SINGDANCE: COMPOSITIONAL ZERO-SHOT SINGING-AND-DANCING VIDEO GENERATION WITH ROLE-AWARE AUDIO CONDITIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating personalized dance videos from a reference image, text prompt, and audio track requires music-conditioned body motion. Singing-and-dancing adds a second requirement: the visible subject must also articulate the vocals. Existing music-conditioned methods focus primarily on choreography, while speech-driven models generally assume that the visible subject produces the input voice, leaving this combined setting largely underexplored. We introduce SingDance, a unified video diffusion framework that formulates controllable vocal articulation as a semantic role: the visible subject is either the *source*, who produces the vocal signal, or the *listener*, who receives it from an off-screen performer. Task-defined hard-compact routing selects task-relevant speech, music, and role conditions, which are composed through frame-wise joint audio injection; source and listener retain the same speech pathway. Training uses asymmetric supervision: on-screen speaking and curated off-screen conversational-response videos establish role control, while instrumental and song-based dancing-only videos establish music-conditioned body motion. The target Song/Source configuration is held out from SingDance task training. At inference, assigning the source role to a song composes separately learned articulation and song-conditioned dance capabilities, enabling compositional zero-shot singing-and-dancing. Experiments demonstrate strong motion–beat alignment and visual fidelity, reliable paired switching of vocal articulation while preserving music-aligned body motion, and highly competitive lip synchronization with substantially fewer generation-time parameters than the strongest speech-driven baseline evaluated. Video samples are available at `https://anonymous4submit-pages.github.io/submission-demo/`.

1 INTRODUCTION

Short-form dance videos are a major form of online visual content. They broadly take two forms: *dancing-only*, where an on-screen subject moves to instrumental music, or to a song without lip-syncing to its vocals, and *singing-and-dancing*, where dance motion and vocal articulation occur together. Given the same subject and song, creators may wish to produce either form. A single framework supporting both from a reference image, text prompt, and audio track would substantially simplify personalized video creation. Despite its practical relevance, singing-and-dancing remains largely unexplored in music-conditioned human video generation.

Existing research addresses adjacent parts of this problem under different assumptions. Music-conditioned dance methods focus on body choreography and rhythmic alignment, and generally do not model vocal articulation by the dancer Chen et al. (2025c); Hong et al. (2026); Yang et al. (2026a); Huang et al. (2026). Speech-driven human video models achieve accurate speech lip synchronization and expressive motion, but conventionally assume that the input voice is produced by the visible subject Gao et al. (2025); Gan et al. (2025); Yang et al. (2025b); Meng et al. (2026). Recent interactive avatar systems model speaking and listening as distinct conversational phases, often with separated or phase-specific audio pathways Zeng et al. (2026); Sun et al. (2026). Song-conditioned dance raises a different question: how to control whether the visible subject produces the vocals while retaining the same vocal content as part of the conditioning for body motion.

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

Our key insight is to model controllable vocal articulation as an explicit semantic role rather than an isolated mouth-motion instruction. We define the visible subject as either the *source*, who produces the vocal signal, or the *listener*, who receives it from an off-screen performer. This distinction prevents the presence of vocal content from automatically assigning articulation to the visible subject. Genuine listening/reacting examples provide non-speaking supervision for the listener role, rather than defining Lip-Sync OFF as a frozen-mouth instruction. This semantic distinction transfers naturally from speech to songs, organizing four behaviors—speaking, listening, dancing-only, and singing-and-dancing—within a single generation problem.

We instantiate this formulation in SingDance, a video diffusion transformer conditioned on a reference image, text, representations specialized for speech and music, and an explicit vocal role. Task-defined hard-compact routing retains the task-relevant token groups: speech interaction uses speech and role tokens, instrumental dancing-only uses music and role tokens, and song-based dancing-only uses all three. Importantly, the two vocal roles share the same speech pathway; compact role conditions work with the prompt's *vocal role* field to specify whether the visible subject is assigned the vocals. For song-based dancing-only, vocal features remain available because lyrics and vocal phrasing may still inform the performance even when they are not articulated by the visible subject.

Training follows a staged design with asymmetric supervision. We first learn role control from on-screen speaking clips and carefully curated conversational clips in which the visible subject genuinely listens or reacts to an off-screen interlocutor. We then learn music-conditioned dance from instrumental and song-based dancing-only videos, without singing-and-dancing supervision. At inference, singing-and-dancing reuses the complete speech-plus-music configuration learned from song-based dancing-only and changes only the vocal-role state from listener to source. This state is expressed consistently through the learned role conditions and the prompt's *vocal role* field. This compact intervention enables paired role switching and instantiates the held-out Song/Source configuration through compositional zero-shot inference.

Our contributions are threefold:

- We formulate dance video generation with an explicit vocal role, organizing speaking, listening, dancing-only, and singing-and-dancing as four related generation behaviors.
- We introduce a role-aware conditioning design in which source and listener share the same speech representation and injection pathway, while task-routed speech, music, and role tokens are composed through frame-wise joint audio injection.
- We learn source-role articulation and song-conditioned dance from asymmetric supervision and demonstrate their compositional zero-shot generalization to the held-out Song/Source configuration. Paired role intervention verifies controllable articulation, while motion–beat evaluation and music-conditioning ablation validate music-responsive body motion.

2 Related Work

Music-conditioned dance generation. Music-conditioned dance generation is inherently one-to-many, and prior work adopts different intermediate representations to structure this ambiguity. Early methods generate 3D body motion before rendering, including FACT on the AIST++ dataset Li et al. (2021), Bailando Siyao et al. (2022), EDGE Tseng et al. (2022), and FineDance Li et al. (2023). More recent systems move toward RGB video while retaining explicit motion intermediates: X-Dancer first generates music-aligned 2D skeleton motion, represented as stick-figure pose maps, and then uses that motion with a reference image to synthesize RGB video in a second stage Chen et al. (2025c), whereas MACE-Dance cascades motion and appearance experts Yang et al. (2025a). A complementary line directly adapts video diffusion models. MusicInfuser introduces music conditioning into a text-to-video backbone Hong et al. (2026). Concurrent studies further explore different allocations between text and music: OmniDance combines image, text, and MERT features in a 5B video backbone Yang et al. (2026a), while the more recent Wan-Dancer uses coarse textual attributes and hierarchical global-to-local music conditioning for long-form generation Huang et al. (2026). These concurrent methods explore different allocations of text and music for choreographic control. SingDance studies an additional compositional axis: whether the visible dancer produces the vocals while retaining music-conditioned body motion.

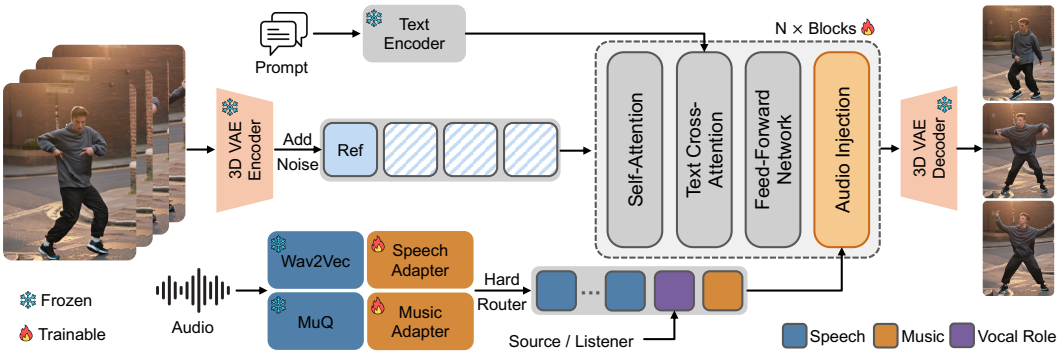


Figure 1: Training architecture of SingDance. The first video frame is separately encoded as a clean reference latent, while subsequent frames are temporally compressed and noised in latent space. Wav2Vec 2.0 and MuQ features, together with the vocal-role condition, are selected by hard-compact routing and supplied to frame-wise joint audio injection. The repeated block is schematic: joint audio injection is applied immediately after each of 10 selected blocks in the 30-block DiT backbone. The denoised latents are decoded into output frames. Snowflakes and flames denote frozen and trainable modules, respectively.

Speech-driven human video generation. Recent diffusion-based speech-driven human video models have substantially improved identity preservation, speech lip synchronization, and natural co-speech motion Chen et al. (2025b); Gan et al. (2025); Gao et al. (2025); Yang et al. (2025b); Meng et al. (2026); Meituan LongCat Team (2026). Most systems, however, assume that the visible subject produces the input utterance and rely on pretrained speech encoders, such as Wav2Vec or Whisper, for temporally aligned conditioning Baevski et al. (2020); Radford et al. (2023). Although these representations capture phonetic and prosodic cues for articulation, they are not explicitly optimized for music-specific structures such as beat, timbre, and longer-range rhythm. Interactive avatar and 3D facial animation systems instead model speaking and listening through dual-speaker context, listener-specific motion priors, or phase-specific audio pathways Peng et al. (2025); Chu et al. (2026); Wang et al. (2026); Weng et al. (2026); Zhang et al. (2026). DualTalk models two-speaker interaction Peng et al. (2025); UniLS separates an internal listener-motion prior from external audio modulation Chu et al. (2026); LPM 1.0 uses independently parameterized speaking and listening streams Zeng et al. (2026); and StreamAvatar routes speech features through dedicated Talk and Listen attention modules Sun et al. (2026). These designs account for conversational motion distributions. SingDance instead keeps the speech representation and conditioning pathway shared across roles, changing textual and learned role conditions. This permits held-out Song/Source composition while retaining the speech and music conditions used for song-based dancing-only.

Multimodal conditioning in video foundation models. Pretrained video diffusion models provide reusable foundations that can be extended with additional control modalities. Wan-S2V combines text-based global motion specification with frame-aligned speech injection Gao et al. (2025). ActAvatar builds upon Wan2.2-TI2V-5B and combines temporally structured text conditioning for phase-level actions with Wav2Vec 2.0-based speech cross-attention for articulation Peng et al. (2026). For music-conditioned generation, OmniDance augments the same TI2V backbone with MERT-based music cross-attention, whereas Wan-Dancer adapts Wan-I2V using compact Librosa features in a hierarchical generation pipeline Yang et al. (2026a); Huang et al. (2026). These approaches illustrate different ways of assigning control responsibilities to text and audio while preserving the generative prior of a pretrained video model. SingDance assigns each modality a primary responsibility according to the information it represents most naturally: the reference image anchors appearance and the initial scene; text specifies intended action and camera evolution; speech captures vocal articulation and phrasing cues; and music complements the prompt with rhythm and broader musical structure that are difficult to express textually. This allocation motivates separate speech and music representations, composed through frame-wise joint audio injection, while explicit vocal-role control determines whether the visible subject articulates the vocals.

3 METHOD

In this section, we first introduce the audio representations and vocal-role control, then describe hard-compact routing and joint audio injection, and finally present staged training for zero-shot composition and text–audio classifier-free guidance.

3.1 AUDIO REPRESENTATIONS AND VOCAL-ROLE CONTROL

To coordinate vocal articulation with music-conditioned body motion within a single generation framework, SingDance employs representations specialized for speech and music together with coordinated textual and learned vocal-role controls.

As illustrated in Figure 1, we use Wav2Vec 2.0 Baevski et al. (2020) to extract speech-oriented features for vocal articulation. Following Wan-S2V Gao et al. (2025), we aggregate its layer-wise representations using learnable weights and map them through a temporal encoder to a global speech feature and frame-aligned local speech tokens. In parallel, MuQ Zhu et al. (2025) provides music-oriented acoustic and structural features, which a separate temporal encoder maps to frame-aligned local music tokens for body-motion conditioning. Both temporal encoders use non-causal 1D convolutions, matching the non-causal context of the pretrained features.

At the level of generated behavior, the two vocal roles correspond to the Lip-Sync ON and OFF states in Figure 2. Semantically, *source* indicates that the vocal signal is produced by the visible subject, whereas *listener* indicates that it originates from an external source. We treat lip synchronization as an outcome of vocal-role assignment rather than an isolated mouth-motion instruction.

To expose this distinction to the text encoder while separating vocal behavior from choreography, we organize each prompt into five fields: *vocal role*, *style*, *action*, *first frame*, and *camera*. The standardized *vocal role* field assigns the vocal signal either to the visible subject or to an external source, corresponding to *source* for Lip-Sync ON and *listener* for Lip-Sync OFF. During prompt construction, we remove references to lip movement, speaking, and singing from the *action* field, retaining only body-motion descriptions. Within the text prompt, vocal behavior is therefore specified exclusively by the *vocal role* field, separating its textual control from the intended body-action sequence.

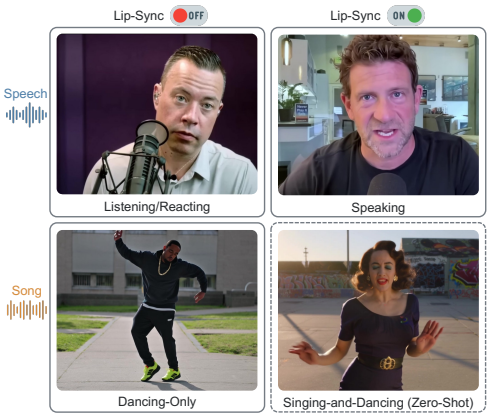


Figure 2: Four generation tasks supported by SingDance. The dashed box marks singing-and-dancing, the held-out Song/Source configuration generated through compositional zero-shot inference.

In parallel, the same binary source/listener state is encoded by a learned global role embedding and a local role token. Each textual role assignment is paired consistently with its corresponding learned role conditions throughout training and inference. Speech and vocal role therefore provide both global and local conditions, whereas music is introduced only through local tokens. When switching roles, SingDance changes only the vocal-role state; its textual and learned representations are updated together, while the role-independent prompt content, speech and music representations, and shared conditioning pathway remain unchanged. This coordinated but compact intervention provides the basis for the held-out composition described later.

3.2 HARD-COMPACT ROUTING AND JOINT AUDIO INJECTION

Different generation tasks require different subsets of the available conditions. We therefore use *task-defined hard-compact routing*: the route is specified by the generation task rather than inferred from audio content, and only the task-relevant speech, music, and role tokens are retained as the routed tokens. Table 1 summarizes the five configurations.

Table 1: Hard-compact routing across generation tasks. S, M, and R denote speech, music, and role tokens, respectively. The first four routes are used during training; Song/Source is activated only at inference as the held-out compositional configuration.

Audio type	Role	Routed tokens	Task
Speech	Source	$S + R_{\text{src}}$	Speaking
Speech	Listener	$S + R_{\text{lis}}$	Listening/Reacting
Instrumental	Listener	$R_{\text{lis}} + M$	Dancing-Only
Song	Listener	$S + R_{\text{lis}} + M$	Dancing-Only
Song	Source	$S + R_{\text{src}} + M$	Singing-and-Dancing

Speech examples retain speech and role tokens. Instrumental dance retains music and role tokens, making music the only time-varying audio input and providing a direct learning signal to the newly introduced music pathway. Song-based dance retains all three token groups under the listener role, because lyrics, vocal phrasing, and phrase boundaries may inform choreographic timing even when the visible subject does not articulate the vocals. At inference, the held-out Song/Source configuration preserves the speech and music tokens and changes only the vocal-role state, updating its learned role conditions and the prompt's *vocal role* field consistently.

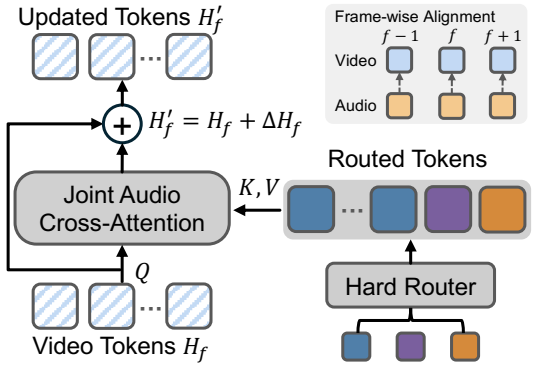


Figure 3: Frame-wise joint audio injection. For each generated latent frame, visual tokens act as queries and the temporally aligned routed tokens serve as joint keys and values; the cross-attention output is added residually to update the visual representation.

Figure 3 details frame-wise joint audio injection, where the routed tokens form a joint key–value sequence temporally aligned with each generated latent frame. For global modulation, speech and song examples use the global speech feature with the global role embedding, whereas instrumental examples use the global role embedding alone; music is supplied through the routed local tokens. Text remains in the backbone's original text cross-attention, and joint audio injection updates only generated-frame tokens, leaving the encoded reference latent unchanged.

3.3 Staged Training for Zero-Shot Composition

SingDance adopts a staged training strategy that first establishes vocal-role control and then introduces music-conditioned body motion. This organization reflects the asymmetric supervision available to the two capabilities: speech videos provide source/listener supervision, whereas dancing-only videos provide music-conditioned body-motion supervision. Importantly, no paired singing-and-dancing videos are used in either stage.

In the speech stage, the music pathway is absent. On-screen vocal production grounds the *source* role, while genuine listening and reacting behavior toward an off-screen interlocutor grounds the *listener* role. Both roles share the same speech representation and conditioning pathway, with their assignment expressed through the prompt's *vocal role* field and global/local role conditions. The model therefore learns role-dependent interpretations of a shared vocal signal from these complementary speech training examples before music conditioning is introduced.

The dance stage starts from the speech-stage model and introduces the music pathway using instrumental and song-based dancing-only videos. Instrumental examples retain only music and role conditions, making music the sole time-varying audio input. Song examples retain speech and music conditions because vocal phrasing and phrase boundaries may remain relevant to choreography even without visible vocal articulation. During this stage, speech examples from both roles constitute 20% of the training samples to mitigate forgetting of articulation and role control.

Across the two stages, SingDance observes four configurations: Speech/Source, Speech/Listener, Instrumental/Listener, and Song/Listener. The target Song/Source configuration is held out from SingDance task training. At inference, we instantiate it by changing the *vocal role* field from an off-screen to an on-screen source assignment and switching the corresponding global and local role conditions from listener to source. The reference image, soundtrack, remaining prompt fields, and extracted speech and music features remain unchanged. This composition combines the source-role articulation learned from speech with the joint speech–music conditioning learned from dancing-only videos, without introducing a new generation module or a post-hoc lip-synchronization model. Both stages use the backbone's standard flow-matching objective Lipman et al. (2023) over generated latent positions, without specialized lip-synchronization or beat-alignment losses.

3.4 TEXT–AUDIO CLASSIFIER-FREE GUIDANCE

At inference, we use three-pass classifier-free guidance (CFG) Ho & Salimans (2022) with separate text and audio guidance scales. Let $\mathbf{v}_{\bar{t},0}$, $\mathbf{v}_{t,0}$, and $\mathbf{v}_{t,a}$ denote the predictions obtained with the negative prompt and null audio, the input prompt and null audio, and the input prompt with routed audio, respectively. We compute

$$\widehat{\mathbf{v}} = \mathbf{v}_{\bar{t},0} + s_t \left(\mathbf{v}_{t,0} - \mathbf{v}_{\bar{t},0}\right) + s_a \left(\mathbf{v}_{t,a} - \mathbf{v}_{t,0}\right), \tag{1}$$

where s_t and s_a are the text and audio guidance scales, respectively. Both the learned global role embedding and the local role token remain active in all three passes, including the null-audio passes. The textual *vocal role* field is part of the input prompt and is therefore included in $\mathbf{v}_{t,0}$ and $\mathbf{v}_{t,a}$.

To enable this three-pass CFG, we apply condition dropout during training: text alone is dropped with probability 10%, audio alone with probability 10%, and both are dropped with probability 5%; all remaining samples retain both conditions. Text dropout removes the entire prompt, including its *vocal role* field, whereas the learned vocal-role embeddings are retained in all dropout cases.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Data. Speech training combines HuMoSet Chen et al. (2025a) with proprietary speech videos, and dance training combines cleaned MA-Data Yang et al. (2025a) with proprietary dance videos. The dance data include instrumental clips and songs; potentially lip-synchronized song clips are conservatively removed, so retained songs supervise Song/Listener while Song/Source remains held out. No singing-and-dancing videos are used in SingDance task training. We evaluate on *SingDance-Test*, *Dance-Only-Test*, and EMTD Meng et al. (2025). The task-specific test sets are disjoint from training, and the paired vocal-role intervention fixes each case's inputs and inference settings except for the role assignment. Dataset composition and filtering are detailed in Appendix C.

Baselines. Wan-S2V Gao et al. (2025) accepts the same text, reference-image, and audio inputs as SingDance and provides an input-matched speech-driven comparison on both test sets. For dancing-only, we also evaluate two complementary music-driven methods: MusicInfuser Hong et al. (2026) generates video directly from text and music, whereas X-Dancer Chen et al. (2025c) predicts skeleton motion before rendering RGB video from a reference image. These methods use their supported input interfaces; Appendix E details the comparison scope.

Implementation Details. All SingDance results are produced by a single model initialized from Wan2.2-TI2V-5B Wan Team et al. (2025). We train on 64 NVIDIA A100 GPUs with an effective batch size of 64, using AdamW Loshchilov & Hutter (2019) with a learning rate of 1×10^{-5}. The speech and dance phases are each trained for eight epochs at 480p, after which the dance phase is further fine-tuned for four epochs, approximately 3K optimization steps, at 704×1280. At inference, talking videos follow the corresponding EMTD input resolution, whereas dance videos are generated at 704×1280. We generate 121 frames at 24 FPS using 50 flow-matching sampling steps. Talking uses two-pass CFG with scale 5, while the two dance tasks use the three-pass CFG in Equation 1, with text and audio scales of 5 and 4, respectively.

Table 2: Quantitative results on *SingDance-Test* and *Dance-Only-Test*. Human scores are SingDance-versus-Wan-S2V preference rates (%). Best and second-best generated results are bold and underlined; GT is shown only as a reference. † MusicInfuser does not accept a reference image, so its reference-fidelity scores are not applicable.

	Reference Fidelity		Motion–Beat Alignment		Visual Quality			Human Preference	
Method	Subject ↑	Background ↑	BeatAlign ↑	MBCR ↑	Aesthetic ↑	Imaging ↑	Smoothness ↑	Rhythm ↑	Visual ↑
Singing-and-Dancing									
GT	0.9673	0.9735	0.2795	0.2922	0.5720	0.6553	0.9858	–	–
Wan-S2V	0.9565	0.9633	0.2640	0.1983	0.5834	0.6538	0.9860	38.67	45.33
SingDance	**0.9682**	**0.9772**	**0.2723**	**0.2558**	**0.5846**	**0.6688**	**0.9899**	**61.33**	**54.67**
Dancing-Only									
GT	0.9648	0.9705	0.3051	0.3027	0.5738	0.6667	0.9880	–	–
MusicInfuser†	–	–	0.2872	**0.2661**	0.5258	0.6522	**0.9948**	–	–
Wan-S2V	0.9415	0.9493	0.2467	0.2180	0.5776	0.6570	0.9867	32.00	44.33
X-Dancer	0.9697	0.9725	0.2583	0.2330	**0.5997**	0.6378	0.9925	–	–
SingDance	**0.9705**	**0.9766**	**0.3002**	0.2632	0.5718	**0.6738**	0.9921	**68.00**	**55.67**

Evaluation Metrics. We evaluate reference fidelity, visual quality, audio–visual lip synchronization, motion–beat alignment, and human preference.

Reference and visual quality. Following VBench Huang et al. (2024a) and its I2V extension in VBench++ Huang et al. (2024b), we report DINO-based Subject and DreamSim-based Background fidelity to each case's inference input image, interpolation-based Motion Smoothness, LAION Aesthetic Quality, and MUSIQ Imaging Quality (normalized to $[0, 1]$). Subject and Background are the VBench-I2V reference metrics, not the ordinary VBench subject/background consistency metrics.

Lip synchronization. We use the official SyncNet evaluator Chung & Zisserman (2017) to report LSE-C and LSE-D. Higher LSE-C and lower LSE-D indicate stronger synchronization for the vocal-source role; the desired directions are reversed for the listener role, where synchronization with external vocals should be suppressed. Videos without a valid face track are excluded.

Motion–beat alignment. We represent audio and video as beat events on a shared time axis. Musical beat times form $\mathcal{B}$. For video, DWPose Yang et al. (2023) tracks body, foot, and hand landmarks; we compute their scale-normalized inter-frame motion speed and take its local minima after Gaussian smoothing as the kinematic beat set $\mathcal{K}$, following the BeatAlign convention Li et al. (2021). BeatAlign measures the temporal proximity from each kinematic beat to its nearest musical beat. We complement it with an adapted Music-Beat Coverage Rate (MBCR), based on the coverage direction of TMD-Bench Yang et al. (2026b), which measures the fraction of musical beats accompanied by nearby movement. Writing $d(t, \mathcal{T}) = \min_{u\in\mathcal{T}} |t - u|$, we compute

$$\mathrm{BeatAlign} = \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}} \exp\!\left(-\frac{d(k,\mathcal{B})^2}{2\sigma^2}\right), \qquad \mathrm{MBCR} = \frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}} \mathbb{I}(d(b,\mathcal{K}) < \tau)\,,$$

where σ controls temporal proximity and τ defines a beat hit; both are fixed across methods. Motion scores are averaged over matched pose-valid inputs, while reference and visual-quality scores use each method's evaluable videos. Appendix D summarizes the validity protocol.

Human evaluation. We run a blind SingDance–Wan-S2V preference study with matched text, image, and audio. Fifteen participants judge 20 randomly sampled cases from each test set (40 pairs, 24 FPS) in balanced, randomized A/B order, choosing better motion–music rhythm and overall visual quality without ties; each task and criterion has 300 votes. MusicInfuser lacks image input and X-Dancer lacks free-form text control, so their results appear on the demo page rather than in this matched-input study.

4.2 Main Results

Table 2 evaluates compositional zero-shot singing-and-dancing on *SingDance-Test* and directly supervised dancing-only generation on *Dance-Only-Test*. It reports reference fidelity where applicable, motion–beat alignment, visual quality, and pairwise human preference between SingDance and Wan-S2V. The Source rows of Table 3 evaluate lip synchronization for both SingDance and

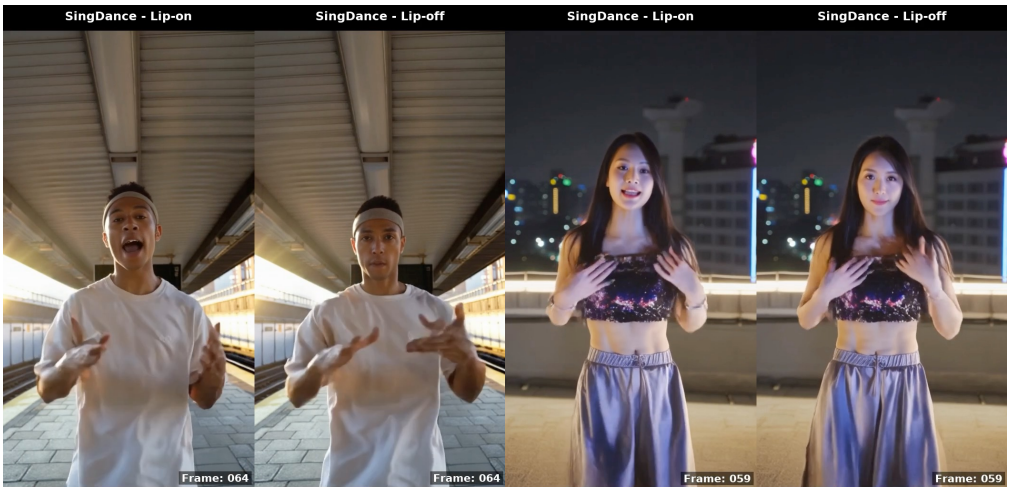


Figure 4: Qualitative paired vocal-role switching on two SingDance examples. Each Lip-on/Lip-off pair shares all role-independent inputs and inference settings; only the vocal role changes.

Wan-S2V on the same singing-and-dancing outputs, so the two tables should be read together to assess dance and vocal articulation in the same videos. BeatAlign reflects the timing precision of detected motion events, while MBCR reflects coverage of musical beats. Sparse but well-timed motion can score highly on BeatAlign while leaving many beats unanswered, so we interpret both metrics together. The routed music token is separately ablated in Table 5. Additional audio-enabled comparisons are available on the demo page linked in the abstract.

Singing-and-dancing. Without singing-and-dancing supervision in either SingDance training stage, the model generates rhythmically coordinated performances that combine vocal articulation and dance while preserving the reference appearance and visual quality. On *SingDance-Test*, SingDance scores higher than the input-matched Wan-S2V baseline in every shared Table 2 metric, including motion–beat alignment and both human preferences. The within-model ablation in Table 5 separately tests the contribution of the routed music token.

Dancing-only. On *Dance-Only-Test*, SingDance leads the generated methods in BeatAlign and Imaging, and leads the reference-conditioned methods in Subject and Background. Its BeatAlign approaches the GT score, while its MBCR is comparable to MusicInfuser's and exceeds X-Dancer's. Together, these scores support precise timing and competitive beat coverage while SingDance also supports held-out Song/Source composition. Human preference is reported against the matched-input Wan-S2V baseline. Automated visual scores do not fully capture motion richness or choreography: interpolation-based Smoothness can also be high for clips with little movement. We encourage readers to inspect the audio-enabled side-by-side videos on the demo page for motion amplitude, rhythm, and visual quality beyond the reported metrics.

4.3 Paired Vocal-Role Switching

Table 3 tests all source/listener pairs on *SingDance-Test*; its source rows directly reuse the singing-and-dancing outputs from Table 2 above. Each listener counterpart keeps exactly the same reference image, song, role-independent prompt fields, checkpoint, seed, and inference settings. Only the vocal-role state itself changes, with its learned conditions and the prompt's *vocal role* field updated. Wan-S2V receives the same textual switch as a prompt-only control.

Table 3: Paired vocal-role switching experiments on *SingDance-Test*. Source and listener correspond to Lip-Sync ON and OFF; LSE-C and LSE-D therefore have opposite desired directions.

Role	Method	LSE-C	LSE-D	BeatAlign ↑	MBCR ↑
Source	Wan-S2V	4.2809	9.4462	0.2640	0.1983
	SingDance	**4.9672**	**8.9363**	**0.2723**	**0.2558**
Listener	Wan-S2V	4.1676	9.5386	0.2480	0.2065
	SingDance	**1.1768**	**12.4059**	**0.2704**	**0.2787**

Wan-S2V's lip-synchronization scores change little under the prompt-only switch. In contrast, switching SingDance to the listener role strongly suppresses vocal synchronization while preserving motion–beat alignment, showing selective articulation control without disrupting music-aligned dance. Figure 4 shows paired frames; the demo page also provides audio-enabled listener-mode videos in speech-driven settings. Read together with Table 2, these results show song-conditioned dance and role-controlled articulation in the same outputs. Appendix B further isolates textual and learned role conditions.

4.4 Speech-Driven Lip Synchronization

Reliable source-role lip synchronization is needed for held-out singing-and-dancing. Table 4 compares SingDance with Hallo3 Cui et al. (2024), Wan-S2V Gao et al. (2025), HY-Avatar Chen et al. (2025b), and InfiniteTalk Yang et al. (2025b) on EMTD. SingDance achieves competitive LSE-C/LSE-D with roughly one-third of InfiniteTalk's generation-time parameters and one-ninth of its inference time, while leading in Aesthetic and Imaging. The same checkpoint generates the dance results in Table 2, showing that a unified model can retain strong speech articulation without a separate lip-sync module. This matters for Song/Source inference, where vocal articulation and music-responsive movement must appear in the same video rather than be combined after generation. Speaking/listening examples appear in Appendix B (Figure 5) and as audio-enabled videos on the demo page.

Table 4: Speech-driven lip-sync and visual quality on EMTD. Params. are generation-time billions; best and second-best are bold and underlined.

Method	Params. ↓	LSE-C ↑	LSE-D ↓	Subj. ↑	Aes. ↑	Img. ↑
Hallo3	14.55	5.352	9.828	0.9746	0.511	0.677
Wan-S2V	16.30	6.999	8.292	0.9812	<u>0.580</u>	<u>0.698</u>
HY-Avatar	<u>12.99</u>	7.210	8.494	0.9791	0.535	0.661
InfiniteTalk	18.88	**8.535**	**7.108**	**0.9868**	0.545	0.680
SingDance	**5.63**	<u>7.995</u>	<u>7.551</u>	<u>0.9859</u>	**0.593**	**0.722**

4.5 Ablation Study

Table 5 shows that removing the routed music token lowers BeatAlign and MBCR on both test sets while Aesthetic and Imaging remain broadly comparable. The consistent trend across singing-and-dancing and dancing-only inputs supports an incremental rhythmic contribution from the music pathway, with little change in visual quality. Improving both metrics matters: the gain concerns the timing of detected movement as well as coverage of musical beats, rather than only a few isolated accents. Because the ablation retains the other conditions, it isolates the added value of the routed music token within SingDance. Audio-enabled song and instrumental ablations on the demo page show how this change affects the generated dance motion.

Table 5: Music-token ablation. Aes. and Img. denote Aesthetic and Imaging Quality; bold marks the better variant within each test set.

Variant	BeatAlign ↑	MBCR ↑	Aes. ↑	Img. ↑
SingDance-Test				
w/o M	0.2462	0.2386	**0.5864**	0.6598
Full	**0.2723**	**0.2558**	0.5846	**0.6688**
Dance-Only-Test				
w/o M	0.2853	0.2457	0.5687	0.6734
Full	**0.3002**	**0.2632**	**0.5718**	**0.6738**

5 Conclusion

We introduced SingDance, a role-aware video generation framework that separates vocal content from the visible subject's role while coordinating text-guided action and music-conditioned body motion. Hard-compact routing and frame-wise joint audio injection compose speech, music, and role conditions within a shared video diffusion backbone. From asymmetric speaking/listening and dancing-only supervision, SingDance realizes compositional zero-shot singing-and-dancing in the held-out Song/Source configuration. Experiments validate music-aligned dance, paired vocal-role control, and parameter-efficient lip synchronization.

AI USE STATEMENT

During the preparation of this manuscript, large language models assisted with document and literature retrieval and served as writing aids to improve linguistic clarity. Multimodal and text-based AI models also assisted with data screening, with repeated human review of sampled clips across the filtering stages, as described in Appendix C. All scientific contributions, including the research ideas, experimental design, final analyses, and conclusions presented here, are the original work of the human authors. The authors checked the cited literature and AI-assisted text and take responsibility for the final content of this paper.

ETHICS STATEMENT

The human images, voices, and music used in training, evaluation, and the demo come from public datasets and internal data sources. SingDance aims to advance research on controllable audio–visual generation. We recognize the risk of impersonation and unauthorized use of a person's likeness or voice, and will not use SingDance for such purposes.

REPRODUCIBILITY STATEMENT

We provide implementation code and configuration files as supplementary material with this submission. Appendix C describes the data-selection and filtering procedures, and the experimental settings and evaluation metrics are specified in the paper. The anonymous demo page linked in the abstract presents audio-enabled generated videos and comparisons.

REFERENCES

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, 2020.

Liyang Chen, Tianxiang Ma, Jiawei Liu, Bingchuan Li, Zhuowei Chen, Lijie Liu, Xu He, Gen Li, Qian He, and Zhiyong Wu. HuMo: Human-centric video generation via collaborative multi-modal conditioning, 2025a. URL https://arxiv.org/abs/2509.08519.

Yi Chen, Sen Liang, Zixiang Zhou, Ziyao Huang, Yifeng Ma, Junshu Tang, Qin Lin, Yuan Zhou, and Qinglin Lu. Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters, 2025b.

Zeyuan Chen, Hongyi Xu, Guoxian Song, You Xie, Chenxu Zhang, Xin Chen, Chao Wang, Di Chang, and Linjie Luo. X-dancer: Expressive music to human dance video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025c.

Xuangeng Chu, Ruicong Liu, Yifei Huang, Yun Liu, Yichen Peng, and Bo Zheng. Unils: End-to-end audio-driven avatars for unified listening and speaking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops*, pp. 251–263. Springer, 2017.

Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer, 2024.

Qijun Gan, Ruizi Yang, Jianke Zhu, Shaofei Xue, and Steven Hoi. Omniavatar: Efficient audio-driven avatar video generation with adaptive body animation, 2025.

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

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

Susung Hong, Ira Kemelmacher-Shlizerman, Brian Curless, and Steven M. Seitz. Musicinfuser: Making video diffusion listen and dance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Mingyang Huang, Peng Zhang, Li Hu, Guangyuan Wang, Ruoshi Zhang, Yi Lu, Gang Cheng, and Bang Zhang. Wan-dancer: A hierarchical framework for minute-scale coherent music-to-dance generation, 2026.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024a.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models, 2024b.

Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, Yansong Tang, and Xiu Li. Finedance: A fine-grained choreography dataset for 3d full body dance generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++, 2021.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Meituan LongCat Team. Longcat-video-avatar 1.5 technical report, 2026.

Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. Echomimicv2: Towards striking, simplified, and semi-body human animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Rang Meng, Yan Wang, Weipeng Wu, Ruobing Zheng, Yuming Li, and Chenguang Ma. Echomimicv3: 1.3b parameters are all you need for unified multi-modal and multi-task human animation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Ziqiao Peng, Yanbo Fan, Haoyu Wu, Xuan Wang, Hongyan Liu, Jun He, and Zhaoxin Fan. Dualtalk: Dual-speaker interaction for 3d talking head conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21055–21064, 2025.

Ziqiao Peng, Yi Chen, Yifeng Ma, Guozhen Zhang, Zhiyao Sun, Zixiang Zhou, Youliang Zhang, Zhengguang Zhou, Zhaoxin Fan, Hongyan Liu, Yuan Zhou, Qinglin Lu, and Jun He. Actavatar: Temporally-aware precise action control for talking avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning*, 2023.

Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Zhiyao Sun, Ziqiao Peng, Yifeng Ma, Yi Chen, Zhengguang Zhou, Zixiang Zhou, Guozhen Zhang, Youliang Zhang, Yuan Zhou, Qinglin Lu, and Yong-Jin Liu. Streamavatar: Streaming diffusion models for real-time interactive human avatars, 2026.

Jonathan Tseng, Rodrigo Castellon, and C. Karen Liu. Edge: Editable dance generation from music, 2022.

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.

Lizhen Wang, Yongming Zhu, Zhipeng Ge, Youwei Zheng, Longhao Zhang, Tianshu Hu, Shiyang Qin, Mingshuang Luo, Jiaxu Zhang, Xin Chen, et al. Flowact-r1: Towards interactive humanoid video generation, 2026.

Yuzhe Weng, Haotian Wang, Xinyi Yu, Xiaoyan Wu, Haoran Xu, Shan He, and Jun Du. Beyond monologue: Interactive talking-listening avatar generation with conversational audio context-aware kernels, 2026.

Kaixing Yang, Jiashu Zhu, Xulong Tang, Ziqiao Peng, Xiangyue Zhang, Puwei Wang, Jiahong Wu, Xiangxiang Chu, Hongyan Liu, and Jun He. Mace-dance: Motion-appearance cascaded experts for music-driven dance video generation, 2025a.

Kaixing Yang, Jiashu Zhu, Xulong Tang, Ziqiao Peng, Xiangyue Zhang, Chubin Chen, Puwei Wang, Jiahong Wu, Xiangxiang Chu, Hongyan Liu, and Jun He. Omnidance: Multimodal driven dance video generation with large-scale internet data, 2026a.

Shaoshu Yang, Zhe Kong, Feng Gao, Meng Cheng, Xiangyu Liu, Yong Zhang, Zhuoliang Kang, Wenhan Luo, Xunliang Cai, Ran He, and Xiaoming Wei. Infinitetalk: Audio-driven video generation for sparse-frame video dubbing, 2025b.

Xiaoda Yang, Majun Zhang, Changhao Pan, Nick Huang, Yuguang Yang, Fan Zhuo, Pengfei Zhou, Jin Zhou, Sizhe Shan, Shan Yang, Miles Yang, Yang You, and Zhou Zhao. TMD-Bench: A multi-level evaluation paradigm for music–dance co-generation, 2026b.

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

Ailing Zeng, Casper Yang, Chauncey Ge, Eddie Zhang, Garvey Xu, Gavin Lin, Gilbert Gu, Jeremy Pi, Leo Li, Mingyi Shi, et al. Lpm 1.0: Video-based character performance model, 2026.

Yu Zhang, Kaiyuan Shen, and Yang Li. Embodiedhead: Real-time listening and speaking avatar for conversational agents, 2026.

Haina Zhu, Yizhi Zhou, Hangting Chen, Jianwei Yu, Ziyang Ma, Rongzhi Gu, Yi Luo, Wei Tan, and Xie Chen. Muq: Self-supervised music representation learning with mel residual vector quantization, 2025.

A Supplementary Overview

This appendix disentangles textual and learned vocal-role control through a factorial intervention, describes the datasets and leakage controls, specifies the motion–beat evaluation protocol, and explains the scope of baseline comparisons.

B Textual and Learned Vocal-Role Intervention

We conduct a 2×2 intervention study to examine how the textual role and the learned vocal-role condition contribute to controllable articulation. We independently switch the prompt's *vocal role* field and the learned vocal-role condition between ON and OFF. The learned condition comprises the global role embedding and local role token described in the main paper. ON assigns the vocals to the visible subject, whereas OFF assigns them to an off-screen performer. The reference image, song, role-independent prompt fields, random seed, model checkpoint, and all other inference settings remain fixed.

We evaluate the generated videos using the official SyncNet protocol Chung & Zisserman (2017). Higher LSE-C and lower LSE-D indicate stronger audio–visual synchronization.

Table 6: Factorial intervention of the textual vocal-role field and learned vocal-role condition. ON assigns the vocals to the visible subject, whereas OFF assigns them to an off-screen performer.

Vocal-role assignment		SyncNet	
Textual field	Learned condition	LSE-C	LSE-D
ON	ON	4.97	8.94
ON	OFF	1.77	11.49
OFF	ON	4.98	8.98
OFF	OFF	1.18	12.41

As shown in Table 6, articulation primarily follows the learned condition. Switching it from ON to OFF while keeping the textual role ON substantially weakens synchronization, whereas changing only the textual role has little effect when the learned condition remains ON. The OFF/OFF setting yields the strongest suppression.

The textual role nevertheless provides additional semantic guidance. Under the OFF learned condition, changing the textual role from ON to OFF further suppresses synchronization. The learned condition therefore provides the dominant articulation control, while the textual field supplies a complementary signal when the visible subject is assigned the listener role.



Figure 5: Representative source- and listener-role frames generated from the same reference image and speech audio. Audio-enabled videos on the demo page linked in the abstract show the corresponding temporal behavior.

Figure 5 compares representative frames from the source and listener roles. The audio-enabled demo videos allow readers to assess articulation and listening behavior over time.

C Datasets, Data Curation, and Leakage Control

To support compositional zero-shot singing-and-dancing generation, we must ensure that the held-out Song/Source behavior is not accidentally observed during task training. We therefore curate the training data with a precision-oriented policy: any sample that may introduce visible singing, lip-sync leakage, unclear vocal ownership, or an invalid role assignment is treated conservatively. In practice, we use a multi-round filtering pipeline for dance data, a separate off-screen speech-listener pipeline, and human-in-the-loop auditing throughout the process.

C.1 Training and Evaluation Datasets

The speech phase combines HuMoSet Chen et al. (2025a) with a proprietary web-video collection, totaling approximately 300,000 five-second clips. Approximately 12,000 clips depict genuine listening or reacting toward an off-screen interlocutor; the remainder primarily depict on-screen speaking. The dance phase combines cleaned MA-Data Yang et al. (2025a) with a proprietary dance-video collection, yielding approximately 48,000 five-second clips: 18,000 with instrumental music and 30,000 with songs. We use SyncNet Chung & Zisserman (2017) to remove song-conditioned clips with detectable or uncertain visible lip synchronization, retaining the remainder as dancing-only supervision. No singing-and-dancing videos are used in SingDance task training.

Existing public music–dance benchmarks do not fully cover our task Li et al. (2021); Yang et al. (2026b). We therefore construct two task-specific evaluation sets. Both are disjoint from SingDance task-training data: *SingDance-Test* contains 50 face-visible vocal-song cases for singing-and-dancing, and *Dance-Only-Test* contains 100 diverse dancing-only cases. Talking is separately evaluated on EMTD Meng et al. (2025). The vocal-role intervention reuses *SingDance-Test*; each listener counterpart keeps the reference image, soundtrack, role-independent prompt fields, sampling seed, checkpoint, and other inference settings fixed while changing the vocal-role assignment.

C.2 Dance Data Curation

We apply the same in-house audio–visual captioner to public and proprietary dance candidates. It produces structured descriptions of subjects, actions, visual style, camera motion, and audio content, including speech and multilingual lyrics. We then apply the same semantic filtering rules to all dance sources.

Round 1: semantic filtering from structured captions. The first round parses the structured captions and determines whether the audible vocals or lyrics may come from the visible subject. The filtering rules are iteratively refined through repeated human review of sampled clips. After this refinement, the merged first-round statistics over two dance candidate subsets are shown in Table 7.

Table 7: First-stage semantic filtering statistics for the merged dance candidate subsets.

First-stage category	Count / Ratio	Handling
Off-screen vocals or lyrics	57,140 / 68.21%	Song/Listener candidate
Instrumental or no vocals	19,436 / 23.20%	Instrumental/Listener candidate
Visible speaking or singing	690 / 0.82%	Excluded
Ambiguous description	3,764 / 4.49%	Risk candidate
Unknown vocal visibility	1,665 / 1.99%	Excluded or quarantined
Instrumental/vocal conflict	598 / 0.71%	Excluded or quarantined
On-screen/off-screen conflict	482 / 0.58%	Excluded or quarantined
Total	83,775 / 100%	Unified first-stage statistics

The two candidate categories that can potentially be used for training, off-screen vocals/lyrics and instrumental/no-vocal music, account for 76,576 clips, or 91.41% of the first-stage candidates. For

the vast majority of candidates, the first-stage captions did not describe visible singing. However, to minimize possible Song/Source leakage, we still adopt an aggressive filtering strategy. Visible speaking or singing, ambiguous vocal ownership, unknown visibility, and conflicting descriptions are treated as risk signals and are not automatically used for training.

During this stage, human review of sampled clips also reveals recurring failure cases that can be missed by caption-based rules. We therefore add a set of high-risk exclusion rules to remove ordinary talking videos, interviews, tutorials, promotional videos, daily-life clips, in-car navigation scenes, non-dance scenes, caption–visual contradictions, unsuitable subject layouts, and other repeated failure patterns. This additional rule layer removes only about 1–2% of high-confidence candidates, but it is important for dataset precision.

Round 2: audio–visual synchronization filtering. The second round uses SyncNet to detect whether the mouth motion of the visible subject is synchronized with the audio. We first use the S3FD face detector in SyncNet to obtain face tracks and compute lip-sync scores. For samples with reliable face tracks, we use a conservative non-lip-sync criterion:

$$\text{confidence} \leq 1.0 \quad \text{and} \quad |\text{offset}| \geq 12 \text{ frames}.$$

This threshold is chosen based on repeated human review of sampled clips and is intentionally conservative. It prioritizes precision over recall and therefore removes many clips that may in fact be valid non-lip-sync dance samples.

For samples where S3FD fails to detect a reliable face track, we apply MediaPipe as an additional face detector and re-evaluate the sample. If neither detector can provide reliable evidence for lip-sync assessment, the sample is removed rather than treated as safe. Among 32,094 high-confidence instrumental/off-screen-vocal candidates with complete synchronization-audit records, 26,143 satisfy the strict non-lip-sync criterion, corresponding to 81.5%.

Round 3: text-only LLM audit of captions. The third round performs a text-only LLM audit on a caption-audit pool constructed after first-stage semantic filtering and rule-based exclusions. This audit is not a second multimodal captioning pass; it operates on the structured captions and checks for remaining semantic risks. Specifically, it assesses what the captions indicate about vocal ownership, visible speaking or lip-syncing, dance-scene validity, subject count, and known failure modes such as ordinary talking or non-dance content.

This text-only audit is complementary to the audio–visual synchronization check rather than a strict count cascade after the SyncNet pass set. The SyncNet strict-pass subset and the text-audit pool have separate denominators, and both provide conservative evidence for the final acceptance decision. Table 8 summarizes the text-audit scope and outcome. In the audited pool of approximately 31.7K candidates, about 94% are accepted and about 6% are excluded. For samples labeled ambiguous in the first stage, only about 7% are eventually retained with high confidence. Thus, ambiguity is not treated as a permissive route into training.

Table 8: Text-only LLM audit scope and outcome for dance candidates.

Audit group	Criterion and outcome
Text-audit pool	$\sim$31.7K candidates involving vocal ownership, visible articulation, or semantic uncertainty; $\sim$94% accepted and $\sim$6% excluded
First-stage ambiguous samples	Only $\sim$7% are retained with high confidence

After all filtering stages and source merging, the final dance training set contains approximately 48K clips. Table 9 summarizes the final route assignment. All accepted dance clips are assigned to either Instrumental/Listener or Song/Listener. The Song/Source configuration is absent from the SingDance task-training data.

Table 9: Final dance training routes; Song/Source is held out.

Training route	Meaning	Scale
Instrumental/Listener	Dance to music without vocals	∼18K
Song/Listener	Dance to songs with off-screen vocals	∼30K
Song/Source	Visible singing-and-dancing	0

For instrumental clips, the *Listener* label denotes the non-source role state; no vocal signal is present to assign to the visible subject.

C.3 Off-Screen Speech-Listener Curation

The speech stage provides explicit supervision for the Source and Listener roles. Speech/Source samples require the visible subject to produce the input speech, with consistent identity, mouth motion, and audio content. Speech/Listener samples impose the opposite condition: the audio comes from a real off-screen speaker, while the visible subject remains silent and responds naturally through facial expression, gaze, head motion, or body motion.

The key goal of the Speech/Listener split is not to collect silent faces. Instead, we aim to collect clips in which the visible subject is plausibly receiving, attending to, or reacting to off-screen speech. This distinction is important because a silent person with no relation to the audio does not provide meaningful Listener-role supervision.

Table 10: Curation stages for off-screen Speech/Listener reaction data.

Stage	Retention and exclusion criteria
Candidate retrieval	Retrieve candidates from large-scale speech-video descriptions with an off-screen speaker, a single visible human subject, and no visible speech by the on-screen subject.
Semantic audit	Retain direct dialogue, phone or device audio, and environmental off-screen speech with natural silent reactions. Exclude narration, commentary, sales or tutorial speech, news-style broadcasting, low-quality dubbing, inner monologue, stage dubbing, visible speaking, non-human subjects, partial bodies, multi-person scenes, and unstable faces.
Face quality	Require the visible face to be stable and sufficiently large in sampled frames, so that facial expression, gaze, and head motion can provide reliable Listener-role supervision.
Source merging and deduplication	Merge candidates from multiple sources using the same rules and deduplicate by video identity.
Caption normalization	Assign the speaker to the off-screen source and describe the visible subject only through silent reactions. Avoid describing the visible subject as speaking, singing, or lip-syncing.
Media validation	Validate structured descriptions, role state, video decodability, frame rate, and 24-FPS conversion before final training selection.

We construct this split through a multi-stage pipeline summarized in Table 10. We first retrieve candidates from a large speech-video corpus using structured descriptions. The retrieval stage searches for clips with an off-screen speaker, a single visible human subject, and no visible speech by the on-screen subject. We then apply a targeted semantic audit to distinguish genuine off-screen interaction from unrelated voice-over or narration. The audit retains direct dialogue, phone or device audio, and environmental off-screen speech accompanied by natural silent reactions. It removes narration, commentary, sales or tutorial speech, news-style broadcasting, low-quality platform dubbing, inner monologue, stage dubbing, visible speaking, non-human subjects, partial bodies, multi-person scenes, and unstable faces.

After semantic auditing, we require the visible face to be stable and sufficiently large in sampled frames. This face-quality step ensures that the retained clips can provide useful supervision for

facial expression, gaze, and head motion. Candidates from multiple sources are then merged under the same rules and deduplicated by video identity.

The retained samples are further normalized into a unified training description. The speaker is fixed as the off-screen speaker, while the visible subject is described only through silent reactions such as expression, gaze, head motion, and body motion. The rewritten description does not characterize the visible subject as speaking, singing, or lip-syncing. This prevents the text condition from assigning the Source role to the visible subject.

Finally, we validate the structured description, role state, video decodability, frame rate, and 24-FPS conversion. The full pipeline begins with more than 80K initial candidates and retains approximately 16K after semantic auditing. Following face-quality filtering, source merging, caption normalization, media validation, and final training selection, approximately 12K off-screen Speech/Listener reaction clips are used for training. SyncNet measurements for this split are retained as diagnostic information and are not used as a final hard-filtering criterion.

C.4 Human-in-the-Loop Auditing

Both data pipelines are constructed with human-in-the-loop quality control. At every major stage, we review sampled clips to verify the current filtering results and to discover recurring failure modes. When a systematic error pattern is found, we convert it into an additional exclusion rule and re-apply the rule to the candidate pool. The pipeline is therefore not a single automatic pass, but an iterative filter–audit–refine process.

This human-in-the-loop process is especially important for the dance data. The first-stage statistics show that most candidates are categorized as either off-screen-vocal or instrumental/no-vocal cases, and only a small fraction are explicitly identified as visible speaking or singing. Nevertheless, first-stage semantic ownership is not sufficient to guarantee zero leakage. We therefore use additional synchronization checks, text-only caption audits, and manually derived exclusion rules to prioritize precision over recall.

As a result, the pipeline intentionally sacrifices recall for precision. Samples with unresolved vocal ownership, unknown visibility, synchronization risk, caption–visual contradictions, ordinary talking, non-dance content, unsuitable subject layouts, or manually identified failure patterns are excluded or quarantined. This conservative policy is designed to make the Song/Listener split as clean as possible and to support the compositional zero-shot Song/Source claim.

D Motion–Beat Evaluation Protocol

All methods use the same audio–video time window and the DWPose body, foot, and hand landmarks described in the main text. We normalize mean inter-frame displacement by a clip-level body scale, smooth the resulting signal, and take its strict local minima as kinematic events. Clips with unreliable pose tracks or missing beat events are excluded rather than assigned zero. Motion means use matched evaluable videos, while reference and visual-quality metrics use each method's evaluable outputs.

E Scope of Baseline Comparisons

We organize baseline comparisons by task and supported input interface. Wan-S2V Gao et al. (2025) accepts the same text, reference-image, and audio inputs as SingDance and can generate full-body video from a detailed action prompt. Its Wan foundation-model family Wan Team et al. (2025) and speech-conditioning design provide a matched-input comparison on both test sets; SingDance additionally has a dedicated music pathway and explicit vocal-role state. The paired vocal-role experiment further uses Wan-S2V as a prompt-only control: its LSE-C/LSE-D change little when the prompt assigns the vocals to an off-screen performer.

For *Dance-Only-Test*, we additionally compare with two music-driven video methods. MusicInfuser Hong et al. (2026) directly generates video from text and music without a reference image. X-Dancer Chen et al. (2025c) accepts a reference image and music but generates video through a skeleton-motion intermediate: its first stage predicts music-driven 2D stick-figure pose maps, and

its second stage renders RGB video from those poses and the reference image. SingDance directly generates reference-conditioned video without an explicit motion intermediate. Each method is evaluated through its supported inputs and prompt format. Table 4 separately compares dedicated speech-driven avatar methods on EMTD.

OmniDance and Wan-Dancer Yang et al. (2026a); Huang et al. (2026) are recent music-conditioned dance-video systems. OmniDance supports joint image, text, and music conditioning, but its model weights were not publicly available during our evaluation. Wan-Dancer focuses on long-form music-to-dance generation through a hierarchical pipeline and a fixed set of dance-style prompts. Its text prompts mainly specify dance styles rather than detailed actions, making its control setting closer to X-Dancer's image-and-audio-driven dance generation than to SingDance's free-form text–image–audio control. SingDance additionally controls whether the visible subject produces the vocals, including in the held-out Song/Source setting. We therefore discuss both in related work and restrict our quantitative claims to the baselines evaluated in Table 2.