the experiments as a progressive ablation over training objectives that isolates the contribution of reconstruction, semantic distillation, and audio-visual contrastive learning: **(i) Recon**—only the per-modality reconstruction losses, with the video and audio VAEs trained independently; **(ii) Recon+Distill**—adds modality-specific semantic distillation, still training the two VAEs independently; **(iii) Recon+AVCLIP**—instead adds the segment-level audio-visual contrastive loss and trains the two VAEs jointly; and **(iv) OmniVAE** (full model)—trains the two VAEs jointly with reconstruction, distillation, and contrastive objectives enabled simultaneously. The video and audio branches of OmniVAE are initialized from the Recon+Distill checkpoints, while the contrastive head is trained from scratch.

The per-modality variants (i, ii) are trained for 250k steps on 121-frame clips, and the joint variants (iii, iv) for 82k steps on 193-frame clips with a global batch size of 256. Joint training follows a two-stage schedule that first freezes the video VAE to prioritize audio reconstruction and cross-modal alignment, and then unfreezes it to jointly optimize all objectives. Since joint training omits the audio discriminators for memory efficiency, we briefly fine-tune the audio decoder of the AVCLIP-based variants with adversarial supervision, removing mild electronic artifacts with negligible changes in objective metrics. The relative scales of the loss groups follow the contrastive-anchored loss-magnitude weighting discussed in §4.5.2. All variants use the same video and audio VAE backbones, training data, and optimization settings.

4.2 Reconstruction Evaluation

We first compare the reconstruction quality of different training configurations. The video VAE is evaluated on UCF-101 [42] and Panda-70M [43] at 24 fps, 121 frames, and 256×256. The audio VAE is evaluated on LibriSpeech [44] test-clean for speech, AudioSet [40] for general audio, and MUSDB18 [45] for music.

Table 2 reports the results. On the video side, all training variants remain close to the reconstruction-only baseline across the reconstruction metrics, indicating that the additional objectives largely preserve video reconstruction quality. Our Recon model is initialized from the official Wan2.2 VAE checkpoint and then fine-tuned using only reconstruction objectives. The official Wan VAEs are trained across diverse video settings and are therefore included for reference only. On the audio side, our models build upon the open-source HunyuanVideo-Foley VAE. Semantic distillation largely preserves reconstruction quality, while contrastive alignment leads to only a slight degradation. Even so, OmniVAE remains competitive with or outperforms the external audio VAEs on most reported metrics. Overall, OmniVAE retains strong reconstruction quality across video, speech, general audio, and music.

(a) Video Reconstruction *UCF-101 / Panda-70M*

Configuration	PSNR ↑	SSIM ↑	LPIPS ↓	L1 ↓	rFVD ↓
External VAEs					
Wan2.1 VAE [18]	34.50 / 32.62	0.9510 / 0.9448	0.0201 / 0.0169	0.0117 / 0.0130	2.46 / 1.49
Wan2.2 VAE [35]	34.87 / 33.21	0.9584 / 0.9541	0.0177 / 0.0137	0.0110 / 0.0120	3.72 / 1.38
Ours					
Recon	36.93 / 36.56	0.9656 / 0.9697	0.0102 / 0.0067	0.0090 / 0.0086	2.40 / 1.25
Recon + Distill	36.70 / 36.22	0.9649 / 0.9688	0.0107 / 0.0071	0.0092 / 0.0089	2.53 / 1.26
Recon + AVCLIP	36.22 / 35.13	0.9616 / 0.9631	0.0115 / 0.0079	0.0097 / 0.0098	2.66 / 1.35
OmniVAE	36.27 / 35.24	0.9616 / 0.9632	0.0113 / 0.0077	0.0097 / 0.0098	2.81 / 1.31

(b) Audio Reconstruction

	Speech				Audio / Music		
Model	SIM ↑	STOI ↑	P-NB ↑	P-WB ↑	Mel ↓	STFT ↓	ViSQOL ↑
External VAEs							
HunyuanVideo-Foley VAE [7]	0.9899	0.9979	4.4780	4.5337	0.52 / 0.60	1.82 / 1.84	4.47 / 4.45
MMAudio VAE [6]	0.9122	0.9502	3.4931	2.9797	1.23 / 1.36	2.14 / 2.44	4.42 / 4.42
Stable Audio 3 SAME-S [46]	0.7960	0.9139	2.8073	2.2991	1.46 / 1.37	3.19 / 3.67	3.68 / 3.99
Stable Audio 3 SAME-L [46]	0.8755	0.9539	3.4615	3.0739	1.37 / 1.22	3.30 / 3.55	3.76 / 4.17
Ours							
Recon	0.9893	0.9981	4.4916	4.5557	0.44 / 0.50	1.75 / 1.72	4.55 / 4.60
Recon + Distill	0.9893	0.9970	4.4453	4.5101	0.56 / 0.64	1.96 / 1.95	4.44 / 4.42
Recon + AVCLIP	0.9666	0.9946	4.4346	4.3638	0.76 / 0.75	1.99 / 2.03	4.03 / 4.18
OmniVAE	0.9737	0.9948	4.4400	4.4085	0.74 / 0.73	2.02 / 2.08	4.08 / 4.31

Table 2 Video and audio reconstruction quality. Video values follow UCF-101 / Panda-70M, while Audio / Music values follow AudioSet / MUSDB18. All our video variants are initialized from Wan2.2, and our audio variants from the HunyuanVideo-Foley VAE. External models serve as reference points rather than training-matched baselines.

4.3 Audio-Video Sync Probing

Evaluating fine-grained audio-video alignment through downstream joint generation is expensive, as each comparison requires training a full generative model. We therefore adopt audio-video sync probing as an efficient and sensitive proxy. The probe predicts the temporal offset between audio and video, requiring it to associate sound events with their visual sources at fine temporal granularity.

We evaluate on VGGSound-Sparse (VGS-Sp), which contains sparse, temporally localizable events [8, 47, 48]. Following Synchformer, we shift audio by $N \times 0.2$ seconds and train a lightweight head to classify 21 offsets from -2 to $+2$ seconds. We report A@1, A@5, and A@1_{tol} with a ±1-class tolerance. Under **frozen**, only the classifier is trained; **+ft** also tunes the contrastive aggregators. Table 3 shows that contrastive learning provides the main gain in temporal alignment, with OmniVAE performing best under both protocols.

Method	**A@1**	**A@5**	**A@1$_{\text{tol}}$**
Frozen			
Recon	6.4	31.5	16.2
Recon + Distill	7.1	33.0	17.5
Recon + AVCLIP	18.1	47.8	34.5
OmniVAE (ours)	**20.2**	**50.1**	**37.2**
+ft			
Recon + ft	7.0	32.4	17.1
Recon + Distill + ft	7.8	33.9	18.4
Recon + AVCLIP + ft	21.8	52.4	39.6
OmniVAE + ft	**22.3**	**53.2**	**40.4**

Table 3 Sync probing on VGS-Sp (%); A@1_{tol} uses ±1 tolerance.

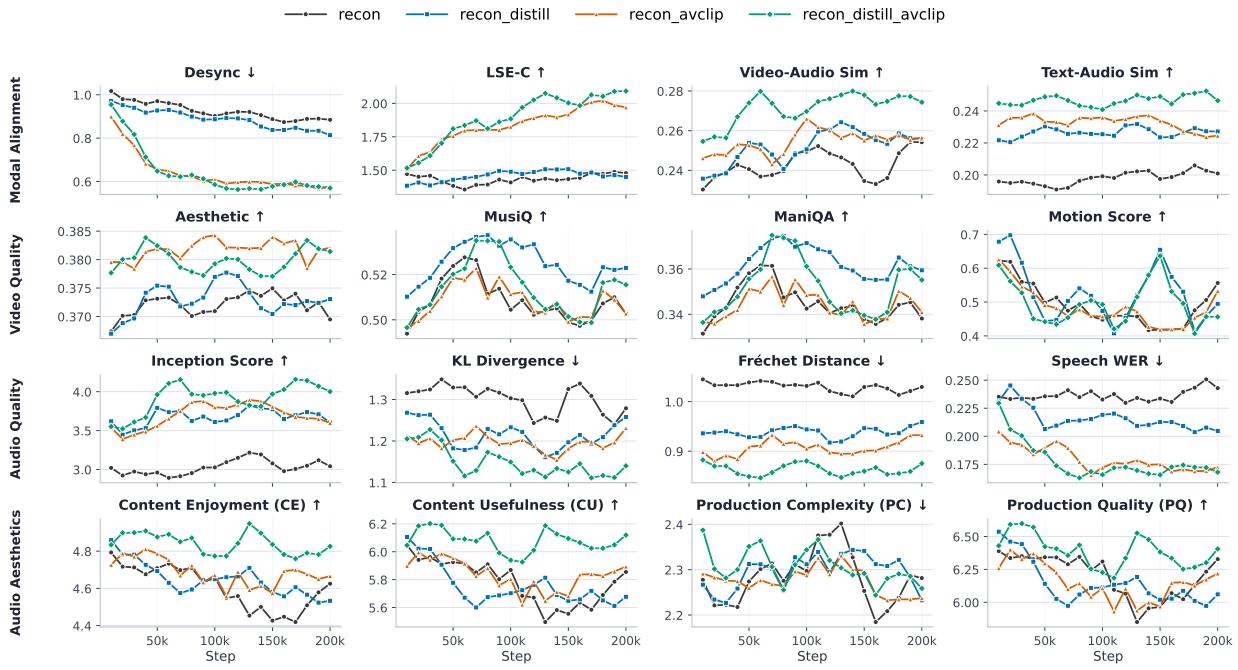

Figure 2 Evaluation metrics of T2AV models trained on the four VAE configurations, tracked across training steps. Rows, top to bottom: cross-modal alignment (Desync, LSE-C, Video-Audio Sim, Text-Audio Sim), video quality (Aesthetic, MusiQ, ManiQA, Motion Score), AudioBox aesthetics [49] (Content Enjoyment, Content Usefulness, Production Complexity, Production Quality), and audio quality (IS, KL Divergence, Fréchet Distance, Speech WER). The AVCLIP-equipped configurations consistently lead on all cross-modal alignment metrics, both semantic objectives improve audio quality, and the full OmniVAE (*recon_distill_avclip*) achieves the best overall results.

4.4 Text-to-Audio-Video Generation

Benchmark. We evaluate T2AV generation on Verse-Bench [50] using the unified T2AV prompts introduced in MOVA [34]. Verse-Bench comprises three subsets: Set1-I (205 samples sourced from still images), Set2-V (295 samples sourced from web videos), and Set3-TED (100 TED-talk samples used for lip-sync evaluation). Because our T2AV models generate at 256×256 resolution, faces are often too small for SyncNet [17] to detect, leaving only a small number of samples for which a lip-sync score can be computed. We therefore use an LLM to rewrite each Set3 visual prompt into a medium close-up or close-up shot while preserving its original intent and speech content, so that faces occupy a larger fraction of the frame. All baselines are evaluated on the rewritten prompts for fair comparison. Set3 results under this protocol are not directly comparable to those reported using the original Verse-Bench prompts.

Results. Figure 2 compares the four VAE configurations throughout T2AV training. For a checkpoint-independent quantitative comparison, Table 4 reports the mean over the final three checkpoints (180k, 190k, and 200k) under CFG = 5, using the same checkpoints for every configuration. On **cross-modal alignment**, Desync measures fine-grained temporal synchronization and LSE-C measures lip synchronization, while Video-Audio Sim and Text-Audio Sim evaluate overall semantic consistency. The AVCLIP-based variants consistently improve both groups of metrics, with OmniVAE achieving the strongest overall alignment. On **audio quality**, the AudioBox scores jointly assess content and production quality, IS, KL divergence, and Fréchet distance evaluate non-speech audio generation, and WER measures speech intelligibility. Both semantic distillation and AVCLIP improve these metrics individually, while their combination achieves the strongest overall audio performance. On **video quality**, AVCLIP improves visual aesthetics, while semantic distillation yields gains on MusiQ and ManiQA; motion quality remains broadly comparable across configurations. Together, these results show that AVCLIP improves both cross-modal alignment and audio genera-

Cross-Modal Alignment				
Configuration	DeSync ↓	LSE-C ↑	Video-Audio Sim ↑	Text-Audio Sim ↑
Recon	0.884	1.479	0.254	0.201
Recon + Distill	0.814	1.450	0.256	0.227
Recon + AVCLIP	0.576	1.970	0.257	0.225
OmniVAE	**0.570**	**2.093**	**0.274**	**0.246**
Video Quality				
Configuration	Aesthetic ↑	MusiQ ↑	ManiQA ↑	Motion Score ↑
Recon	0.369	0.503	0.338	**0.556**
Recon + Distill	0.373	**0.523**	**0.359**	0.495
Recon + AVCLIP	**0.382**	0.503	0.341	0.533
OmniVAE	0.381	0.515	0.355	0.456
Audio Quality				
Configuration	IS ↑	KL ↓	FD ↓	WER ↓
Recon	3.041	1.279	1.029	0.243
Recon + Distill	3.604	1.258	0.959	0.205
Recon + AVCLIP	3.598	1.231	0.932	0.172
OmniVAE	**4.001**	**1.140**	**0.875**	**0.168**
AudioBox Aesthetics				
Configuration	CE ↑	CU ↑	PC ↓	PQ ↑
Recon	4.625	5.855	2.282	6.328
Recon + Distill	4.533	5.677	**2.232**	6.060
Recon + AVCLIP	4.666	5.894	2.237	6.219
OmniVAE	**4.826**	**6.119**	2.259	**6.406**

Table 4 T2AV generation results averaged over the final three checkpoints (180k, 190k, and 200k) with CFG = 5. All metrics are aggregated over the supported Verse-Bench subsets; LSE-C is evaluated on the face-detectable Set3 samples. CE, CU, PC, and PQ denote Content Enjoyment, Content Usefulness, Production Complexity, and Production Quality, respectively. Bold indicates the best result for each metric.

tion quality, while semantic distillation further strengthens modality-specific generation quality. Combining the two objectives yields complementary gains and the strongest overall performance.

4.5 Ablation Study

We ablate two key designs: the temporal granularity of contrastive learning (§4.5.1) and the loss-balancing strategy (§4.5.2). We evaluate alignment using sync probing for the former and segment-retrieval accuracy for the latter, while also reporting reconstruction quality.

4.5.1 Contrastive Pool Size and Temporal Granularity

The contrastive pool is determined by the video frame rate and clip duration, while the segment length controls temporal granularity. As shown in Table 5, using a higher frame rate together with finer temporal segments, or increasing the clip duration, substantially improves sync probing accuracy, demonstrating the importance of a sufficiently large negative pool. By contrast, performance is relatively stable across segment lengths. We therefore use 24 fps, 8-second clips, and 0.17-second segments in the main experiments, which provide a large negative pool at the finest evaluated temporal granularity.

FPS	Clip	Segment	A@1 (%)
8	8 s	0.50 s	19.0
24	—	0.17 s	7.0
24	2 s	0.17 s	6.0
24	8 s	0.17 s	**30.0**
24	8 s	0.34 s	27.0
24	8 s	0.50 s	29.0

Table 5 Ablation of contrastive pool size and temporal granularity. A@1 is measured using the VGS-Sp temporal-offset probe in Table 3. The dash denotes the reconstruction-only baseline without contrastive training.

4.5.2 Balancing Reconstruction and Cross-Modal Alignment

Joint training combines video reconstruction, audio reconstruction, and contrastive alignment losses with substantially different scales. Without appropriate balancing, the dense reconstruction objectives can overwhelm the comparatively sparse alignment signal. We therefore compare three strategies for balancing these loss groups: loss-magnitude weighting, gradient-based weighting using the last encoder layer as a proxy, and uncertainty-based weighting [51].

Because all strategies use the same contrastive setup, we evaluate alignment through segment retrieval. **Intra A@48** retrieves the paired segment within the same video, whereas **Overall A@64** additionally includes negatives sampled from other videos in the batch. As shown in Table 6, all three strategies achieve comparable reconstruction quality, but loss-magnitude weighting performs substantially better on both retrieval metrics. We therefore adopt it for joint training.

Balancing strategy	Segment retrieval		Audio recon.			Video recon.
	Intra A@48	Overall A@64	STOI	PESQ-WB	PESQ-NB	PSNR
Loss-magnitude (ours)	**0.51**	**0.24**	0.994	4.41	4.38	34.1
GradNorm (last-layer proxy)	0.44	0.17	**0.995**	4.42	4.40	**35.4**
Uncertainty-based adaptive	0.29	0.13	0.994	**4.45**	**4.41**	33.4

Table 6 Ablation of loss-balancing strategies. Segment retrieval measures fine-grained alignment, while STOI, PESQ, and PSNR measure reconstruction quality.

5 Conclusion

We presented OmniVAE, a jointly trained audio-video VAE that embeds modality-specific semantics and fine-grained cross-modal correspondence into the latent representations. OmniVAE augments standard reconstruction training with segment-level audio-video contrastive learning and modality-specific semantic distillation. Our experiments show that the two objectives provide complementary benefits. Contrastive learning substantially improves fine-grained audio-video alignment, while semantic distillation strengthens the learnability of the modality-specific representations. These enhanced representations enable downstream text-to-audio-video generation models to achieve better generation quality and more accurate cross-modal synchronization without materially compromising reconstruction quality. Taken together, these findings highlight the importance of incorporating both modality-specific semantic structure and cross-modal correspondence directly into learned representations, providing a stronger foundation for unified multimodal

generation. Future work should explore more effective methods for modality-specific and cross-modal semantic learning and validate their benefits at larger scales.

References

[1] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6114.

[2] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. *arXiv preprint arXiv:2501.01423*, 2025. URL https://arxiv.org/abs/2501.01423.

[3] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders, 2025. URL https://arxiv.org/abs/2510.11690.

[4] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. SpeechTokenizer: Unified speech tokenizer for speech large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2308.16692.

[5] OpenMOSS Team. MOSS-Audio-Tokenizer: Scaling audio tokenizers for future audio foundation models, 2026. URL https://arxiv.org/abs/2602.10934.

[6] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28901–28911, 2025. doi: 10.1109/CVPR52734.2025.02691.

[7] Sizhe Shan, Qiulin Li, Yutao Cui, Miles Yang, Yuehai Wang, Qun Yang, Jin Zhou, and Zhao Zhong. Hunyuanvideo-foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation, 2025. URL https://arxiv.org/abs/2508.16930.

[8] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5325–5329, 2024. doi: 10.1109/ICASSP48485.2024.10448489.

[9] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. URL https://arxiv.org/abs/1807.03748.

[10] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, et al. Qwen3-Omni technical report, 2025. URL https://arxiv.org/abs/2509.17765.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, 2021. URL https://arxiv.org/abs/2103.00020.

[12] 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. URL https://arxiv.org/abs/2305.05665.

[13] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. LanguageBind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.01852.

[14] Alexander H. Liu, SouYoung Jin, Cheng-I Lai, Andrew Rouditchenko, Aude Oliva, and James Glass. Cross-modal discrete representation learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3013–3035, 2022. URL https://aclanthology.org/2022.acl-long.215/.

[15] Yan Xia, Hai Huang, Jieming Zhu, and Zhou Zhao. Achieving cross modal generalization with multimodal unified representation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/c89f09849eb5af489abb122394ff0f0b-Abstract-Conference.html.

[16] Hai Huang, Yan Xia, Shengpeng Ji, Shulei Wang, Hanting Wang, Minghui Fang, Jieming Zhu, Zhenhua Dong, Sashuai Zhou, and Zhou Zhao. Enhancing multimodal unified representations for cross modal generalization. In

Findings of the Association for Computational Linguistics (ACL Findings), 2025. URL https://aclanthology.org/2025.findings-acl.119/.

[17] Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Computer Vision – ACCV 2016 Workshops*, pages 251–263, 2017. doi: 10.1007/978-3-319-54427-4_19.

[18] Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, et al. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

[19] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models, 2024. URL https://arxiv.org/abs/2410.13720.

[20] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGAN. In *Advances in Neural Information Processing Systems*, volume 36, pages 27980–27993, 2023. URL https://arxiv.org/abs/2306.06546.

[21] Zach Evans, CJ Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2402.04825.

[22] Ziyue Jiang, Dake Guo, Zekai Zhang, Hangrui Hu, Ting He, Xinfa Zhu, Xiong Wang, Yongqi Wang, Jiapeng Wang, Wenxiang Guo, Zhifang Guo, Chenfei Wu, Dayiheng Liu, and Jin Xu. Qwen-audio-vae technical report. *arXiv preprint arXiv:2607.11738*, 2026.

[23] Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. REPA-E: Unlocking VAE for end-to-end tuning with latent diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2504.10483.

[24] Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. UniTok: A unified tokenizer for visual generation and understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2502.20321.

[25] Jingfeng Yao, Yuda Song, Yucong Zhou, and Xinggang Wang. Towards scalable pre-training of visual tokenizers for generation, 2025. URL https://arxiv.org/abs/2512.13687.

[26] 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. doi: 10.1109/CVPR52729.2023.00985.

[27] Kai Wang, Shijian Deng, Jing Shi, Dimitrios Hatzinakos, and Yapeng Tian. AV-DiT: Efficient audio-visual diffusion transformer for joint audio and video generation, 2024. URL https://arxiv.org/abs/2406.07686.

[28] OpenAI. Sora 2 system card. https://openai.com/index/sora-2-system-card/, 2025. Accessed: 2026-07.

[29] Google DeepMind. Veo 3. https://deepmind.google/models/veo/, 2025. Accessed: 2026-07.

[30] Seedance Team, ByteDance. Seedance 2.0: Advancing video generation for world complexity, 2026. URL https://arxiv.org/abs/2604.14148.

[31] Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Fei, and Tat-Seng Chua. JavisDiT: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization, 2025. URL https://arxiv.org/abs/2503.23377.

[32] 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.

[33] Yoav HaCohen et al. LTX-2: Efficient joint audio-visual foundation model, 2026. URL https://arxiv.org/abs/2601.03233.

[34] SII-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, 2026. URL https://arxiv.org/abs/2602.08794.

[35] Wan-AI. Wan2.2-T2V-A14B model card. Hugging Face model card, 2025. URL https://huggingface.co/Wan-AI/Wan2.2-T2V-A14B. Accessed: 2026-07-25.

[36] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=DJSZGGZYVi.

[37] Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure?, 2025. URL https://arxiv.org/abs/2512.10794.

[38] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

[39] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, pages 5178–5193, 2023. URL https://arxiv.org/abs/2212.09058.

[40] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 776–780, 2017. doi: 10.1109/ICASSP.2017.7952261.

[41] Qwen Team. Qwen3.5: Towards native multimodal agents, 2026. URL https://qwen.ai/blog?id=qwen3.5.

[42] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild, 2012. URL https://arxiv.org/abs/1212.0402.

[43] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-Wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70M: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. doi: 10.1109/CVPR52733.2024.01265.

[44] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

[45] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. The MUSDB18 corpus for music separation. In *International Society for Music Information Retrieval Conference, Late-Breaking/Demo Session*, 2017. URL https://sigsep.github.io/datasets/musdb.html.

[46] Zach Evans, Julian D. Parker, Matthew Rice, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio 3, 2026. URL https://arxiv.org/abs/2605.17991.

[47] 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*, pages 721–725, 2020. doi: 10.1109/ICASSP40776.2020.9053174.

[48] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Sparse in space and time: Audio-visual synchronisation with trainable selectors. In *British Machine Vision Conference*, 2022. URL https://arxiv.org/abs/2210.07055.

[49] 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. URL https://arxiv.org/abs/2502.05139.

[50] 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, 2025. URL https://arxiv.org/abs/2509.06155.

[51] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018. doi: 10.1109/CVPR.2018.00781.