Beyond Co-Generation: Rethinking Jointness in Joint Audio–Video Generation

Jihoo Jung[1] Ji-Hoon Kim[2] Suyeon Lee[1] Gia Thinh Pham[1] Joon Son Chung[1]
[1]KAIST [2]Chung-Ang University

Abstract

Our aim is to advance the understanding of genuine audio-visual coordination in joint text-to-audio-video (T2AV) generation, with particular emphasis on semantic consistency. Evaluating audio-visual consistency in T2AV generation is nontrivial, as detailed text prompts obscure whether such consistency stems from genuine cross-modal coordination or merely from independent text following. To isolate the former, we introduce intrinsic jointness: the ability of audio and video to produce the same realization for details left underspecified by the prompt. Using semantically underspecified prompts (e.g., "musical instrument playing" rather than "electric guitar playing"), we investigate four recent models and reveal their limited intrinsic jointness. To address this limitation, we propose JointREPA, a simple representation alignment-based post-training method to facilitate cross-modal coordination. Since existing benchmarks use semantically detailed prompts, we introduce UNSAIDBench, a benchmark designed to evaluate intrinsic jointness under semantic underspecification. Experiments across two distinct models and three benchmarks, including UNSAIDBench, demonstrate that JointREPA consistently improves semantic consistency.

1. Introduction

Recent advances in text-to-image (T2I) and text-to-video (T2V) generation have enabled high-quality visual synthesis from natural language descriptions [3, 8, 17, 20, 43, 46]. Building on this progress, recent works have explored joint text-to-audio-video (T2AV) generation, in which both audio and video are synthesized from a shared text prompt [10, 15, 23, 26, 27, 29, 32, 35, 36, 39, 40]. Unlike single-modality generation tasks such as T2I and T2V, where faithfully following the text prompt is the main focus, T2AV additionally requires consistency between the two generated modalities. This cross-modal consistency encompasses two key aspects: *semantic consistency*, which requires audio and video to depict the same events, and *temporal consistency*, which requires corresponding events to occur at aligned times.

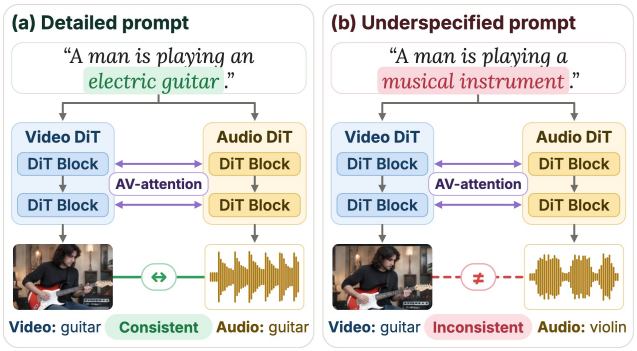


Figure 1. **Intrinsic jointness.** Under detailed prompts, output consistency can arise from each modality independently following the text prompt alone, without audio–visual interaction. In contrast, underspecified prompts enable disentangling whether this consistency stems from mere prompt following or genuine audio–visual coordination. We define intrinsic jointness as the ability to maintain consistency under such underspecified prompts.

While recent T2AV models appear to exhibit such consistency, this does not necessarily imply that the two modalities are genuinely coordinated. When the prompt specifies every detail, such consistency can arise simply from each modality following the prompt independently, even without cross-modal coordination. To evaluate cross-modal coordination beyond text following, we therefore examine **intrinsic jointness**: the ability of audio and video to converge on the same realization for details that the prompt leaves open to multiple possibilities. For example, an underspecified prompt such as "A man is playing a *musical instrument*" (Fig. 1(b)), which admits multiple valid realizations (*e.g.*, guitar or violin), requires the two modalities to actively coordinate their choice to reliably produce consistent outputs. Intrinsic jointness matters in practice, as users rarely specify every detail in their prompts [7, 16]. It is also important for physically coherent generation, where unstated object properties (*e.g.*, size or material) should jointly shape both sound and visuals [21].

Existing benchmarks do not equally evaluate the semantic and temporal dimensions of intrinsic jointness. Since natural text prompts rarely state temporal properties such as exact timestamps [42, 44], temporal coordination beyond text following has been widely examined. In contrast, seman-

tic information is often specified in detail within standard benchmarks [2, 19, 26], leaving the semantic dimension largely underexplored. In this work, we focus specifically on this overlooked dimension of intrinsic jointness, hereafter referred to as *semantic jointness*.

We investigate semantic jointness in four recent open-source T2AV models: LTX-2.3, LTX-2 [15], JavisDiT++ [27], and JavisDiT [26]. We construct small diagnosis prompts across three levels of semantic specificity by progressively abstracting event identities (*e.g.*, electric guitar → string instrument → musical instrument). Across all four models, semantic consistency between the generated audio and video drops substantially as prompts become less specific, revealing limited semantic jointness. Probing intermediate features shows that this inconsistency is not confined to the final output but persists throughout internal representations. By analyzing audio–visual attention, which enables cross-modal interaction, we further reveal that this pathway is insufficiently utilized to facilitate semantic coordination.

Motivated by these findings, we propose **JointREPA**, a simple representation alignment [47] based post-training method for T2AV models. JointREPA aligns each modality's intermediate features with pretrained representations from both its own modality and the other modality, facilitating cross-modal coordination. We also introduce **UNSAIDBench** (**UN**derspecified **S**emantics in **A**udio-v**ID**eo generation), a benchmark designed to evaluate semantic jointness. Unlike existing benchmarks that often explicitly specify semantic details, UNSAIDBench features semantically underspecified prompts that admit multiple valid event realizations. Evaluated on UNSAIDBench as well as the standard JavisBench [26] and Verse-Bench [39], JointREPA yields consistent gains in audio–visual alignment across two distinct models.

Our contributions are threefold:

- To assess genuine audio–visual coordination in T2AV models, we introduce intrinsic jointness, defined as the ability of audio and video to converge on the same realization for details that remain underspecified in the text prompt.
- Using semantically underspecified prompts, we systematically analyze four recent T2AV models, uncovering substantial limitations in semantic jointness.
- We propose JointREPA, a lightweight post-training method to promote audio–visual coordination, improving semantic consistency across two models and three benchmarks.

2. Related Works

2.1. Joint Text-to-audio-video Generation (T2AV)

T2AV generation aims to synthesize audio and video simultaneously from a text prompt [10, 15, 23, 26, 27, 29, 32, 35, 36, 39, 40]. Dual-tower models (e.g., LTX-2, LTX-2.3 [15], Ovi [29], and JavisDiT [26]) employ separate Diffusion Transformer (DiT) [33] towers for audio and video. Each DiT block comprises unimodal self-attention, text cross-attention, and audio–visual cross-attention. Alternatively, single-tower models such as JavisDiT++ [27] flatten and concatenate audio and video tokens, feeding them into a shared DiT tower. Its DiT blocks consist of joint audio–visual self-attention, text cross-attention, and modality-specific FFNs. Most models adopt flow matching [24] as their primary training objective. While single-modality generation such as text-to-image (T2I) or text-to-video (T2V) primarily focuses on text alignment, T2AV generation additionally requires cross-modal consistency, both semantically and temporally. Although current models appear to achieve such consistency, we investigate whether this alignment truly stems from genuine cross-modal coordination.

2.2. Prompt Specificity in Generative Models

Prompt specificity has been widely studied in T2I and T2V generation from two main perspectives. One line of work investigates the impact of prompt specificity on generation performance. Zhang et al. [53] analyze how prompt complexity influences quality, diversity, and consistency in T2I generation. Liu et al. [28] categorizes prompts along multiple dimensions, including complexity, and evaluates T2V model performance across these dimensions. Another line of work views prompt underspecification as an issue to be mitigated. These approaches seek to resolve ambiguities left by the prompts by soliciting user clarifications [16, 31] or developing models for automatic prompt optimization [7]. In T2AV generation, such underspecification poses a different challenge: when multiple valid realizations exist, audio and video are required to converge on the same one. In this work, we use underspecified prompts as a probe to investigate whether models genuinely coordinate across modalities.

2.3. Representation Alignment for Generation

Aligning internal representations with pretrained encoders has recently been widely explored in generative models. REPA [47] aligns hidden states in denoising networks with clean image representations from pretrained visual encoders during T2I model training. This encourages the model to capture useful semantic information, improving both training efficiency and generation quality. Recent studies have further adapted this paradigm to other generative domains: VideoREPA [52] extends it to T2V generation and proposes Token Relation Distillation to improve physical plausibility; AG-REPA [51] applies representation alignment to text-to-audio generation, introducing a causal layer selection strategy to enhance audio quality; and TARO [38] and MGAudio [50] adapt it to video-to-audio generation to improve cross-modal coherence and audio realism. Our work extends this paradigm to T2AV generation, addressing the semantic

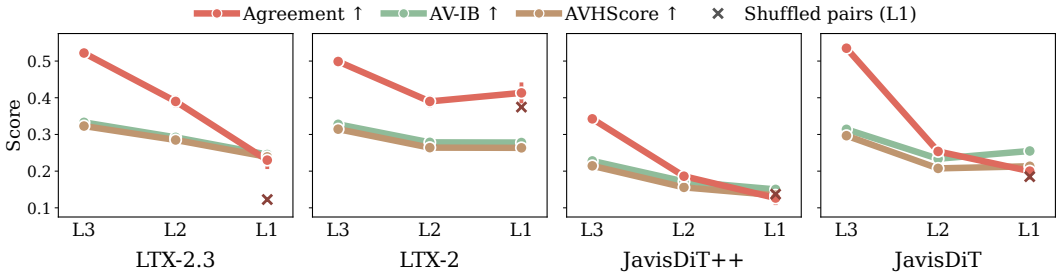


Figure 2. **Q1. Semantic jointness under varying prompt specificity.** We evaluate semantic consistency using three metrics (Agreement, AV-IB, and AVHScore) across different levels of prompt specificity. All metrics generally decrease as prompt specificity decreases (L3 → L2 → L1). At L1, Agreement scores for most models are close to the shuffled-pair baseline (marked by ×), obtained by randomly pairing audio and video generated with different seeds under the same prompt. This suggests that current T2AV models exhibit limited semantic jointness.

jointness bottleneck by aligning features both within and between the two modalities.

3. Diagnosing Semantic Jointness

In this section, we first formalize intrinsic jointness in Sec. 3.1, outline our analysis setup in Sec. 3.2, and present our findings in Sec. 3.3.

3.1. Defining Intrinsic Jointness

Given a text prompt c, let $p_\theta(a, v \mid c)$ denote the distribution of generated audio–video pairs. Since prompts rarely specify every detail, sampling from p_θ necessitates realizing the details left unspecified by c. Some such details are modality-specific (e.g., clothing color), while others must be resolved consistently across audio and video. We formalize the latter as a set of *underspecified shared factors* $\mathcal{Z}(c) = \{Z_1, \dots, Z_K\}$. For example, in Fig. 1(b), these factors include instrument type (Z_1), playing timing (Z_2) and others. Each Z_k is drawn from a set of possible realizations $\mathcal{R}_k(c)$ (*e.g.*, $\mathcal{R}_1(c) = \{\text{guitar}, \text{violin}, \dots\}$), forming the joint realization space $\mathcal{R}(c) = \mathcal{R}_1(c) \times \cdots \times \mathcal{R}_K(c)$.

Let $\mathbf{z}^a = (z_1^a, \dots, z_K^a)$ and $\mathbf{z}^v = (z_1^v, \dots, z_K^v) \in \mathcal{R}(c)$ denote their realizations in audio and video, respectively. We view joint generation $p_\theta(a, v \mid c)$ as realizing the underspecified shared factors as $\mathbf{z}^a$ and $\mathbf{z}^v$ in each modality. Audio and video are generated conditioned on c and their respective realizations:

$$p_\theta(a, v \mid c) = \sum_{\mathbf{z}^a, \mathbf{z}^v} \pi_\theta(\mathbf{z}^a, \mathbf{z}^v \mid c) \cdot p_\theta(a \mid c, \mathbf{z}^a) p_\theta(v \mid c, \mathbf{z}^v), \tag{1}$$

where $\pi_\theta(\mathbf{z}^a, \mathbf{z}^v \mid c)$ is the joint distribution over the underspecified shared factors realized by the two modalities.

Ideally, a T2AV model resolves underspecified shared factors consistently across audio and video, yielding the same realization in both modalities, i.e., $\mathbf{z}^a = \mathbf{z}^v$. We define **intrinsic jointness** under prompt c as the probability that audio and video agree on the underspecified shared factors:

$$J_\theta(c) = \Pr_\theta(\mathbf{z}^a = \mathbf{z}^v \mid c) = \sum_{\mathbf{z} \in \mathcal{R}(c)} \pi_\theta(\mathbf{z}, \mathbf{z} \mid c). \tag{2}$$

We focus on the semantic aspect of intrinsic jointness. That is, underspecified shared factors in prompt c concern semantic attributes, such as source or event identity, rather than temporal properties, such as event timing. Because natural text prompts often omit exact timestamps [42, 44], temporal coordination has been widely examined. In contrast, semantic information is typically specified in detail [2, 19, 26], leaving this semantic dimension—hereafter referred to as *semantic jointness*—largely underexplored.

3.2. Analysis Setup

Data. To evaluate semantic jointness, we construct prompts c with exactly one underspecified shared factor ($\mathcal{Z}(c) = \{Z_1\}$). For this factor, we focus on events clearly distinguishable in both audio and video [54], choosing three categories: *musical instrument playing*, *human speech*, and *animal vocalization* (Tab. 1). Guided by the AudioSet ontology [11], we organize each event into three levels of semantic specificity (L1–L3) such that the size of the valid realization set $|\mathcal{R}(c)|$ progressively shrinks: from a broad category (L1, *e.g.*, *musical instrument*) to a coarse class (L2, *e.g.*, *string instrument*) and a fine-grained event (L3, *e.g.*, *guitar*). We construct the test prompts by specifying a single underspecified shared factor at each level while keeping the surrounding text fixed (*e.g.*, *"A person is playing a ___."*).

Models. For each prompt, we generate 100 samples using different random seeds with four recent open-source T2AV models: three dual-tower DiTs (LTX-2.3, LTX-2 [15], JavisDiT [26]) and one single-tower DiT (JavisDiT++ [27]). Dual-tower models couple the two modalities through audio–visual cross-attention, whereas the single-tower model applies joint self-attention over concatenated audio and video tokens. For simplicity, we refer to these attention between audio and video tokens in either architecture as *AV-attention*. All models use flow matching [24] as their base objective.

Metrics. To evaluate semantic jointness, we operationalize $\mathbf{z}^a$ and $\mathbf{z}^v$ in Eq. (2) via classification over predefined realization space $\mathcal{R}(c)$. Specifically, we employ category- and modality-specific classifiers (*e.g.*, SigLIP [48] and LAION-

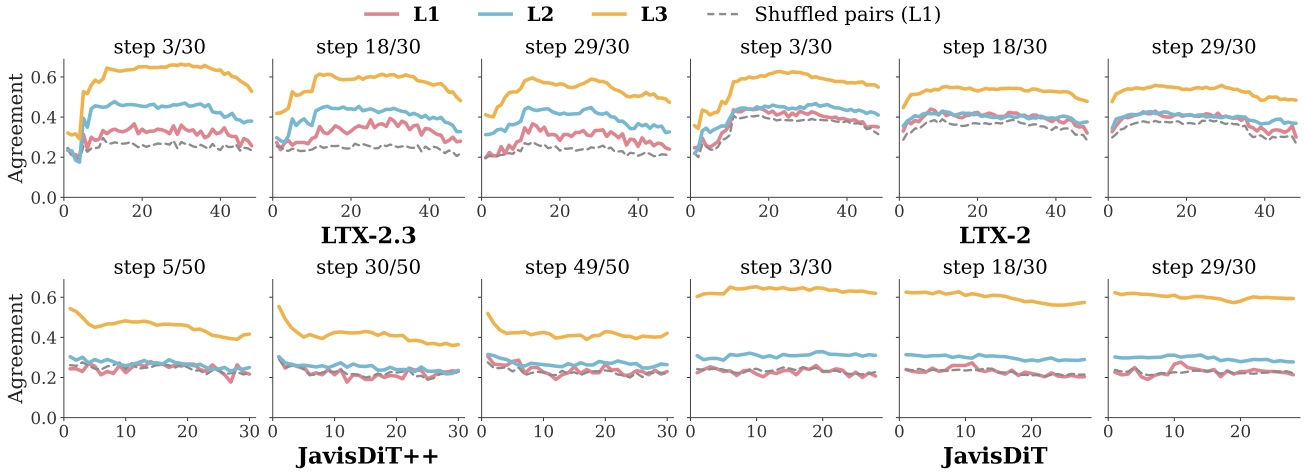


Figure 3. **Q2. Semantic jointness in internal representations.** We evaluate probe agreement across Transformer blocks (horizontal axis) at different generation steps and levels of prompt specificity. Probe agreement generally decreases as prompt specificity decreases (L3 → L2 → L1). At L1, probe agreement is close to the shuffled-pair baseline (dashed line). This suggests that limited semantic jointness is also reflected in the models' internal representations.

L1	L2	L3
"A/An ___ is vocalizing."		
animal	pet	dog / cat
	farm animal	cow / sheep
	wild animal	lion / wolf
"A person is playing a ___."		
musical instrument	string inst.	guitar / violin
	percussion inst.	drum / xylophone
	wind inst.	flute / trumpet
"A/An ___ is speaking."		
person	man	adult man / young boy
	woman	adult woman / young girl

Table 1. **Simple diagnostic set for semantic jointness analysis.** Prompts at L1, L2, and L3 share the same template but differ in the level of semantic specificity.

CLAP [41] for the video and audio streams of musical instruments, respectively) to independently predict the realized identities $\hat{\mathbf{z}}^a, \hat{\mathbf{z}}^v$. We then measure identity *agreement* over N generated samples as $\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{\mathbf{z}}_i^a = \hat{\mathbf{z}}_i^v)$. We also report two standard audio–visual semantic consistency metrics, AV-IB [13] and AVHScore [30].

3.3. Findings

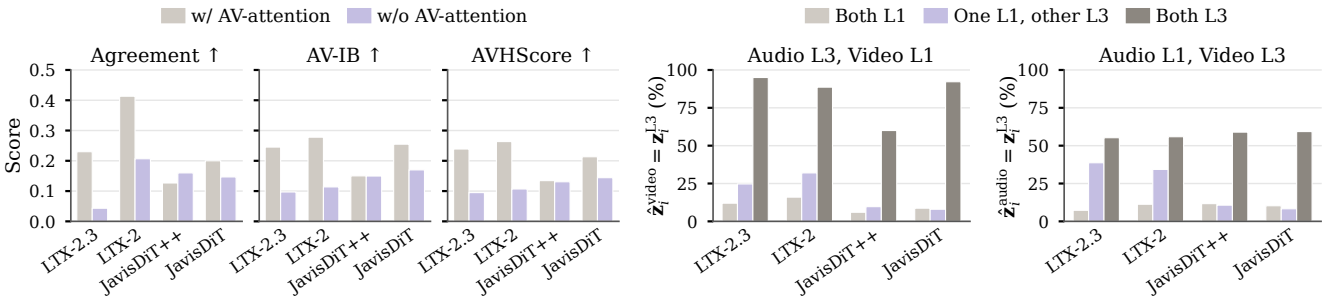
Q1. Do existing T2AV models exhibit semantic jointness? Fig. 2 plots Agreement, AV-IB, and AVHScore across specificity levels (L1–L3) for four models. We make two observations. First, all metrics generally degrade as prompts become less specific (L3 → L2 → L1). Second, for most models, L1 Agreement closely approaches a shuffled baseline that randomly pairs audio and video generated from different seeds within the same text prompt. Together, these results indicate that the two modalities resolve unspecified identities nearly independently, struggling to achieve true semantic jointness.

Q2. Does limited semantic jointness persist in internal representations? We examine whether this disagreement arises only at the final output or persists within internal representations. To this end, we measure internal agreement by probing intermediate features across Transformer blocks at various denoising steps. Specifically, at each block and step, we separately mean-pool audio and video tokens and train modality-specific linear probes to predict each modality's output identity label. We then measure the agreement between the two modalities' probe predictions (Fig. 3). Probe agreement generally decreases as prompts become less specific, with the L1 curves remaining close to the shuffled baseline. These findings indicate that limited semantic jointness persists in internal representations.

Q3. How effective is AV-attention for semantic coordination? Since cross-modal information exchange in T2AV models occurs through AV-attention, we conduct two experiments to examine its role in semantic coordination.

AV-attention knockout. First, we examine whether T2AV models actively utilize AV-attention during generation. Following attention knockout [12], we block AV-attention at inference time and measure how much semantic consistency degrades. As shown in Fig. 4(a), blocking AV-attention causes substantial drops across all metrics for the LTX family. In contrast, JavisDiT exhibits only a slight decline, and JavisDiT++ remains nearly unchanged. This indicates that the audio-visual pathway is active in the LTX family, but largely underutilized in the JavisDiT family.

Asymmetric prompting. Next, we test how effectively this pathway carries semantic information across modalities. Specifically, we condition one modality m_{L3} on an L3 prompt specifying a target identity $\mathbf{z}^{\mathrm{L3}}$ and the other



(a) AV-attention knockout (b) Asymmetric prompting

Figure 4. **Q3. Analysis of the role of AV-attention in semantic coordination.** (a) Semantic consistency with and without AV-attention. Blocking AV-attention causes substantial drops across all metrics in the LTX family, while inducing only minor changes in the JavisDiT family. (b) Semantic identity following under asymmetric prompting (L3 for one modality, L1 for the other). For both model families, the L1-conditioned modality m_{L1} follows the target identity specified $\mathbf{z}^{\mathrm{L3}}$ substantially less reliably than under symmetric L3 prompting (both-L3). Together, these findings suggest that AV-attention is not effectively utilized for semantic coordination.

modality m_{L1} on a broader L1 prompt, and measure the *follow rate*—the fraction of samples in which the predicted identity of L1-conditioned modality $\hat{\mathbf{z}}_i^{m_{\mathrm{L1}}}$ matches $\mathbf{z}_i^{\mathrm{L3}}$: $\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{\mathbf{z}}_i^{m_{\mathrm{L1}}} = \mathbf{z}_i^{\mathrm{L3}})$. In Fig. 4(b), the left plot feeds L3 prompt to audio and L1 to video to measure $\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{\mathbf{z}}_i^{\mathrm{video}} = \mathbf{z}_i^{\mathrm{L3}})$, and vice versa on the right. For comparison, we also reports symmetric prompting baselines, where both modalities are conditioned on the same prompt level: either L1 (both-L1) or L3 (both-L3). As in the asymmetric setting, we evaluate the fraction of samples where the predicted identity $\hat{\mathbf{z}}_i^{m}$ of each modality matches the target $\mathbf{z}_i^{\mathrm{L3}}$, given by $\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{\mathbf{z}}_i^{m} = \mathbf{z}_i^{\mathrm{L3}})$ (m = video for the left panel and m = audio for the right panel).

The LTX family exhibits a higher follow rate than the both-L1 baseline, indicating that semantic information can be transferred across modalities through AV-attention. However, the follow rate remains substantially lower than in the both-L3 setting, suggesting that AV-attention is not fully effective in facilitating semantic coordination. The JavisDiT family exhibits negligible gains over the both-L1 baseline in either direction, with follow rates remaining substantially below those in the both-L3 setting.

Overall, despite varying degrees of effectiveness across models, AV-attention remains limited in its ability to transfer semantic information across modalities in all four models.

4. Improving Semantic Jointness

Motivated by the findings in Sec. 3.3, we propose **JointREPA**, which extends REPA [47] to T2AV generation. JointREPA aligns each modality's representation not only with target features from its own pretrained encoder, but also with those from the other modality. The former builds semantically rich representations that provide clearer signals for AV-attention to leverage, while the latter actively promotes cross-modal semantic exchange.

As illustrated in Fig. 5, let h_ℓ^a and h_ℓ^v denote the pooled audio and video hidden states at layer ℓ, and r^a and r^v the

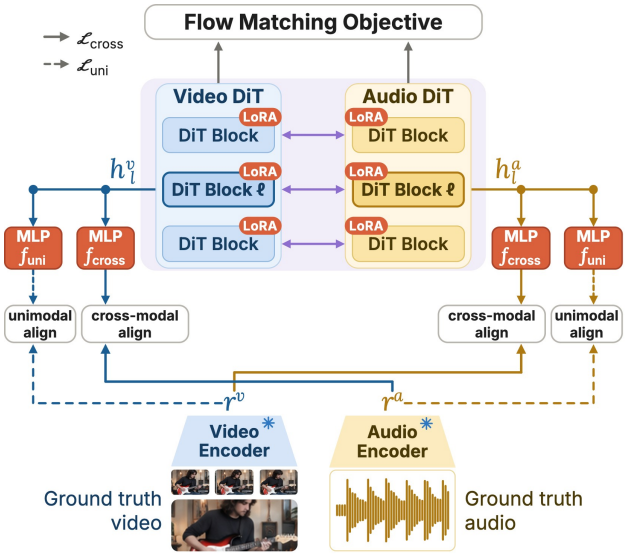


Figure 5. **Overview of JointREPA.** JointREPA aligns each modality's features with pretrained encoder representations from both its own modality and the other modality.

target features from frozen pretrained encoders (CLAP [9] for audio and V-JEPA [1] for video). Each modality uses two MLP projectors, f_{uni} and f_{cross}, to align its hidden states with the unimodal target and cross-modal counterpart, respectively. The unimodal ($\mathcal{L}_{\mathrm{uni}}$) and cross-modal ($\mathcal{L}_{\mathrm{cross}}$) alignment losses are formulated as:

$$\begin{aligned}\mathcal{L}_{\mathrm{uni}} &= -\mathbb{E}\Big[\,\mathrm{sim}\left(f^a_{\mathrm{uni}}(h^a_\ell), r^a\right) \\ &\qquad + \mathrm{sim}\left(f^v_{\mathrm{uni}}(h^v_\ell), r^v\right)\Big], \\ \mathcal{L}_{\mathrm{cross}} &= -\mathbb{E}\Big[\,\mathrm{sim}\left(f^a_{\mathrm{cross}}(h^a_\ell), r^v\right) \\ &\qquad + \mathrm{sim}\left(f^v_{\mathrm{cross}}(h^v_\ell), r^a\right)\Big],\end{aligned} \tag{3}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes a similarity function.

Table 2. **Main results on UNSAIDBench, JavisBench-mini, and VerseBench.** We apply JointREPA to JavisDiT++ and LTX-2.3 and evaluate AV-Consistency, AV-Synchrony, Text-Consistency, and AV-Quality on three benchmarks. JointREPA consistently improves AV-Consistency across all benchmarks and backbone models.

		AV-Consistency		AV-Synchrony		Text-Consistency				AV-Quality	
Model	Size	AV-IB ↑	AVHScore ↑	JavisScore ↑	DeSync ↓	TV-IB ↑	TA-IB ↑	CLIP ↑	CLAP ↑	VQ ↑	AQ ↑
UNSAIDBench											
JavisDiT++ [27]	2.1B	0.176	0.163	0.133	0.815	**0.180**	0.138	**0.313**	0.343	1.455	4.595
+ **JointREPA**	2.1B	**0.207**	**0.196**	**0.164**	**0.804**	0.174	**0.163**	0.310	**0.369**	**1.655**	**4.628**
LTX-2.3 [15]	22B	0.244	0.236	0.209	0.468	**0.173**	0.101	0.298	0.287	**2.354**	**5.050**
+ **JointREPA**	22B	**0.302**	**0.288**	**0.261**	**0.440**	0.169	**0.124**	0.298	**0.327**	2.256	4.852
JavisBench-mini [26]											
JavisDiT++ [27]	2.1B	0.200	0.182	0.157	0.784	**0.283**	0.170	0.316	0.429	1.538	**4.987**
+ **JointREPA**	2.1B	**0.221**	**0.204**	**0.179**	**0.782**	0.278	**0.187**	0.316	**0.447**	**1.715**	4.917
LTX-2.3 [15]	22B	0.249	0.238	0.214	**0.555**	**0.280**	0.171	**0.306**	0.412	**2.458**	**5.348**
+ **JointREPA**	22B	**0.285**	**0.271**	**0.247**	0.575	0.277	**0.186**	0.304	**0.433**	2.339	5.157
VerseBench [39]											
JavisDiT++ [27]	2.1B	0.213	0.201	0.172	**0.905**	**0.182**	0.189	**0.325**	0.479	1.580	**4.280**
+ **JointREPA**	2.1B	**0.237**	**0.223**	**0.195**	0.958	0.175	**0.216**	0.323	**0.489**	**1.850**	4.260
LTX-2.3 [15]	22B	0.332	0.317	0.287	0.284	**0.167**	0.191	0.317	0.447	**2.760**	**5.004**
+ **JointREPA**	22B	**0.369**	**0.351**	**0.322**	**0.241**	0.165	**0.211**	0.317	**0.472**	2.677	4.809

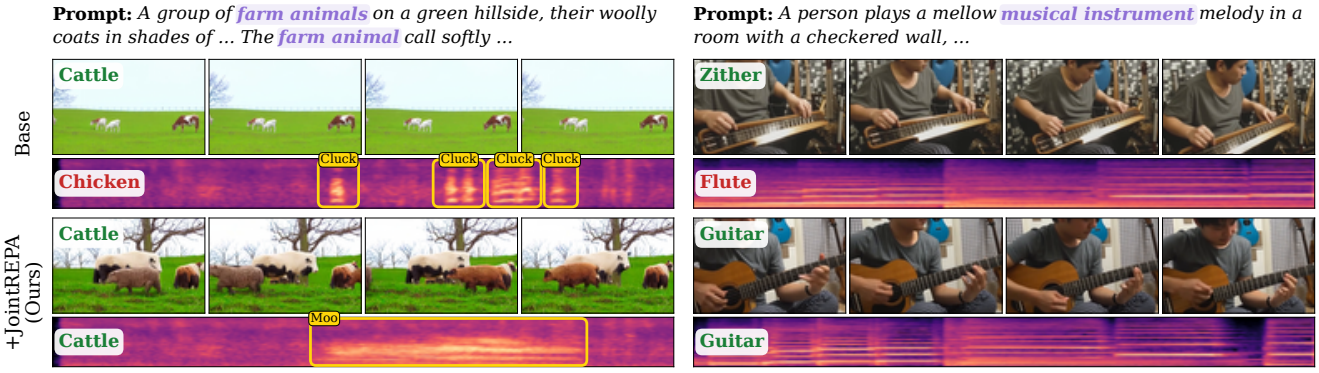


Figure 6. **Qualitative results on UNSAIDBench.** Generated video frames and audio spectrograms from the base model (top) and its JointREPA variant (bottom). Given an *underspecified shared factor* per prompt, the base model realizes different semantic identities in audio and video, whereas JointREPA converges on the same identity.

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{FM}} + \lambda_{\text{uni}}\mathcal{L}_{\text{uni}} + \lambda_{\text{cross}}\mathcal{L}_{\text{cross}}, \tag{4}$$

where $\mathcal{L}_{\text{FM}}$ is the flow-matching loss, and λ_{uni} and λ_{cross} weight the alignment terms.

5. Experiments

5.1. Experimental Setup

Datasets. For training, we use a balanced subset of VGGSound [4] across 52 classes. We exclude the categories used in Sec. 3.2 to ensure a fair evaluation when revisiting the analysis in Sec. 5.3. Existing benchmarks provide detailed semantic information, making it difficult to evaluate semantic jointness [2, 19, 26]. To this end, we introduce **UNSAIDBench** (**UN**derspecified **S**emantics in **A**udio-v**ID**eo generation). Moving beyond the simple diagnostic prompts in Tab. 1, we adapt the complex, real-world prompts of JavisBench [26] to inject one underspecified shared factor per prompt. Specifically, following the AudioSet ontology [11], we abstract fine-grained event identities into their parent categories while keeping the surrounding context intact, yielding 524 instances that remain natural in context. To assess the generalizability of JointREPA, we additionally evaluate it on two standard benchmarks, JavisBench-mini and VerseBench [39]. Since Verse-Bench uses both text and image conditioning, which JavisDiT++ does not support, we evaluate all models using only its text prompts.

Evaluation metrics. Following prior works [45, 49], we evaluate four complementary dimensions: (1) *AV-Consistency*, measuring cross-modal semantic coherence using AV-IB [13] and AVHScore [30]; (2) *AV-Synchronization*, evaluating temporal alignment using JavisScore [26] and DeSync [6, 22]. (3) *AV-Quality*, assessing the perceptual quality of each modality independently using Visual Quality (VQ) from VideoAlign [25] and Audio Quality (AQ) from Audiobox-Aesthetics [37]; (4) *Text-Consistency*, assessing text-to-modality alignment using ImageBind [13], CLIP [34], and CLAP [9].

Implementation details. We select one representative model from each model family, JavisDiT++ and LTX-2.3, as our backbones. We apply JointREPA to both models using LoRA [18] with a rank of 32. We train for one epoch using AdamW with a learning rate of 10^{-4}. We use effective batch sizes of 16 for JavisDiT++ and 4 for LTX-2.3. For the similarity function in Eq. (3), we employ token-relation distillation for unimodal video alignment following VideoREPA [52], while using cosine similarity for the remaining terms. We apply JointREPA at the middle Transformer layer ($\ell = 15$ out of 30 layers for JavisDiT++ and $\ell = 24$ out of 48 layers for LTX-2.3), with $\lambda_{\text{uni}} = \lambda_{\text{cross}} = 0.5$.

5.2. Main Results

Quantitative results. Tab. 2 summarizes the quantitative results across UNSAIDBench, JavisBench, and Verse-Bench. Across both JavisDiT++ and LTX-2.3, JointREPA consistently improves AV-Consistency metrics over the respective base models. The gains are most pronounced on UNSAIDBench, suggesting that JointREPA effectively guides both modalities to converge on the same identity among the multiple valid realizations. Notably, these improvements also extend to JavisBench and Verse-Bench, highlighting the general effectiveness of JointREPA in enhancing semantic consistency across diverse prompting conditions. Furthermore, JointREPA improves AV-Synchrony metrics in most settings, while yielding mixed results in Text-Consistency and AV-Quality. These results suggest that the benefits of JointREPA primarily manifest in cross-modal coordination.

Qualitative results. Fig. 6 presents qualitative results on UNSAIDBench. In the animal vocalization example, the base model generates a video of cattle accompanied by chicken clucking sounds. Similarly, in the musical instrument example, the video depicts a zither being played, while the audio contains flute sounds. Although each modality independently satisfies the underspecified prompt, the generated pairs are semantically inconsistent. In contrast, JointREPA generates matching cattle vocalizations and guitar sounds for the corresponding visual events, yielding semantically coherent audio–video pairs. Additional samples are provided in the supplementary material.

Table 3. **Ablations of fine-tuning losses on JavisDiT++.** While the FM loss yields modest improvements and unimodal REPA provides larger gains, JointREPA achieves the best performance.

Model	AV-Consistency		AV-Synchrony	
	AV-IB↑	AVHScore↑	JavisScore↑	DeSync↓
UNSAIDBench				
JavisDiT++ [27]	0.176	0.163	0.133	<u>0.815</u>
+ FM loss	0.181	0.171	0.141	0.826
+ Unimodal REPA	<u>0.199</u>	<u>0.188</u>	<u>0.158</u>	0.832
+ **JointREPA** (Ours)	**0.207**	**0.196**	**0.164**	**0.804**
+ BEATs JointREPA	0.213	0.201	0.170	0.818
JavisBench-mini				
JavisDiT++ [27]	0.200	0.182	0.157	0.784
+ FM loss	0.204	0.187	0.160	**0.780**
+ Unimodal REPA	<u>0.215</u>	<u>0.199</u>	<u>0.174</u>	0.787
+ **JointREPA** (Ours)	**0.221**	**0.204**	**0.179**	<u>0.782</u>
+ BEATs JointREPA	0.221	0.204	0.179	0.772

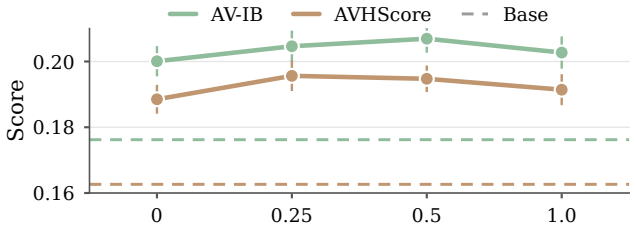


Table 4. **Ablation on the cross-modal alignment weight λ_{cross}.** All non-zero weights consistently improve over $\lambda_{\text{cross}} = 0$, demonstrating robustness to λ_{cross}.

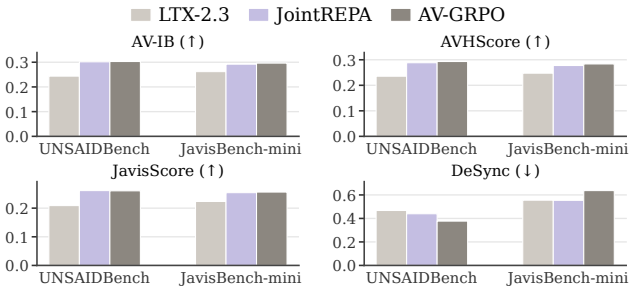
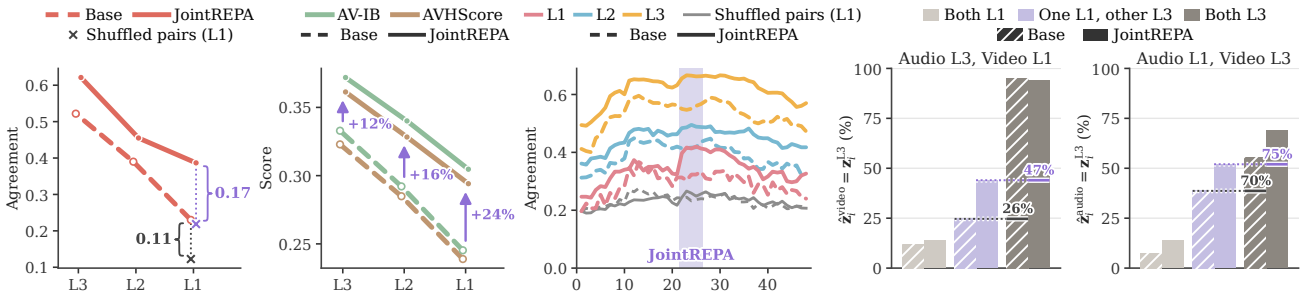


Table 5. **Comparison of JointREPA and AV-GRPO [45] on LTX-2.3.** Despite relying on lightweight LoRA-based representation alignment rather than full-model reinforcement learning, JointREPA achieves comparable results with AV-GRPO.

5.3. Additional Analysis

Ablations on fine-tuning loss. In Tab. 3, we conduct an ablation study of JointREPA on JavisDiT++ across UNSAIDBench and JavisBench-mini. Fine-tuning with the flow-matching objective alone yields only marginal improvements. While applying unimodal alignment clearly improves AV-Consistency, combining it with cross-modal alignment in JointREPA achieves the best overall performance. To further



(a) Q1. Prompt specificity (b) Q2. Internal probe (c) Q3. Asymmetric prompting

Figure 7. **Revisiting Q1, Q2 and Q3 after applying JointREPA.** (a) Q1. JointREPA improves all jointness metrics across specificity levels, peaking at L1—widening the Agreement margin over the shuffled baseline ($0.11 \rightarrow 0.17$) and achieving larger average gains in AV-IB and AVHScore at lower specificity (L3 $<$ L2 $<$ L1). (b) Q2. Intermediate representation probing reveals a pronounced increase in L1 Agreement around the middle layers where JointREPA is applied. (c) JointREPA increases the follow rate under asymmetric prompting (relative to Both-L3) from 26% to 47% and from 70% to 75%.

examine the impact of the target representation, we substitute the default CLAP audio encoder with BEATs [5]. The resulting performance remains comparable, indicating that JointREPA is agnostic to different types of audio encoders.

Ablations on hyperparameter selection. To examine the effect of cross-modal alignment, we train JavisDiT++ with varying λ_{cross}. We report AV-Consistency on UNSAIDBench in Tab. 4. Specifically, we vary $\lambda_{\text{cross}} \in \{0, 0.25, 0.5, 1.0\}$ while fixing $\lambda_{\text{uni}} = 0.5$. Removing the cross-modal alignment term ($\lambda_{\text{cross}} = 0$) yields the weakest performance among the tested settings. In contrast, all non-zero weights yield consistently higher performance, with $\lambda_{\text{cross}} = 0.5$ achieving the best average performance across the two metrics. This indicates that JointREPA is robust to the selection of the cross-modal alignment weight λ_{cross}.

Comparison with other post-training methods. To assess the effectiveness of JointREPA relative to existing post-training approaches, we compare it with AV-GRPO [45] on both UNSAIDBench and JavisBench-mini. AV-GRPO applies Group Relative Policy Optimization [14] using modality quality, prompt alignment, and audio–video synchronization scores as rewards. We use its publicly released, fully fine-tuned LTX-2.3 checkpoint. As shown in Tab. 5, even without explicit reward definitions or full-model fine-tuning, JointREPA achieves comparable AV-Consistency and AV-Synchrony on both benchmarks.

Revisiting Q1, Q2, and Q3. We repeat the analyses of Q1, Q2, and Q3 from Sec. 3.3 after applying JointREPA to LTX-2.3. For Q1 (Fig. 7(a)), JointREPA improves all semantic consistency metrics across specificity levels with the largest gain at L1. Also, it widens the Agreement margin over the shuffled baseline from 0.11 to 0.17. Intermediate representation probing for Q2 (Fig. 7(b)) shows a consistent trend, with a pronounced increase in L1 Agreement around the middle layers where JointREPA is applied. For asymmetric prompting in Q3 (Fig. 7(c)), JointREPA improves the follow rates in both directions. Specifically, the follow rates relative to Both-L3 increase from 26% to 47% for video following audio and from 70% to 75% for audio following video. The slight increases observed in the Both-L1 and Both-L3 conditions stem from stronger overall semantic representations induced by unimodal alignment. Together, these results demonstrate that JointREPA effectively promotes genuine cross-modal semantic coordination.

6. Conclusion

In this work, we rethink joint audio-video generation networks through the lens of *intrinsic jointness*, to look beyond the seemingly strong audio–visual consistency reported under standard evaluations. Using semantically underspecified prompts which leave multiple valid events possible, we show that recent T2AV models largely fail to coordinate audio and video toward a shared semantic realization. To address this, we propose *JointREPA*, a lightweight post-training method that aligns internal representations both within and across modalities. In addition, to rigorously evaluate semantic jointness, we introduce *UNSAIDBench*, which—unlike existing benchmarks with detailed prompts—abstracts fine-grained event identities into general categories. Applying JointREPA to two models with distinct architectures and model sizes yields consistent improvements in semantic consistency across three benchmarks, including UNSAIDBench. Although our study of intrinsic jointness focuses on explicit semantic identities, extending it to physical attributes (e.g., material and weight) offers a promising avenue for physically-coherent T2AV generation. Looking ahead, we envision *intrinsic jointness* as a core design principle for T2AV models, moving beyond mere co-generation toward truly unified audio–video synthesis.

References

[1] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. 5

[2] Zhe Cao, Tao Wang, Jiaming Wang, Yanghai Wang, Yuanxing Zhang, Jiahao Wang, Jialu Chen, Miao Deng, Chenxi Liao, Yize Zhang, Yubin Guo, Zhaoxiang Zhang, and Jiaheng Liu. T2AV-Compass: Towards unified evaluation for text-to-audio-video generation. In *ICML*, 2026. 2, 3, 6

[3] Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. VideoJAM: Joint appearance-motion representations for enhanced motion generation in video models. In *ICML*, 2025. 1

[4] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 6

[5] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *ICML*, 2023. 8

[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 *CVPR*, 2025. 7

[7] Jiale Cheng, Ruiliang Lyu, Xiaotao Gu, Xiao Liu, Jiazheng Xu, Yida Lu, Jiayan Teng, Zhuoyi Yang, Yuxiao Dong, Jie Tang, et al. VPO: Aligning text-to-video generation models with prompt optimization. In *ICCV*, 2025. 1, 2

[8] Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. In *ICLR*, 2026. 1

[9] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *ICASSP*, 2023. 5, 7

[10] Jess Gallegos, Thomas Iljic, and Google DeepMind. Introducing Veo 3.1 and advanced capabilities in Flow. https://blog.google/technology/ai/veo-updates-flow/, 2025. Google Blog, October 15, 2025. Technical details inherited from the Veo 3 Tech Report, https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf. 1, 2

[11] 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 *ICASSP*, 2017. 3, 6

[12] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *EMNLP*, 2023. 4

[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 *CVPR*, 2023. 4, 7

[14] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 8

[15] 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. 1, 2, 3, 6

[16] Meera Hahn, Wenjun Zeng, Nithish Kannen, Rich Galt, Kartikeya Badola, Been Kim, and Zi Wang. Proactive agents for multi-turn text-to-image generation under uncertainty. In *ICML*, 2025. 1, 2

[17] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *CVPR*, 2025. 1

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 7

[19] Daili Hua, Xizhi Wang, Bohan Zeng, Xinyi Huang, Hao Liang, Junbo Niu, Xinlong Chen, Quanqing Xu, and Wentao Zhang. VABench: A comprehensive benchmark for audio-video generation. In *CVPR*, 2026. 2, 3, 6

[20] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *NeurIPS*, 2025. 1

[21] Oh Hyun-Bin, Yuhta Takida, Toshimitsu Uesaka, Tae-Hyun Oh, and Yuki Mitsufuji. PAVAS: Physics-aware video-to-audio synthesis. In *CVPR*, 2026. 1

[22] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP*, 2024. 7

[23] Team Kandinsky, Julia Agafonova, Bulat Akhmatov, Mikhail Aksyutin, Grigorii Alekseenko, Anastasia Aliaskina, Olga Androsova, Vladimir Arkhipkin, Anna Averchenkova, Alexander Belykh, et al. Kandinsky 6.0 Video: Foundation models for synchronized video and audio generation. *arXiv preprint arXiv:2610.05608*, 2026. 1, 2

[24] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023. 2, 3

[25] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. In *NeurIPS*, 2025. 7

[26] Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Scofield Fei, et al. JavisDiT: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. In *ICLR*, 2026. 1, 2, 3, 6, 7

[27] 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. In *ICLR*, 2026. 1, 2, 3, 6, 7

[28] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. FETV: A benchmark for fine-grained evaluation of open-domain text-to-video generation. In *NeurIPS*, 2023. 2

[29] Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation. *arXiv preprint arXiv:2510.01284*, 2025. 1, 2

[30] Yuxin Mao, Xuyang Shen, Jing Zhang, Zhen Qin, Jinxing Zhou, Mochu Xiang, Yiran Zhong, and Yuchao Dai. TAVGBench: Benchmarking text to audible-video generation. In *ACM MM*, 2024. 4, 7

[31] Ninareh Mehrabi, Palash Goyal, Apurv Verma, Jwala Dhamala, Varun Kumar, Qian Hu, Kai-Wei Chang, Richard Zemel, Aram Galstyan, and Rahul Gupta. Resolving ambiguities in text-to-image generative models. In *ACL*, 2023. 2

[32] OpenAI. Sora 2, 2025. Accessed: 2026-11-05. 1, 2

[33] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 2

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 7

[35] Team Seedance, Heyi Chen, Siyan Chen, Xin Chen, Yanfei Chen, Ying Chen, Zhuo Chen, Feng Cheng, Tianheng Cheng, Xinqi Cheng, et al. Seedance 1.5 pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 2025. 1, 2

[36] 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. 1, 2

[37] Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, et al. Meta Audiobox Aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025. 7

[38] Tri Ton, Ji Woo Hong, and Chang D. Yoo. TARO: Timestep-adaptive representation alignment with onset-aware conditioning for synchronized video-to-audio synthesis. In *ICCV*, 2025. 2

[39] 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. 1, 2, 6

[40] Jun Wang, Chunyu Qiang, Yuxin Guo, Yiran Wang, Xijuan Zeng, and Feng Deng. Apollo: Unified multi-task audio-video joint generation. *arXiv preprint arXiv:2601.04151*, 2026. 1, 2

[41] 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. 4

[42] Ziyi Wu, Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Yuwei Fang, Varnith Chordia, Igor Gilitschenski, and Sergey Tulyakov. Mind the Time: Temporally-controlled multi-event video generation. In *CVPR*, 2025. 1, 3

[43] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *ICLR*, 2025. 1

[44] Zeyu Xie, Xuenan Xu, Zhizheng Wu, and Mengyue Wu. PicoAudio: Enabling precise timestamp and frequency controllability of audio events in text-to-audio generation. *arXiv preprint arXiv:2407.02869*, 2024. 1, 3

[45] Zhiyu Xu, Weilong Yan, Yufei Shi, Shiyang Li, Yihao Liu, Kin-Man Lam, and Yuewen Cao. AV-GRPO: Modality-anchored decoupling diffusion reinforcement learning for joint audio-video generation. *arXiv preprint arXiv:2609.29816*, 2026. 7, 8

[46] Tian Ye, Song Fei, and Lei Zhu. UltraFlux: Data-model co-design for high-quality native 4k text-to-image generation across diverse aspect ratios. In *CVPR*, 2026. 1

[47] 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 *ICLR*, 2025. 2, 5

[48] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 3

[49] Guohui Zhang, XiaoXiao Ma, Jie Huang, Hang Xu, Hu Yu, Siming Fu, Yuming Li, Zeyue Xue, Lin Song, Haoyang Huang, et al. OmniNFT: Modality-wise omni diffusion reinforcement for joint audio-video generation. *arXiv preprint arXiv:2605.12480*, 2026. 7

[50] Kang Zhang, Trung X. Pham, Suyeon Lee, Axi Niu, Arda Senocak, and Joon Son Chung. Model-guided dual-role alignment for high-fidelity open-domain video-to-audio generation. In *NeurIPS*, 2025. 2

[51] Pengfei ZHANG, Tianxin Xie, Minghao Yang, and Li Liu. AG-REPA: Causal layer selection for representation alignment in audio flow matching. In *ICML*, 2026. 2

[52] Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. VideoREPA: Learning physics for video generation through relational alignment with foundation models. In *NeurIPS*, 2025. 2, 7

[53] Xiaofeng Zhang, Aaron Courville, Michal Drozdzal, and Adriana Romero-Soriano. The intricate dance of prompt complexity, quality, diversity and consistency in T2I models. In *ICLR*, 2026. 2

[54] Jinxing Zhou, Dan Guo, Ruohao Guo, Yuxin Mao, Jingjing Hu, Yiran Zhong, Xiaojun Chang, and Meng Wang. Towards open-vocabulary audio-visual event localization. In *CVPR*, 2025. 3