Meric: A Unified Framework for Multimodal Music Generation and Retrieval via Representation Space Anchoring

Xihua Wang*, Yinbo Wang*, Jingchao Zhang*, and Ruihua Song†

Gaoling School of Artificial Intelligence, Renmin University of China
xhwang78@gmail.com, songruihua_bloon@outlook.com
Project page: https://audiofool934.github.io/meric/

Abstract. Cross-modal music generation from text and visual inputs is uniquely challenging due to its subjective, one-to-many nature. Current methods face three major limitations: they are bottlenecked by scarce paired data, lack explicit one-to-many modeling, and isolate generation from retrieval. To address these issues, we propose Meric, a unified framework for multimodal music generation and retrieval. Our core innovation is a music semantic anchor that decouples multimodal understanding from acoustic synthesis. We employ a two-stage training pipeline. First, a Flow Matching decoder learns to synthesize audio from anchor embeddings using large, unpaired music corpora. Second, a lightweight generative diffusion module maps multimodal inputs into this continuous anchor space using limited paired data. This decoupled design explicitly models cross-modal ambiguity and exploits unpaired data to bypass the paired data bottleneck. During inference, the predicted anchor embedding simultaneously acts as a condition for acoustic synthesis and a query for music retrieval. Extensive experiments demonstrate that Meric achieves state-of-the-art performance across text-to-music and vision-to-music generation and retrieval benchmarks.

1 Introduction

The past few years have witnessed remarkable advances in multimodal generative models across images [10,48], video [3,46], natural audio [11,28], and speech [20]. Music, however, occupies a singular position among these modalities: as an inherently artistic and abstract medium, its connection to visual scenes or text is subjective, culturally situated, and one-to-many, making cross-modal music generation and retrieval uniquely challenging. This challenge is increasingly recognized at scale: Google's Lyria 3 [14], a multimodal music generation model integrated into Gemini [7], highlights unified music-vision-language understanding and generation as a core capability for next AGI systems.

*Equal contribution.
†Corresponding author.

Existing academic work on cross-modal music generation mainly focuses on individual input modalities. A dominant research line explores text-to-music generation [1, 6, 34], while a separate body of work targets video-to-music synthesis [12, 42, 57]. Some recent methods have attempted to unify multiple input modalities [5, 29, 40]. Some works extract per-modality features and train a generative model conditioned on these features using paired data [40]; others align non-text modalities into the conditioning space of a text-to-music model [5, 29].

Despite recent progress, we identify three fundamental limitations that remain unaddressed in academic multimodal music generation. **(i) Limited scalability from paired data bottleneck.** Current approaches rely heavily on scarce paired data (visual–music or text–music pairs) for end-to-end training [5, 28, 29, 40]. This constraint severely limits data scale and, consequently, the acoustic quality of generated music. Larger, unlabeled, single-modality music corpora remains largely unexploited. **(ii) Lack of explicit one-to-many modeling.** The correspondence between a multimodal prompt and music is inherently ambiguous: a sunset video could inspire a melancholic piano piece as readily as an uplifting orchestral swell. Existing models implicitly handle this ambiguity through stochastic noise in their generative process [40], without any explicit or interpretable mechanism for representing semantic diversity. **(iii) Absence of unified generation and retrieval.** Retrieval and generation serve complementary roles in music applications [47, 53]. Yet current systems treat these as entirely separate models [49], missing the opportunity to leverage a shared cross-modal understanding for both tasks, and to use retrieved candidates to guide or augment generation in challenging long-tail scenarios [51].

Motivated by these challenges, we propose MERIC, a unified framework for cross-modal music generation and retrieval from text, image, and video inputs. The key insight behind MERIC draws from recent findings in representation-aligned generation (REPA [52] and RAE [54]): *a well-structured semantic representation of the target modality greatly facilitates its generative modeling*. We therefore introduce a **music semantic anchor**, a semantic embedding space defined by a dedicated music semantic encoder, and use it to decouple multimodal understanding from acoustic synthesis. Understanding is modeled as a mapping from arbitrary multimodal inputs into the anchor space (a one-to-many process); synthesis is modeled as a mapping from the anchor space to acoustic waveforms.

This decomposition yields a two-stage training pipeline that directly addresses all three limitations. **Stage I** trains a Flow based DiT [33] decoder to reconstruct music from anchor embeddings using *unpaired, single-modality music data*, decoupling acoustic quality from paired data availability. **Stage II** trains a generative Music Head, a lightweight diffusion module to map multimodal embeddings extracted by Qwen3-VL-Embedding [25] into the anchor space, using paired multimodal–music data. Because the anchor is a continuous semantic space rather than a deterministic point, the Music Head naturally models the one-to-many structure of cross-modal music association. At inference, the anchor embedding produced by the Music Head serves *simultaneously* as the condition for acoustic synthesis and as the query embedding for retrieval against a mu-

sic library indexed in the same anchor space, unifying generation and retrieval within a single forward pass.

We summarize our main contributions as follows:

- We propose MERIC, the first unified framework supporting both *generation* and *retrieval* of music from multimodal inputs (text, image, video).
- We introduce a music semantic anchor that decouples multimodal understanding from acoustic synthesis, enabling joint exploitation of large unpaired music corpora and scarce paired multimodal data—significantly improving both data scalability and generation quality.
- Our anchor-based architecture provides an explicit, interpretable one-to-many semantic mapping, going beyond noise-driven diversity to capture genuine cross-modal ambiguity.
- Extensive experiments demonstrate that MERIC achieves state-of-the-art or competitive performance on text-to-music generation, vision-to-music generation, text–music retrieval, and vision–music retrieval benchmarks. Project page: https://audiofool934.github.io/meric.

2 Related Work

2.1 Flow Matching

Flow Matching (FM) is a generative modeling framework that constructs continuous probability paths from a simple prior to complex data. Built on Continuous Normalizing Flows, it uses simulation-free training by directly regressing conditional vector fields, avoiding costly ODE solving. It combines diffusion models' strengths with faster sampling and more stable training. Lipman et al. [27] introduced this framework and demonstrated strong performance on image generation benchmarks, while Holderrieth et al. [18] further elaborated on its mathematical foundations, covering both ODE and SDE perspectives.

2.2 Multimodal Music Generation

Recent advances have propelled music generation from symbolic representations to waveform synthesis, drawing on Transformer architectures [45] and diffusion models [17]. Multimodal approaches extend this by conditioning on diverse inputs such as text, images, and videos [1].

Text-to-Music Text-to-music generation synthesizes audio from descriptive prompts specifying genre, instruments, or emotions. Copet et al. [6] introduced MusicGen, a single-stage sequence-to-sequence Transformer for high-fidelity and controllable generation. Latent diffusion variants, such as AudioLDM2 [28], compress Mel-spectrograms via VAEs [23] and progressively refine intermediates for improved coherence. These models excel at capturing explicit musical attributes but struggle with non-textual conditioning, motivating multimodal extensions.

Vision-to-Music Vision-to-music generation translates visual inputs into music by capturing emotional tones or scenic contexts. Early methods [9, 43] used pre-trained networks like VGG [39] to extract affective features for MIDI or waveform synthesis. Art2Mus [36] generates music from artworks via latent diffusion, bypassing intermediate text. GVMGen [57] establishes video-to-music baselines but shows limitations in emotional accuracy.

Multimodal-to-Music Multimodal-to-music generation combines heterogeneous inputs for holistic music synthesis. M2UGen [29] leverages large language models to unify multimodal representations before feeding into generators like MusicGen. Chowdhury et al. [5] augmented text conditioning with image inputs, and MusFlow [40] pioneered flow matching with MLP adapters [19] to align CLIP/CLAP [35, 49] embeddings from image–story–caption–music quadruplets. AudioX [44] offers a unified anything-to-audio framework with improved cross-modal alignment. Our proposed Meric advances this line by unifying generation and retrieval within a single pipeline through a shared music semantic anchor. By decoupling cross-modal mapping from waveform synthesis, it provides an explicit one-to-many mechanism and overcomes the paired-data bottleneck of prior end-to-end approaches.

2.3 Music Generation Evaluation

Single-Modality Evaluation Single-modality metrics assess audio fidelity and distributional similarity in isolation: Fréchet Audio Distance (FAD) [22] measures distributional similarity between generated and reference audio embeddings using backbones such as VGGish [16], CLAP [49], or MERT [26], while Inception Score (IS) analogs [37] and KL divergence further quantify generation diversity and distributional discrepancy [24].

Multimodal Evaluation Multimodal evaluation additionally requires assessing semantic alignment between heterogeneous conditioning inputs and generated audio. Existing cross-modal metrics such as the CLAP score [49], ImageBind Score [13], and IMSM [5] provide partial proxies for text-audio and vision-audio alignment, yet remain fragmented, modality-specific, and poorly correlated with holistic human judgment [21]. Meanwhile, the LLM-as-a-judge paradigm has demonstrated strong scalability in related domains: GPT-4 achieves over 80% agreement with human preferences on dialogue benchmarks [55], and G-Eval [31] enables fine-grained assessment via chain-of-thought prompting, with similar approaches extended to vision-language tasks [4]. However, LLM-based evaluation for multimodal music generation remains largely unexplored. This gap motivates us to incorporate a multimodal LLM judge capable of jointly reasoning over audio, visual, and textual signals as a complementary evaluation strategy.

3 Method

We present **Meric**, a unified framework that addresses multimodal-to-music generation and retrieval within a single architecture. Given an arbitrary query from text, image, or video modality, Meric produces either a music waveform (generation) or a ranked list of music items (retrieval).

The key insight is that generation and retrieval share the same semantic bottleneck: a *music semantic embedding* predicted from the multimodal query. For generation, this embedding is decoded into music waveform via a flow decoder and a waveform decoder; for retrieval, it serves directly as the query vector against a pre-indexed music embedding (encoded by a music semantic encoder). This design eliminates the need for separate models or training objectives for the two tasks.

The remainder of this section is organized as follows. Section 3.1 reviews the generative model preliminaries: diffusion models and flow matching, that underpin our acoustic decoder. Section 3.2 describes the Meric framework and its inference-time behavior for both tasks. Section 3.3 details our two-stage decoupled training strategy. Section 3.4 introduces our LLM-as-Judge evaluation protocol, which uses an multimodal LLM to assess the semantic correspondence between multimodal inputs and generated music.

3.1 Preliminary

Diffusion Models. Diffusion models [17, 41] learn to reverse a gradual noising process. Given a data sample $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, a forward Markov chain adds Gaussian noise over T steps: $q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t;\, \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0,\, (1-\bar{\alpha}_t)\mathbf{I})$, where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ and $\{\beta_s\}$ is a predefined noise schedule. A neural network $\boldsymbol{\epsilon}_\theta$ is trained to predict the added noise via:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{x}_0,\, \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}),\, t}\left[\, \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})\|^2 \right], \tag{1}$$

where $\mathbf{c}$ denotes an optional conditioning signal. We exploit this objective to train our Music Head module (Section 3.2) to bridge multimodal semantics and music semantics.

Flow Matching. Flow Matching (FM) [27,30] defines a probability path between a simple source distribution $p_0 = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a target data distribution p_1 via an ordinary differential equation (ODE): $\frac{d\mathbf{x}}{dt} = \mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c})$, where $\mathbf{v}_\theta$ is a neural velocity field. The conditional FM objective regresses $\mathbf{v}_\theta$ onto the ground-truth velocity along linear interpolants $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,\, \mathbf{x}_0,\, \mathbf{x}_1}\left[\|\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}) - (\mathbf{x}_1 - \mathbf{x}_0)\|^2\right]. \tag{2}$$

Compared to vanilla diffusion, FM yields straighter sampling trajectories and requires fewer function evaluations at inference, making it well-suited for high-fidelity audio latent decoding. We use a Diffusion Transformer (DiT) [33] backbone as the velocity field, referred to as the **Flow-DiT Decoder**.

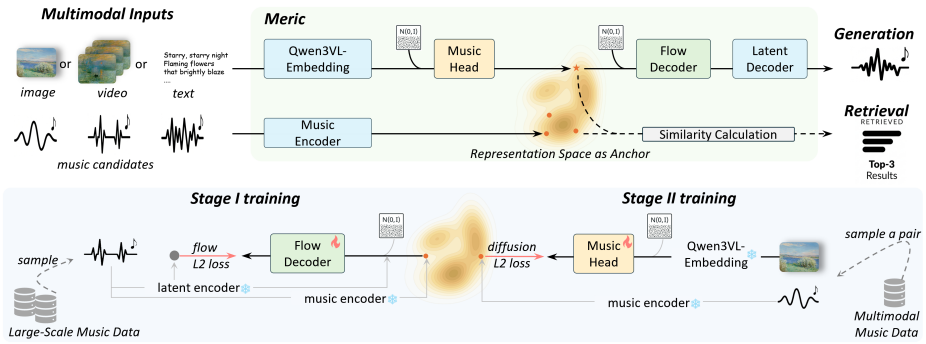


Fig. 1: Meric framework and training stages. At inference, multimodal queries (image / video / text) are encoded by a frozen Qwen3-VL-Embedding model. The Music Head maps the resulting query embedding to a music semantic embedding, which serves as the common interface for both generation (decoded via Flow-DiT and a latent decoder) and retrieval (matched against pre-indexed music embeddings).

3.2 Meric Framework

Meric consists of four components: (i) a multimodal encoder, (ii) a Music Head, (iii) a Flow-DiT Decoder, and (iv) a Music Semantic Encoder. Figure 1 illustrates the full pipeline.

Multimodal Encoder. We adopt **Qwen3-VL-Embedding** [25] as a frozen universal encoder to extract semantic representations from all supported query modalities. Given a query q of any modality (text string, image, or video frames), the encoder produces a fixed-dimensional embedding:

$$\mathbf{e}_q = f_{\text{QVLE}}(q) \in \mathbb{R}^d. \tag{3}$$

Using a single pre-trained vision–language embedding model for all modalities ensures consistent semantic alignment across modalities without requiring modality-specific encoders.

Music Head. The Music Head g_ϕ is a lightweight diffusion-based module that translates the multimodal query embedding $\mathbf{e}_q$ into a *music semantic embedding* $\hat{\mathbf{z}}_m \in \mathbb{R}^{d_m}$:

$$\hat{\mathbf{z}}_m = g_\phi(\mathbf{e}_q). \tag{4}$$

Rather than performing a deterministic mapping, we formulate this translation as a conditional diffusion process conditioned on $\mathbf{e}_q$, which allows the model to capture the inherent one-to-many correspondence between a visual/textual concept and suitable music semantics. The Music Head is trained with $\mathcal{L}_{\text{diff}}$ using paired multimodal–music data (Section 3.3).

Flow-DiT Decoder. The Flow-DiT Decoder h_ψ maps the music semantic embedding $\hat{\mathbf{z}}_m$ to a compressed latent $\mathbf{l} \in \mathbb{R}^{T_l \times C}$:

$$\mathbf{l} = h_\psi(\mathbf{z}_n \xrightarrow{\text{ODE}} \mathbf{l} \mid \hat{\mathbf{z}}_m), \tag{5}$$

where $\mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the initial noise sample and the ODE integration is performed using the learned velocity field. The decoder is trained with $\mathcal{L}_{\text{FM}}$ conditioned on music semantic features extracted by the Music Semantic Encoder.

Latent-to-Wave Decoder. A pre-trained variational autoencoder decoder $\mathcal{D}_{\text{wav}}$ converts the latent $\mathbf{l}$ into a time-domain waveform: wave $= \mathcal{D}_{\text{wav}}(\mathbf{l})$.

Music Semantic Encoder. A dedicated Music Semantic Encoder f_{MSE} encodes a music track m into a semantic embedding $\mathbf{z}_m = f_{\text{MSE}}(m)$ living in the same space as $\hat{\mathbf{z}}_m$. This encoder is used both as a conditioning signal during acoustic decoder training and as the index-side encoder during retrieval.

Unified Inference for Generation and Retrieval. At inference time, for **generation**, the pipeline executes: $q \xrightarrow{f_{\text{QVLE}}} \mathbf{e}_q \xrightarrow{g_\phi} \hat{\mathbf{z}}_m \xrightarrow{h_\psi} \mathbf{l} \xrightarrow{\mathcal{D}_{\text{wav}}}$ wave. For **retrieval**, the pipeline executes: $q \xrightarrow{f_{\text{QVLE}}} \mathbf{e}_q \xrightarrow{g_\phi} \hat{\mathbf{z}}_m$, and the music score for a candidate track m_i is computed as:

$$s(q, m_i) = \cos(\hat{\mathbf{z}}_m,\ f_{\text{MSE}}(m_i)), \tag{6}$$

where the music embeddings $\{f_{\text{MSE}}(m_i)\}$ are pre-computed and indexed offline. Both tasks thus share the identical forward pass up to $\hat{\mathbf{z}}_m$, requiring no task-specific branches or heads at inference time.

3.3 Training of Meric

We decouple the training of Meric into two sequential stages, each formulated as a generative learning problem. This decoupling avoids the interference between acoustic quality objectives and cross-modal alignment objectives, and allows each stage to be scaled independently.

Stage I: Music Acoustic Modeling. In the first stage, we train the Flow-DiT Decoder h_ψ on music-only data, independent of any multimodal input. Each training sample is a music track m. The Music Semantic Encoder f_{MSE} extracts a semantic conditioning embedding: $\mathbf{z}_m = f_{\text{MSE}}(m)$. Separately, a wave-to-latents encoder $\mathcal{E}_{\text{wav}}$ extracts the acoustic latent representation: $\mathbf{l} = \mathcal{E}_{\text{wav}}(m)$. The Flow-DiT Decoder is then trained to reconstruct $\mathbf{l}$ from noise conditioned on $\mathbf{z}_m$:

$$\mathcal{L}_{\text{Stage I}} = \mathbb{E}_{m,\, t,\, \mathbf{z}_n}\left[\left\|h_\psi(\mathbf{x}_t, t, \mathbf{z}_m) - (\mathbf{l} - \mathbf{z}_n)\right\|^2\right], \tag{7}$$

where $\mathbf{x}_t = (1-t)\mathbf{z}_n + t\,\mathbf{l}$ is the linear interpolant. After this stage, h_ψ can faithfully reconstruct high-fidelity audio latents from any music semantic embedding.

Stage II: Cross-Modal Semantic Alignment. In the second stage, we train the Music Head g_ϕ using paired multimodal–music data $\{(q_i, m_i)\}$. Given a multimodal query q_i, the frozen Qwen3-VL-Embedding model provides: $\mathbf{e}_{q_i} = f_{\text{QVLE}}(q_i)$. The paired music m_i is encoded by the (also frozen) Music Semantic Encoder: $\mathbf{z}_{m_i} = f_{\text{MSE}}(m_i)$. The Music Head is trained to translate $\mathbf{e}_{q_i}$ into $\mathbf{z}_{m_i}$ via a diffusion objective, treating $\mathbf{z}_{m_i}$ as the denoising target conditioned on $\mathbf{e}_{q_i}$:

$$\mathcal{L}_{\text{Stage II}} = \mathbb{E}_{(q,m),\,\boldsymbol{\epsilon},\,t}\left[\left\|\boldsymbol{\epsilon} - g_\phi\big(\sqrt{\bar{\alpha}_t}\,\mathbf{z}_m + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon},\ t,\ \mathbf{e}_q\big)\right\|^2\right]. \tag{8}$$

Because $\mathbf{z}_{m_i}$ lives in a compact semantic space rather than the raw audio, training converges rapidly and the learned mapping generalizes well across modalities.

Decoupled Training Enables Both Tasks. After both stages, Meric is equipped for generation: g_ϕ predicts a music semantic embedding from a multimodal query, and h_ψ (with waveform decoder $\mathcal{D}_{\text{wav}}$) renders it into audio. Retrieval emerges as a zero-cost byproduct: since g_ϕ maps queries into the *same* semantic space as f_{MSE}, cosine similarity between $\hat{\mathbf{z}}_m$ and pre-indexed $\mathbf{z}_{m_i}$ directly supports ranking—no retrieval-specific training is required.

3.4 LLM as Judge for Cross-Modal Evaluation

Evaluating the semantic relevance between a multimodal query (text, image, or video) and a generated music piece is inherently subjective, making automatic metrics such as FAD [22] or CLAP score [49] insufficient and imprecise on their own. Human evaluation is reliable but costly and difficult to reproduce at scale.

We propose an **LLM-as-Judge** protocol that leverages a large multimodal language model (specifically **Gemini** [7]) as an automatic evaluator of cross-modal semantic alignment between generated music and its conditioning input.

Given a multimodal query q and a generated audio clip, we prompt Gemini with both modalities simultaneously using a structured four-step reasoning framework: (1) **visual scene description**, (2) **audio attribute extraction**, (3) **alignment/mismatch analysis**, and (4) **calibrated scoring**. The prompt instructs the model to assess correspondences in mood, energy, and thematic atmosphere, outputting a scalar score as well as a detailed chain-of-thought rationale. The judge score is as below, with specific prompts detailed in Supplement.

$$s_{\text{judge}}(q, \text{wave}) = \text{Gemini}(\texttt{[prompt]}, q, \text{wave}) \in [0, 100]. \tag{9}$$

4 Experiments

4.1 Experiment Setup

Datasets. We organize training data into two categories. **Unimodal music data** is used to train the Flow Decoder via standard flow matching, comprising 100K full-length songs from FMA [8], 130K clips from MusicSet, and 400K clips

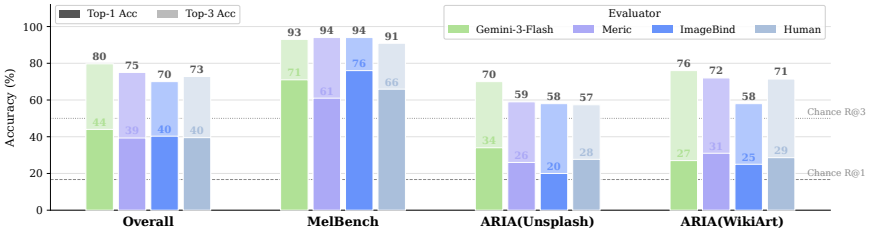


Fig. 2: Performance comparison of different evaluators on cross-modal audio-visual alignment. The test set contains 300 6-way multiple-choice questions (100 MelBench + 100 ARIA-Unsplash + 100 ARIA-WikiArt): given an image, select the matching audio from 1 ground-truth and 5 distractors (random chance: top-1 = 16.7%, top-3 = 50.0%). Metric-based methods (Metric, ImageBind) pick the highest-scoring candidate; Gemini scores each candidate individually (Section 3.4); humans answer directly. Gemini 2.0 Flash achieves the best overall accuracy (44.0%), exceeding human performance.

from AudioSet-Music. **Cross-modal data** for pretraining the Music Head consists of 369K vision-audio pairs drawn from AudioSet-Music (a music subset curated from AudioSet, subsuming the MuImage and MuVideo [29] training splits after YouTube-ID deduplication), MelBench [5], and EmoMV [50]. We further introduce **ARIA** (**A**rt-**R**eferenced **I**nstrumental **A**udio), a curated art-music dataset of 5K images (50% Unsplash photographs, 50% WikiArt artworks), each captioned three times by Gemini 3 Flash and paired with instrumentals from a commercial system (3 tracks, ∼9 10 s segments of per image; 4K/1K train/test split), used for supervised fine-tuning. For evaluation, we benchmark vision-to-music generation on MelBench, MuImage, EmoMV, and ARIA; text-to-music generation on MusicCaps [1] and ARIA; and cross-modal retrieval on MelBench, MusicCaps, and ARIA. For video inputs, up to 3 keyframes are extracted via CLIP-based diversity selection.

Evaluation Metrics. For generation quality we report Fréchet Audio Distance (FAD) [22] with three variants: MERT [26], $\text{CLAP}_{\text{music}}$ [49], and VGGish [16], and paired KL divergence (KL_{mert}). For retrieval we report R@k ($k \in \{1, 10\}$) and MRR. For cross-modal semantic alignment we propose *Sem*, an LLM-based score where Gemini 3 rates each image-audio pair on a 0–100 scale (Sec. 3.4).

Baselines. For vision-to-music generation we compare with Art2Mus [36], GVM-Gen [57], AudioX [44], and two cascade baselines that route Qwen3-VL-8B-Instruct [2] captions through MusicGen [6] or AudioLDM2-Large [28]. For text-to-music generation we compare with MusicGen (Medium/Large), AudioLDM2 (Music/Large), and Stable Audio Open [11]. For retrieval we compare with ImageBind [13] (image-music) and CLAP [49] (text-music).

Implementation Details. Meric uses frozen Qwen3-VL-Embedding-2B [25] (2048-d) followed by a 27M-parameter Music Head projecting into a 512-d music semantic anchor space. Specifically, we utilize a pretrained MuQ [56] encoder to extract the target music semantic representations. The Music Head is pretrained on 369K vision-audio pairs with squared cosine schedule [32] and v-prediction [38], then fine-tuned on 4K ARIA pairs with EMA, min-SNR [15], and

Table 1: Vision-to-music generation performance. We compare Meric against baselines across four benchmarks. The Qwen3VL cascade methods first translate visual input into captions via Qwen3VL, then generates music with text-to-music models. FAD and KL measure audio quality (lower is better); Sem denotes the cross-modal semantic consistency introduced in Sec. 3.4 (higher is better). Best results are **bold**; second-best are underlined. Meric achieves SoTA or near-SoTA across all four benchmarks.

Method	$\text{FAD}_{\text{mert}}\downarrow$	$\text{FAD}_{\text{clap}}\downarrow$	$\text{FAD}_{\text{vgg}}\downarrow$	$\text{KL}_{\text{mert}}\downarrow$	Sem↑
MelBench					
Art2Mus [36]	31.60	0.945	13.72	0.0145	39.1
GVMGen [57]	3.55	0.483	5.78	0.0072	52.9
AudioX [44]	16.48	0.723	13.69	0.0127	53.8
Qwen3VL [2]+MusicGen [6]	4.42	0.497	6.56	0.0076	**80.6**
Qwen3VL [2]+AudioLDM2-L [28]	6.57	0.408	5.92	0.0078	77.6
Meric (Ours)	**1.55**	**0.127**	**1.27**	**0.0059**	74.7
MuImage					
Art2Mus [36]	30.75	0.902	15.28	0.0151	45.3
GVMGen [57]	3.78	0.549	8.73	0.0086	49.6
AudioX [44]	13.90	0.623	14.44	0.0130	51.8
Qwen3VL [2]+MusicGen [6]	2.99	0.418	8.93	0.0079	85.6
Qwen3VL [2]+AudioLDM2-L [28]	4.59	0.274	6.08	0.0079	**87.2**
Meric (Ours)	**1.79**	**0.169**	**3.20**	**0.0078**	87.1
EmoMV					
Art2Mus [36]	33.86	1.109	14.55	0.0147	44.3
GVMGen [57]	3.96	0.617	7.20	**0.0057**	63.6
AudioX [44]	4.75	0.487	5.45	0.0071	79.0
Qwen3VL [2]+MusicGen [6]	7.08	0.736	7.65	0.0083	80.3
Qwen3VL [2]+AudioLDM2-L [28]	8.61	0.609	7.41	0.0086	78.7
Meric (Ours)	**3.06**	**0.253**	**2.45**	0.0060	**81.0**
ARIA					
Art2Mus [36]	15.05	0.560	7.02	0.0077	45.3
GVMGen [57]	4.18	0.337	5.83	**0.0052**	60.0
AudioX [44]	20.15	0.766	11.71	0.0120	38.5
Qwen3VL [2]+MusicGen [6]	5.09	0.283	**2.04**	0.0059	76.6
Qwen3VL [2]+AudioLDM2-L [28]	9.95	0.262	3.66	0.0069	77.6
Meric (Ours)	**3.30**	**0.247**	2.08	0.0055	**79.6**

15% AudioSet replay. The Flow Decoder is trained on large-scale music data, conditioning on the MuQ embedding from each music clip, under standard flow matching training framework. Sampling of Music Head: 20 DDIM steps (guidance 2.0 for generation/ARIA retrieval; 10.0 for MelBench/MusicCaps retrieval). Sampling of Flow Decoder: guidance scale 5.0 with dopri5 solver. Wave decoder decode latents to a 10 s waveform at 44.1 kHz.

4.2 Effectiveness of LLM-Judger

FAD measures distributional quality but not whether generated music semantically matches its visual input. We design an LLM-based protocol and validate it against human annotators.

Table 2: Text-to-music generation performance. FAD (lower is better) and Sem (higher is better) measure audio quality and cross-modal semantic alignment 3.4, respectively. Meric achieves first or second place on both benchmarks.

	MusicCaps			*ARIA*		
Method	**FAD**$_{\text{mert}}$↓	**FAD**$_{\text{clap}}$↓	**Sem**↑	**FAD**$_{\text{mert}}$↓	**FAD**$_{\text{clap}}$↓	**Sem**↑
MusicGen-Medium [6]	3.80	0.421	65.1	3.00	0.272	71.1
MusicGen-Large [6]	<u>3.51</u>	0.409	67.1	<u>2.55</u>	0.248	72.4
AudioLDM2-Music [28]	4.60	0.260	68.0	4.63	0.251	68.6
AudioLDM2-Large [28]	4.17	<u>0.178</u>	**76.7**	3.93	<u>0.211</u>	**77.5**
Stable-Audio-Open [11]	3.85	0.268	68.3	3.50	0.324	57.8
Meric (Ours)	**1.87**	**0.110**	<u>74.2</u>	**2.14**	**0.155**	<u>74.5</u>

Table 3: Vision-music retrieval performance (%, higher is better). Note that MelBench is a one-to-one retrieval task, while ARIA is a one-to-many task (1,000 vision queries → ∼9,000 audio candidates). Recall@K (noted as R@K) is computed as the fraction of ground-truth matches retrieved within the top-K results. I→M denotes vision as query.

	MelBench				*ARIA*			
	I→M		M→I		I→M		M→I	
Method	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
ImageBind [13]	**4.38**	**18.00**	**3.66**	**17.00**	0.30	2.30	0.13	1.77
Meric (Ours)	1.34	8.62	1.12	9.54	**0.40**	**2.80**	**0.39**	**2.40**

Benchmark & Protocol. We construct a 300-task, 6-way forced-choice benchmark across three visual domains (100 MelBench, 100 Unsplash, 100 WikiArt; chance R@1=16.7%). For artistic subsets, all candidates are ARIA-pipeline instrumentals demanding fine-grained discrimination. Gemini 3 Flash scores each image-audio pair (0–100) via chain-of-thought pointwise reasoning; 203 human annotations from 8 annotators serve as ground truth.

Results. As shown in Fig. 2, all evaluators exceed random chance. ImageBind leads on MelBench (76.0%) via pre-trained cross-modal alignment, but Gemini outperforms on the harder artistic domains (34.0% Unsplash, 27.0% WikiArt) by reasoning about abstract correspondences. Gemini achieves the highest overall accuracy (44.0%, MRR=0.644) and follows the same difficulty pattern as humans, validating its use as our Sem evaluator.

4.3 Generation Performance

Vision-to-Music. As shown in Table 1, Meric achieves the lowest FAD_{mert} and FAD_{clap} across all four datasets. For Sem, Meric closely matches or leads cascades on three of four benchmarks while exceeding all direct baselines by >20 points.

Text-to-Music. Because Qwen3VLE encodes both images and text into the same space, Meric supports text-to-music generation without any text-audio

Table 4: Text-music retrieval performance (%, higher is better).

	MusicCaps				*ARIA*			
	T→M		M→T		T→M		M→T	
Method	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
CLAP [49]	**4.31**	**19.48**	**5.54**	**23.53**	0.17	2.30	0.63	3.73
Meric (Ours)	3.27	18.80	2.82	16.42	**0.83**	**5.47**	**0.73**	**6.40**

Table 5: Ablation on Music Head projector architectures. All methods use Qwen3VL-Embedding [25] features. Results of the Music Head are reported without ensemble.

		FAD$_{\text{mert}}$↓		**R@1 / R@10 (%)**↑	
Method	**Loss**	*MelB.*	*ARIA*	*MelB.*	*ARIA*
Linear	Contrastive	12.94	20.04	**1.70** / 9.02	0.70 / **5.50**
Linear	MSE	4.37	7.01	0.82 / 5.56	0.50 / 4.10
BridgeVAE	MSE+KL	4.06	9.35	0.84 / 6.28	0.80 / 3.50
CrossVAE	MSE+KL+Cos	3.48	7.22	0.98 / 5.88	0.80 / 4.30
Music Head	Diffusion	**1.96**	**4.32**	1.52 / **9.30**	0.80 / 4.80

training. As shown in Table 2, Meric achieves the best FAD on both MusicCaps and ARIA, surpassing all dedicated T2M models. Meric's Sem (74.2) also outperforms all baselines except AudioLDM2-Large.

4.4 Retrieval Performance

Since the Music Head projects visual inputs into the music semantic anchor space, Meric can perform cross-modal retrieval via cosine similarity—without contrastive training.

Image-to-Music. Table 3 compares Meric with ImageBind [13]. On MelBench, ImageBind leads owing to its pre-trained cross-modal space; on ARIA (∼9K art-referenced segments), Meric outperforms ImageBind in both directions, suggesting that the anchor space captures artistic correspondences that generic embeddings miss.

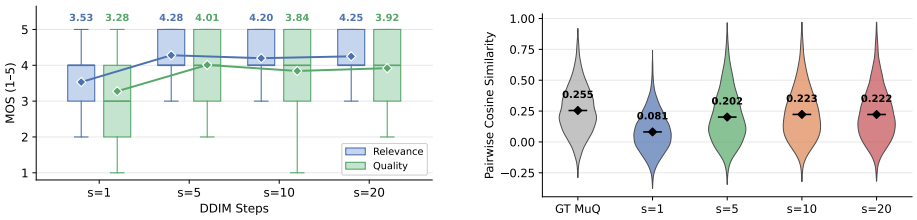
Text-to-Music. Table 4 compares Meric with CLAP [49] (630K text-audio pairs). On MusicCaps, Meric approaches CLAP's T→M recall (R@10: 18.80% vs. 19.48%). On ARIA, Meric substantially outperforms CLAP (T→M R@1: 0.83% vs. 0.17%), demonstrating its superior ability to capture and align fine-grained artistic semantics compared to general-purpose contrastive models.

4.5 Ablative Analysis

Projector Architecture. Table 5 compares five projectors, all using the same frozen Qwen3VLE encoder. The results reveal a *discriminability–fidelity trade-off*: contrastive linear achieves the best retrieval but catastrophic FAD (6.6× the Music Head); MSE linear yields good FAD but poor retrieval by collapsing to

Table 6: Effect of different vision encoders. Qwen3VLE performs better than CLIP.

		$\text{FAD}_{\text{mert}}\downarrow$		R@1 / R@10 (%)↑	
Backbone	**Dim**	MelB.	ARIA	MelB.	ARIA
CLIP-ViT-H-14 [35]	1024	2.54	4.73	**1.60** / **10.00**	0.90 / 4.10
Qwen3VLE [25]	2048	**1.95**	**3.98**	1.52 / 9.30	**0.80** / **4.80**



(a) Human MOS ratings (relevance & quality) **(b)** Intra-group pairwise cosine similarity

Fig. 3: Effect of DDIM steps on Music Head. (a) Human MOS (n=100 images × 4 steps); s=5/10/20 are indistinguishable (p>0.2). (b) Fewer steps produce collapsed embeddings; s=20 matches the GT distribution.

the conditional mean; VAE methods strike a moderate balance but share MSE's deterministic collapse (Sec. 4.5). Only the Music Head resolves this trade-off—best FAD on both benchmarks with competitive retrieval—because iterative denoising produces predictions that are individually accurate yet collectively diverse. Note that to ensure a fair comparison, the Music Head evaluated here uses an earlier version of the pretrained baseline (i.e., trained solely on the 369K pairs without v-prediction) rather than the final Meric model.

Vision Encoder. Table 6 isolates the vision encoder by fixing the Music Head projector. Replacing CLIP ViT-H-14 with Qwen3VLE consistently improves FAD (2.54→1.95 on MelBench), confirming that richer vision-language representations benefit cross-modal translation even with a frozen backbone.

Music Head Sampling Steps. Since the Flow Decoder dominates the acoustic characteristics of the final output, FAD becomes insensitive to the step counts of Music Head; we therefore conduct a human listening study (100 images × 4 step counts, 1–5 MOS). As shown in Fig. 3(a), s=1 is significantly worse (p<0.001), but $s\geq5$ are statistically indistinguishable (p>0.2). The mechanism is revealed by the pairwise cosine similarity distribution of the generated embeddings (Fig. 3(b)): fewer steps produce collapsed, low-diversity embeddings; at s=20 the distribution closely matches GT. We use s=20 as a conservative default, though s=5 suffices at 4× lower latency.

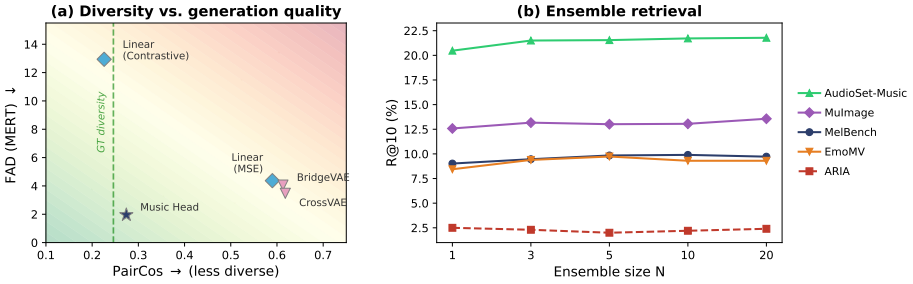


Fig. 4: Benefits of Music Head stochasticity. (a) Projectors in PairCos–FAD space (MelBench). Only the Music Head achieves both GT-level diversity and low FAD. (b) Ensemble retrieval across five benchmarks: N-seed averaging improves R@10 on natural-audio datasets (+4–10%) but not on ARIA.

Stochasticity and Ensemble. The intrinsic uncertainty of human musical creativity means that the same visual inspiration can lead to diverse yet equally valid composition outcomes. The Music Head's iterative denoising elegantly models this by naturally introducing controlled stochasticity: different initial noise vectors yield diverse semantic predictions for the same input.

Output diversity. We measure PairCos (mean pairwise cosine similarity; lower = more diverse). Fig. 4(a) reveals three regimes: (i) *deterministic collapse* (MSE, VAEs; PairCos≈0.61, 2.5× above GT), (ii) *diverse but unfaithful* (contrastive; low PairCos, catastrophic FAD), and (iii) *diverse and faithful* (Music Head; PairCos≈GT, lowest FAD).

Ensemble retrieval. Averaging N seed predictions reduces per-sample noise. Fig. 4(b) shows N=10 improves R@10 by +4–10% on natural-audio datasets but not on ARIA, where the bottleneck is embedding-level discriminability rather than per-sample noise.

5 Conclusion

We present **Meric**, a unified framework for multimodal music generation and retrieval. By introducing a music semantic representation as an anchor, we decouple cross-modal understanding from music synthesis, enabling the two stages to be trained on paired multimodal data and large-scale unimodal music data respectively. This design simultaneously addresses the data scarcity bottleneck, endows the model with explicit one-to-many generation capability, and allows the learned embeddings to serve retrieval without any architectural overhead. Extensive evaluations demonstrate that Meric achieves state-of-the-art acoustic fidelity and semantic alignment in both vision-to-music and text-to-music generation tasks, while also enabling direct zero-shot cross-modal retrieval. Our future work will build upon this Meric framework to jointly generate music, artistic images, and lyric text, moving toward a fully unified multimedia ecosystem.

Acknowledgments

The authors are deeply grateful to Prof. Hongteng Xu for his thoughtful guidance and continuous support. This work is supported by the National Natural Science Foundation of China (No. 62276268).

References

1. Agostinelli, A., Denk, T.I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., et al.: Musiclm: Generating music from text. arXiv preprint arXiv:2301.11325 (2023) 2, 3, 9
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 9, 10
3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023) 1
4. Chen, D., Chen, R., Zhang, S., Wang, Y., Liu, Y., Zhou, H., Zhang, Q., Wan, Y., Zhou, P., Sun, L.: Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In: Salakhutdinov, R., Kolter, Z., Heller, K.A., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. Proceedings of Machine Learning Research, vol. 235, pp. 6562–6595. PMLR / OpenReview.net (2024) 4
5. Chowdhury, S., Nag, S., Joseph, K.J., Srinivasan, B.V., Manocha, D.: Melfusion: Synthesizing music from image and language cues using diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 26816–26825. IEEE (2024) 2, 4, 9
6. Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., Défossez, A.: Simple and controllable music generation. In: NeurIPS (2023) 2, 3, 9, 10, 11
7. DeepMind, G.: Gemini3 (2025), https://blog.google/products-and-platforms/products/gemini/gemini-3-collection/ 1, 8
8. Defferrard, M., Benzi, K., Vandergheynst, P., Bresson, X.: Fma: A dataset for music analysis. arXiv preprint arXiv:1612.01840 (2016) 8
9. Di, S., Jiang, Z., Liu, S., Wang, Z., Zhu, L., He, Z., Liu, H., Yan, S.: Video background music generation with controllable music transformer. In: Shen, H.T., Zhuang, Y., Smith, J.R., Yang, Y., César, P., Metze, F., Prabhakaran, B. (eds.) MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021. pp. 2037–2045. ACM (2021) 4
10. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. In: Salakhutdinov, R., Kolter, Z., Heller, K.A., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. Proceedings of Machine Learning Research, vol. 235, pp. 12606–12633. PMLR / OpenReview.net (2024) 1
11. Evans, Z., Parker, J.D., Carr, C., Zukowski, Z., Taylor, J., Pons, J.: Stable audio open. In: ICASSP. pp. 1–5 (2025) 1, 9, 11

12. Gan, C., Huang, D., Chen, P., Tenenbaum, J.B., Torralba, A.: Foley music: Learning to generate music from videos. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI. Lecture Notes in Computer Science, vol. 12356, pp. 758–775. Springer (2020) 2
13. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind one embedding space to bind them all. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 15180–15190. IEEE (2023) 4, 9, 11, 12
14. Google DeepMind: A new way to express yourself: Gemini can now create music (2026), https://blog.google/innovation-and-ai/products/gemini-app/lyria-3/ 1
15. Hang, T., Gu, S., Li, C., Bao, J., Chen, D., Hu, H., Geng, X., Guo, B.: Efficient diffusion training via min-snr weighting strategy. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 7407–7417. IEEE (2023) 9
16. Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R.J., Wilson, K.W.: CNN architectures for large-scale audio classification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017. pp. 131–135. IEEE (2017) 4, 9
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) NeurIPS (2020) 3, 5
18. Holderrieth, P., Havasi, M., Yim, J., Shaul, N., Gat, I., Jaakkola, T.S., Karrer, B., Chen, R.T.Q., Lipman, Y.: Generator matching: Generative modeling with arbitrary markov processes. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net (2025) 3
19. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research, vol. 97, pp. 2790–2799. PMLR (2019) 4
20. Hu, H., Zhu, X., He, T., Guo, D., Zhang, B., Wang, X., Guo, Z., Jiang, Z., Hao, H., Guo, Z., et al.: Qwen3-tts technical report. arXiv preprint arXiv:2601.15621 (2026) 1
21. Ji, S., Wu, S., Wang, Z., Li, S., Zhang, K.: A comprehensive survey on generative ai for video-to-music generation. arXiv preprint arXiv:2502.12489 (2025) 4
22. Kilgour, K., Zuluaga, M., Roblek, D., Sharifi, M.: Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In: Kubin, G., Kacic, Z. (eds.) 20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019. pp. 2350–2354. ISCA (2019) 4, 8, 9
23. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: Bengio, Y., LeCun, Y. (eds.) 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings (2014) 3
24. Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W.Z., Sotelo, J., de Brébisson, A., Bengio, Y., Courville, A.C.: Melgan: Generative adversarial

networks for conditional waveform synthesis. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. pp. 14881–14892 (2019) 4

25. Li, M., Zhang, Y., Long, D., Chen, K., Song, S., Bai, S., Yang, Z., Xie, P., Yang, A., Liu, D., et al.: Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. arXiv preprint arXiv:2601.04720 (2026) 2, 6, 9, 12, 13
26. Li, Y., Yuan, R., Zhang, G., Ma, Y., Chen, X., Yin, H., Xiao, C., Lin, C., Ragni, A., Benetos, E., Gyenge, N., Dannenberg, R.B., Liu, R., Chen, W., Xia, G., Shi, Y., Huang, W., Wang, Z., Guo, Y., Fu, J.: MERT: acoustic music understanding model with large-scale self-supervised training. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024) 4, 9
27. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net (2023) 3, 5
28. Liu, H., Yuan, Y., Liu, X., Mei, X., Kong, Q., Tian, Q., Wang, Y., Wang, W., Wang, Y., Plumbley, M.D.: Audioldm 2: Learning holistic audio generation with self-supervised pretraining. IEEE ACM Trans. Audio Speech Lang. Process. **32**, 2871–2883 (2024) 1, 2, 3, 9, 10, 11
29. Liu, S., Wu, Q., Sun, C., Hussain, A.S., Shan, Y.: Mumu-llama: Multi-modal music understanding and generation via large language models. Expert Systems with Applications p. 130688 (2025) 2, 4, 9
30. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: ICLR (2023) 5
31. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-eval: NLG evaluation using gpt-4 with better human alignment. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023. pp. 2511–2522. Association for Computational Linguistics (2023) 4
32. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 8162–8171. PMLR (2021) 9
33. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 4172–4182. IEEE (2023) 2, 5
34. Prajwal, K.R., Shi, B., Le, M., Vyas, A., Tjandra, A., Luthra, M., Guo, B., Wang, H., Afouras, T., Kant, D., Hsu, W.: Musicflow: Cascaded flow matching for text guided music generation. In: Salakhutdinov, R., Kolter, Z., Heller, K.A., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. Proceedings of Machine Learning Research, vol. 235, pp. 41052–41063. PMLR / OpenReview.net (2024) 2
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.)

Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021) 4, 13

36. Rinaldi, I., Fanelli, N., Castellano, G., Vessio, G.: Art2mus: Bridging visual arts and music through cross-modal generation. In: Bue, A.D., Canton, C., Pont-Tuset, J., Tommasi, T. (eds.) Computer Vision - ECCV 2024 Workshops - Milan, Italy, September 29-October 4, 2024, Proceedings, Part V. Lecture Notes in Computer Science, vol. 15627, pp. 173–186. Springer (2024) 4, 9, 10
37. Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain. pp. 2226–2234 (2016) 4
38. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022) 9
39. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015) 4
40. Song, J., Wang, Y.: Musflow: Multimodal music generation via conditional flow matching. In: Gurrin, C., Schoeffmann, K., Zhang, M., Rossetto, L., Rudinac, S., Dang-Nguyen, D., Cheng, W., Chen, P., Benois-Pineau, J. (eds.) Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025, Dublin, Ireland, October 27-31, 2025. pp. 10200–10209. ACM (2025) 2, 4
41. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021) 5
42. Su, K., Li, J.Y., Huang, Q., Kuzmin, D., Lee, J., Donahue, C., Sha, F., Jansen, A., Wang, Y., Verzetti, M., Denk, T.I.: V2meow: Meowing to the visual beat via video-to-music generation. In: Wooldridge, M.J., Dy, J.G., Natarajan, S. (eds.) Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada. pp. 4952–4960. AAAI Press (2024) 2
43. Tan, H.H., Herremans, D.: Music fadernets: Controllable music generation based on high-level features via low-level feature modelling. In: Cumming, J., Lee, J.H., McFee, B., Schedl, M., Devaney, J., McKay, C., Zangerle, E., de Reuse, T. (eds.) Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR 2020, Montreal, Canada, October 11-16, 2020. pp. 109–116 (2020) 4
44. Tian, Z., Jin, Y., Liu, Z., Yuan, R., Tan, X., Chen, Q., Xue, W., Guo, Y.: Audiox: Diffusion transformer for anything-to-audio generation. arXiv preprint arXiv:2503.10522 (2025) 4, 9, 10
45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Infor-

mation Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 5998–6008 (2017) 3
46. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025) 1
47. Wang, B., Zhuo, L., Wang, Z., Bao, C., Chengjing, W., Nie, X., Dai, J., Han, J., Liao, Y., Liu, S.: Multimodal music generation with explicit bridges and retrieval augmentation. arXiv preprint arXiv:2412.09428 (2024) 2
48. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025) 1
49. Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., Dubnov, S.: Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In: IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023. pp. 1–5. IEEE (2023) 2, 4, 8, 9, 12
50. Xu, J., Wu, X., Yao, T., Zhang, Z., Bei, J., Wen, W., He, L.: Aesthetic matters in music perception for image stylization: A emotion-driven music-to-visual manipulation. arXiv preprint arXiv:2501.01700 (2025) 9
51. Yin, Z., Li, S., Dong, W.: Video2song: Video-to-song generation via retrieval-augmented prompt generation. In: Komura, T., Noh, J. (eds.) Proceedings of the SIGGRAPH Asia 2025 Posters, SA Posters 2025, Hong Kong, SAR, China, December 15-18, 2025. pp. 15:1–15:2. ACM (2025) 2
52. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net (2025) 2
53. Zhao, Z., Jin, D., Zhou, Z.: Zero-effort image-to-music generation: An interpretable rag-based vlm approach. arXiv preprint arXiv:2509.22378 (2025) 2
54. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. arXiv preprint arXiv:2510.11690 (2025) 2
55. Zheng, L., Chiang, W., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023) 4
56. Zhu, H., Zhou, Y., Chen, H., Yu, J., Ma, Z., Gu, R., Luo, Y., Tan, W., Chen, X.: Muq: Self-supervised music representation learning with mel residual vector quantization. IEEE Transactions on Audio, Speech and Language Processing (2025) 9
57. Zuo, H., You, W., Wu, J., Ren, S., Chen, P., Zhou, M., Lu, Y., Sun, L.: Gvmgen: A general video-to-music generation model with hierarchical attentions. In: Walsh, T., Shah, J., Kolter, Z. (eds.) AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA. pp. 23099–23107. AAAI Press (2025) 2, 4, 9, 10