VMEVAL: A MULTI-DIMENSIONAL AND INTERPRETABLE FRAMEWORK FOR VIDEO-MUSIC HARMONY EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in audio-visual learning have primarily focused on cross-modal generation tasks such as video-to-music generation. However, the field still lacks a systematic and objective framework for evaluating video-music harmony, particularly one that simultaneously supports quantitative scoring and qualitative interpretation. Existing approaches are limited to either global compatibility estimation for cross-modal retrieval or descriptive analysis with multimodal large language models (MLLMs) without targeted supervision for harmony evaluation. In this work, we propose VMEval, a multi-dimensional and interpretable framework for video-music harmony evaluation. We decompose video-music harmony into six structured features including rhythm, emotion, dynamics, structure, atmosphere, and content, which provide complementary views of cross-modal consistency and collectively cover the major perceptual aspects. Based on these features, VMEval learns fine-grained cross-modal consistency and jointly produces quantitative dimension-specific harmony scores and interpretable natural-language explanations. VMEval can be applied to downstream tasks including cross-modal retrieval and the evaluation of video-to-music generation models. Extensive experiments show that VMEval achieves competitive performance on cross-modal retrieval and demonstrates strong agreement with human judgments in evaluating video-music harmony.

1 INTRODUCTION

Background music plays a critical role in shaping the emotional, temporal, and artistic coherence of visual content. The rapid growth of online video platforms and short-form media has dramatically increased the demand for high-quality video soundtrack matching, yet selecting appropriate music for a video remains challenging, particularly for amateur users. Although recent advances in video-to-music generation (Zuo et al., 2025; Tian et al., 2025) have simplified soundtrack creation through automatic music synthesis, the field still lacks a systematic and objective framework for assessing whether a music track is well matched to a given video. Such an evaluation framework is important not only for practical applications in video creation, but also for providing a principled basis for training, comparing, and assessing generative and retrieval-based video-music models.

As shown in Fig. 1, existing research on video-music correspondence has largely focused on learning joint embedding spaces for cross-modal retrieval or recommendation, where compatibility is typically summarized by a single global similarity score. Representative works, such as artistic correspondence learning (Ullah & Choi, 2024; Su et al., 2024), model global compatibility between video and music pairs, but do not provide fine-grained assessments of the perceptual factors underlying the correspondence or interpretable feature-level analysis. More recently, instruction-tuning datasets such as HarmonySet (Zhou et al., 2025) and MLLM-based frameworks (Maaz et al., 2024; Cheng et al., 2024; Tang et al., 2025) have enabled open-ended descriptions of video-music relationships across dimensions such as rhythm, theme, emotion, and cultural relevance. However, these approaches provide descriptive but not quantitative analyzes and do not explicitly learn a calibrated evaluator that models how different aspects jointly contribute to overall harmony. This limits their use for systematic benchmarking and quantitative comparison of video-music pairs.

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.

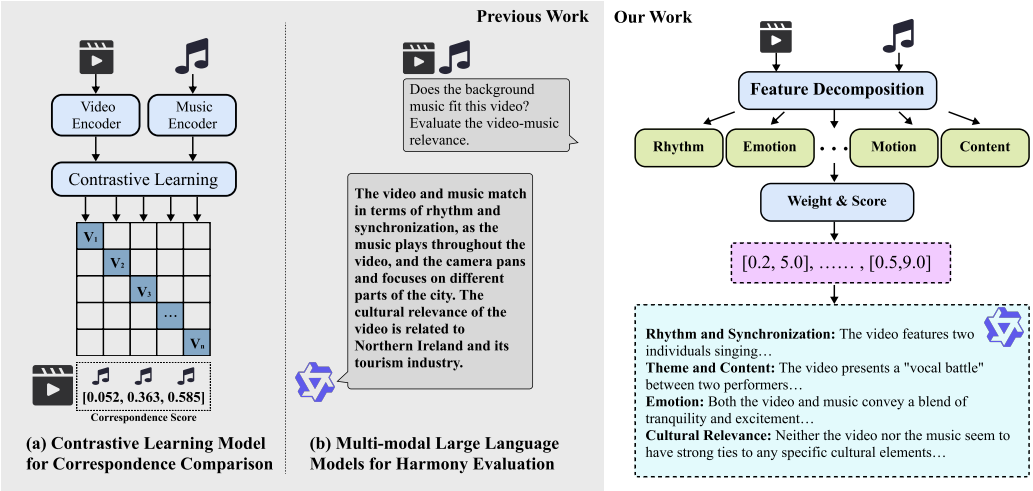

Figure 1: Difference between our work and previous works. Previous works include contrastive learning for correspondence comparison and MLLM for descriptive evaluation, while our work proposes multi-dimensional harmony evaluation with both quantitative and qualitative analysis.

A key challenge is that video-music harmony is multi-dimensional, such as emotional, stylistic, and dynamic correspondence. Moreover, the relative importance of these dimensions varies across video-music pairs: for example, rhythmic alignment may be particularly important in dance or sports videos (Prétet et al., 2021), whereas emotional and atmospheric compatibility may play a larger role in cinematic or narrative scenes. This observation motivates two requirements for a practical harmony evaluator: it should explicitly model fine-grained feature-wise correspondence, while also adapting the relative contribution of different features according to the overall characteristics of the video-music pair. Therefore, we first introduce six features, including rhythm, dynamics, content, emotion, structure, and atmosphere, as complementary components of video-music harmony. These features are designed as a finer-grained representation of perceptual features based on HarmonySet.

Then, we propose **VMEval**, a multi-dimensional and interpretable framework for video-music harmony evaluation. VMEval first projects video and music into six feature-specific subspaces through aligning them with corresponding textual feature descriptions, enabling explicit estimation of feature-wise cross-modal consistency. An adaptive weighting module then assigns pair-specific importance weights to features, allowing the evaluator to account for differences in the dominant perceptual cues across video-music pairs. In addition, a global style module models coarse-grained stylistic compatibility and uses it to regularize the final harmony score beyond local feature-level correspondence. Finally, structured intermediate predictions including feature-wise scores, feature importance weights, and global style assessments, are provided to a LoRA-adapted Qwen-2.5-Omni decoder to generate natural-language explanations of the resulting harmony assessment.

Since existing multimodal datasets do not directly provide the fine-grained feature descriptions, scores, and importance weights, we construct pseudo-supervised process annotations from HarmonySet using Gemini-2.5-Pro (Comanici et al., 2025). These annotations are used only to supervise the intermediate feature decomposition and weighting process, while the overall harmony analysis remains grounded in HarmonySet's original annotations and is further evaluated with independent human judgments. In this way, the pseudo-supervision serves as a structured training signal rather than replacing human evaluation or treating the MLLM annotations as ground-truth judgments.

Extensive experiments demonstrate that VMEval generally improves cross-modal retrieval performance over representative baselines, indicating that explicitly modeling multi-dimensional harmony can produce more discriminative video-music representations. In quantitative harmony evaluation, VMEval achieves the lowest MAE against human judgments among the compared methods. For qualitative analysis, its generated explanations achieve competitive human ranking performance and strong LLM-based evaluation scores across multiple harmony dimensions, while providing a unified interpretation of the underlying quantitative assessment.

Our contributions are summarized as follows:

- We formulate video-music harmony evaluation as a structured multi-dimensional prediction problem and propose **VMEval**, a unified framework that jointly supports quantitative harmony scoring and interpretable qualitative explanation.
- We introduce a fine-grained decomposition of video-music harmony, together with pseudo-supervised feature alignment, adaptive feature weighting, and global style modeling, to explicitly characterize how different perceptual cues contribute to overall harmony.
- Extensive experiments on cross-modal retrieval, human-based quantitative evaluation, and qualitative explanation evaluation demonstrate the effectiveness of VMEval for both objective harmony assessment and interpretable video-music analysis.

2 Related Work

Video-Music Retrieval and Generation. Early research on video-music correspondence primarily focused on learning shared embedding spaces for cross-modal retrieval and recommendation (Surís et al., 2022; Gu et al., 2023; Doh et al., 2026), where the objective is to retrieve music tracks compatible with a given video. These works formulate video-music matching as metric learning or contrastive representation learning, leveraging global semantic alignment between modalities. Recent advances (Dong et al., 2024) further extend this paradigm to personalized and interactive recommendation settings, incorporating user preferences and conversational feedback into the recommendation process. In parallel, video-to-music generation has recently emerged as a fast-growing direction (You et al., 2025; Wu et al., 2025; Li et al., 2024), where generative models aim to synthesize background music conditioned on video semantics (Tian et al., 2025), temporal dynamics (Zuo et al., 2025), and rhythmic cues (Li et al., 2024). While these methods have significantly advanced automatic soundtrack creation, they typically evaluate correspondence through task-specific retrieval or subjective evaluation rather than objective harmony assessment. Our work complements this line of research by introducing a dedicated and systematic video-music harmony evaluation framework that can be applied to both cross-modal retrieval and the assessment of generated music.

Video-Music Correspondence Evaluation. Despite increasing interest in video-music understanding, explicit research on objective audio-visual harmony evaluation remains limited. Existing approaches often rely on retrieval performance for compatibility (Surís et al., 2022; Yang et al., 2024), implicitly assuming that embedding similarity reflects perceptual harmony. Other works (Ullah & Choi, 2024; Su et al., 2024) introduce correspondence scoring and temporal alignment metrics for evaluating generated video-music pairs, yet these methods generally produce only coarse global scores without fine-grained explanation. Such global criteria may also overlook the fact that different aspects of harmony can contribute differently across video-music pairs. Meanwhile, datasets such as HarmonySet (Zhou et al., 2025) and EmoMV (Thao et al., 2023) provide richer language-based annotations of video-music relationships across dimensions like rhythm, theme, and emotion, enabling qualitative analysis of harmony. However, such data remain descriptive rather than evaluative, lacking an evaluator that predicts feature-level consistency, adaptive feature importance, and an overall harmony score. In contrast, our method establishes a unified evaluation framework that jointly performs fine-grained quantitative scoring and interpretable qualitative explanation.

Multimodal Large Language Models for Fine-Grained Cross-Modal Reasoning. Recent MLLMs (Zhang et al., 2025; Sun et al., 2024) have demonstrated strong capabilities in multimodal understanding, reasoning, and instruction following across vision-language and audio-language tasks. By aligning pretrained modality encoders with LLMs (Li et al., 2025; Lin et al., 2024; Lyu et al., 2023), these systems enable open-ended semantic analysis and fine-grained attribute description over complex multimodal inputs. Such capabilities have recently been adopted in music recommendation and harmony understanding tasks to generate textual explanations or semantic annotations for video-music pairs (Carone et al., 2026). However, prior works primarily use MLLMs for description and reasoning, rather than incorporating them into a structured quantitative evaluation pipeline that explicitly models feature-level correspondence and overall harmony. In this work, we leverage MLLMs as pseudo-supervisors, backbones and explanation generators within a structured evaluation framework, enabling them to support fine-grained quantitative harmony assessment.

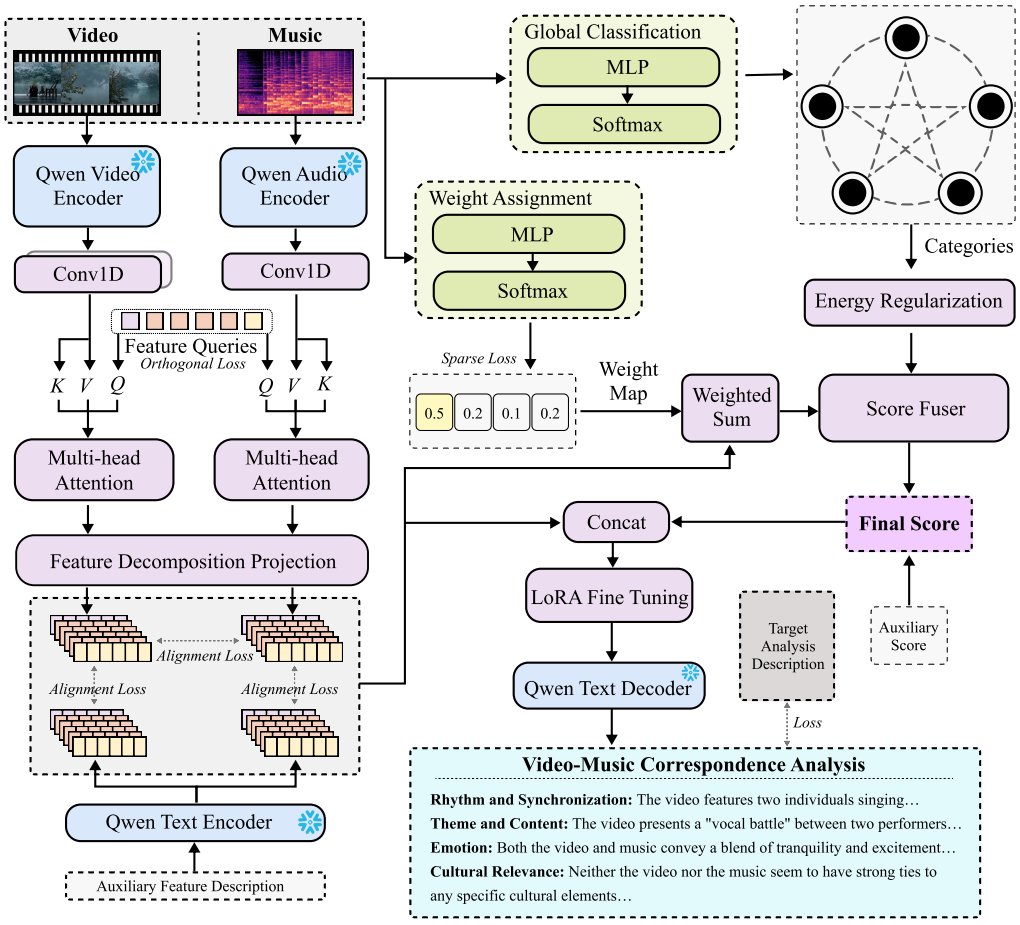

Figure 2: The model architecture of VMEval.

3 Approach

3.1 Problem Definition

Given a video–music pair (V, M), our goal is to learn an audio-visual harmony evaluation model f_θ that predicts both a quantitative harmony score and a qualitative explanation:

$$(\hat{s}, \hat{y}) = f_\theta(V, M), \tag{1}$$

where $\hat{s} \in \mathbb{R}$ denotes the predicted overall harmony score and $\hat{y}$ denotes the generated explanation.

To enable interpretable fine-grained evaluation, we further introduce intermediate variables that represent the harmony assessment process, including feature-level textual descriptions, feature-wise consistency scores and weights. These intermediate variables are treated as intermediate fine-grained outputs of the evaluator and are supervised during training via pseudo annotations.

Given ground-truth harmony analysis y and pseudo-supervised process annotations $\mathcal{A}$, the model is trained to jointly align its predicted score and explanation with the supervision signals:

$$\theta^* = \arg\min_\theta \mathcal{L}\big(f_\theta(V, M),\, y,\, \mathcal{A}\big), \tag{2}$$

where $\mathcal{L}$ combines supervision on the final harmony assessment with auxiliary supervision on the intermediate evaluation process.

3.2 PSEUDO-SUPERVISED DATA ANNOTATION

A major challenge in video-music harmony evaluation is that existing datasets typically provide only overall harmony analysis without intermediate reasoning process. To address this limitation, we construct structured pseudo-supervised annotations that decompose harmony assessment into fine-grained feature-level signals.

Drawing inspiration from established theories in film aesthetics (Wang & Cheong, 2006), cinematic scoring (Hanjalic, 2006; Joshi et al., 2011), and audiovisual art analysis (Vasconcelos & Lippman, 1997; den Heijer, 2012), we identify six semantically structured and complementary features that capture distinct aspects of video–music harmony: *rhythm*, *dynamics*, *content*, *emotion*, *structure*, and *atmosphere*. These features are finer-grained based on the complementary aspects of synchronization, semantic content, emotion, and cultural relevance in HarmonySet. We employ Gemini-2.5-Pro as a multimodal annotator. Given a video-music pair (V, M), Gemini generates a textual auxiliary description t_k for each visual and musical feature k that we have defined. To reduce semantic overlap, the annotation prompt instructs Gemini to focus each description on its designated feature.

In addition, we convert HarmonySet's original free-form harmony analyses into structured numerical pseudo labels. Conditioned on the original HarmonySet annotation and the corresponding video-music pair, Gemini predicts for each feature a consistency score $s_k \in [0, 10]$ and an importance weight $w_k \in [0, 1]$. The resulting process annotation is defined as:

$$\mathcal{A}(V, M) = \{(t_k, s_k, w_k)\}_{k=1}^{6}. \tag{3}$$

These annotations provide supervision only for the intermediate evaluation process and are not treated as ground-truth overall harmony judgments. The final harmony analysis is instead supervised by HarmonySet's original annotations and independently evaluated by human judgments. To assess the quality of the pseudo annotations, we randomly select 200 samples and ask three users per sample to judge whether the generated annotations match the underlying video-music pair. A sample is considered valid when at least two users approve the annotation, yielding a 96% sample-level pass rate and a 92.2% overall pass rate across these 600 judgments.

3.3 MULTI-DIMENSIONAL QUANTITATIVE EVALUATION

Given a video–music pair, we first extract modality-specific embeddings $(\tilde{H}_v, \tilde{H}_m)$ using frozen pretrained Qwen-2.5-Omni (Xu et al., 2025) encoders. Since these representations are temporally long and high-dimensional, we apply lightweight 1D-convolutional layers to compress temporal and channel dimensions, producing compact audiovisual token sequences for feature decomposition.

We then decompose the compressed multimodal representations into the six structured feature subspaces. Inspired by the Q-Former design (Li et al., 2022; 2023), we introduce six learnable query tokens, each corresponding to one harmony feature. These query tokens attend to the compressed audiovisual embedding sequences through cross-attention (Vaswani et al., 2017), aggregating modality-specific information relevant to their designated feature. The attended outputs are projected through feature-specific linear layers to obtain feature-level embeddings:

$$z_v^{(k)} = W_k^v \mathrm{CrossAttn}(q_k, \tilde{H}_v), \qquad z_m^{(k)} = W_k^m \mathrm{CrossAttn}(q_k, \tilde{H}_m), \tag{4}$$

This operation maps the modality-specific representations into structured feature subspaces, allowing different harmony features to be modeled separately. To reduce redundancy among feature queries, we further impose an orthogonality constraint:

$$\mathcal{L}_{\text{ortho}} = \|Q \cdot Q^T - I\|_2^2 \tag{5}$$

To ensure that each projection captures the decomposed feature correctly, we align the projected embeddings with the embeddings of the pseudo-supervised textual feature descriptions. Specifically, we encode the auxiliary textual descriptions using the frozen Qwen text encoder and minimize mean squared error between projected embeddings and corresponding text embeddings:

$$\mathcal{L}_{\text{feat}} = \sum_{k=1}^{6} \left(\|\hat{z}_v^{(k)} - t_v^{(k)}\|_2^2 + \|\hat{z}_m^{(k)} - t_m^{(k)}\|_2^2 \right). \tag{6}$$

After semantic alignment, we compute the harmony consistency of each feature via cosine similarity between video and music feature embeddings:

$$\hat{s_k} = \cos(\hat{z}_v^{(k)}, \hat{z}_m^{(k)}) = \frac{\hat{z}_v^{(k)} \cdot \hat{z}_m^{(k)}}{\|\hat{z}_v^{(k)}\| \|\hat{z}_m^{(k)}\|}. \tag{7}$$

Adaptive Feature Weight Assignment. Although harmony is jointly determined by multiple features, their relative contributions can vary substantially across video-music pairs. Some features may dominate the perceptual judgment for a given artistic pairing, while others provide supporting cues. For example, rhythmic synchronization may dominate in dance videos, whereas emotional resonance may be more important in cinematic narratives. Motivated by this observation, we predict pair-specific feature importance weights conditioned on the global audiovisual content:

$$\hat{w} = \mathrm{Softmax}(\mathrm{MLP}([\bar{H}_v; \bar{H}_m])). \tag{8}$$

The predicted weights and scores are additionally aligned with the corresponding pseudo-supervised annotations. Furthermore, to encourage a sparse but non-degenerate feature weighting pattern, we impose an entropy minimization regularizer over the predicted weight distribution:

$$\mathcal{L}_{\text{weight}} = -\lambda_{sparse} \sum_{k=1}^{6} \hat{w}_k \log(\hat{w}_k + \epsilon) - \lambda_{sharp} \sum_{k=1}^{6} \log(\hat{w}_k + \epsilon), \tag{9}$$

The first term minimizes the entropy of distribution and encourages a sparse assignment in which a small number of features can dominate the harmony judgment. The second term acts as a logarithmic barrier against vanishing weights, preventing the distribution from becoming overly concentrated and preserving the contribution of supporting features. The resulting weights aggregate the feature-wise consistency scores into a preliminary harmony estimate like $s_{\text{base}} = \sum_{k=1}^{6} \hat{w}_k \hat{s}_k$.

Global Style Energy Regularization. Feature-wise correspondence alone may overlook coarse-grained stylistic compatibility between the overall artistic expressions of video and music. To complement the local feature modeling, we introduce a global style evaluator that categorizes video and music into five prototypical expressive style archetypes inspired by common aesthetic taxonomies in film and music theory (Nelson, 1979; Chen, 2025): *vitality*, *intensity*, *serenity*, *grandeur*, and *severity*. These styles describe broad expressive tendencies such as energetic motion, dramatic tension, contemplative calmness, majestic expansiveness, and ambiguous abstraction.

The model predicts soft style assignments for video and music, denoted by $p_v, p_m \in \Delta^5$, and learns a style interaction energy matrix $R \in \mathbb{R}^{5\times5}$ capturing pairwise compatibility between style archetypes. The resulting global style energy is computed as:

$$e_{\text{style}} = p_v^\top R p_m. \tag{10}$$

Finally, we complement feature-level correspondence with a coarse-grained global style refinement using global style energy as a regularization term:

$$s_{\text{final}} = s_{\text{base}} + \lambda \tanh(e_{\text{style}}), \tag{11}$$

which will be scaled up 10 times to approximate $\hat{s}$.

3.4 Interpretable Qualitative Evaluation

Beyond quantitative scoring, we aim to generate human-interpretable explanations of the harmony assessment process. We therefore condition a language decoder on the structured intermediate outputs of the quantitative evaluator, including feature-wise consistency scores, feature importance weights, global style predictions, and compressed audiovisual representations:

$$X_{\text{prompt}} = \mathrm{Concat}\Big(\{s_k\}_{k=1}^{6}, \{w_k\}_{k=1}^{6}, p_v, p_m, \bar{H}_v, \bar{H}_m\Big), \tag{12}$$

which summarizes the quantitative evaluation process. The prompt is then fed into a Qwen text decoder to generate the harmony analysis:

$$\hat{y} = D_{\text{Qwen}}(X_{\text{prompt}}). \tag{13}$$

To improve explanation quality and domain-specific reasoning, we fine-tune the decoder using Low-Rank Adaptation (LoRA) (Hu et al., 2021). The decoder is instructed to act as a professional audiovisual analyst and organize the analysis into four perspectives aligned with HarmonySet: rhythmic synchronization (corresponding to *rhythm* and *dynamics*), thematic/semantic coherence (*content*), emotional resonance (*emotion*), and cultural/aesthetic compatibility (*structure* and *atmosphere*).

We supervise the generated explanation using HarmonySet's textual annotations via autoregressive cross-entropy loss, where y denotes the HarmonySet ground-truth explanation text:

$$\mathcal{L}_{\text{text}} = -\sum_{t=1}^{N} y_t \log P(\hat{y}_t | X_{\text{prompt}}, \hat{y}_{<t}), \tag{14}$$

By conditioning explanation generation on structured intermediate reasoning variables rather than raw audiovisual embeddings alone, the resulting explanations are explicitly grounded in the quantitative evaluation process, providing a more traceable interpretation of the final harmony score.

4 Experiments

4.1 Implementation Details

We build our framework upon the pretrained Qwen-2.5-Omni-7B model, using its video and audio encoders and text decoder. The hidden size of the original embeddings is 3584. For feature alignment, video embeddings are temporally compressed using two-layer 1D-convolution with $(k, s, p) = (7, 2, 3)$ and $(5, 2, 2)$, respectively, reducing the hidden dimension from 3584 to 1024 and then 512. Audio embeddings are compressed using a single 1D convolution layer with kernel/stride/padding $(5, 2, 2)$ and output dimension 512. Feature decomposition is implemented via 4-head cross-attention with hidden size 512 for all six learnable feature queries. For correspondence modeling, the feature weight assignment module is implemented as a two-layer MLP with hidden size 512, while the global style classifier uses a two-layer MLP with hidden size 1024. In the explanation stage, we fine-tune the projection layers of the Qwen-2.5-Omni text decoder using LoRA with rank 16 and scaling factor $\alpha = 32$. We trained our model in a single NVIDIA-A800 for 10 epochs.

4.2 Dataset, Baselines & Metrics

Dataset. We conduct our experiments primarily on HarmonySet (Zhou et al., 2025), a benchmark for video-music harmony understanding that provides paired video-music samples with textual annotations describing correspondence across multiple dimensions. To examine generalization ability, we further evaluate it on a new video–music emotional correspondence dataset EmoMV (Thao et al., 2023), which is unseen for VMEval. We use HarmonySet for both quantitative and qualitative evaluation and EmoMV as an additional benchmark for objective video-to-audio retrieval.

Baselines. We compare our method against representative retrieval-based and MLLM-based video–music correspondence models (re-trained on the same HarmonySet):

- **GVMGen Evaluator**: The evaluation model proposed in GVMGen, which adopts an MVpt-style cross-modal framework to estimate global video-music correspondence.
- **VideoLLaMA2 & VideoLLaMA3**: A general-purpose MLLM with strong temporal video understanding capability, adapted here for direct harmony assessment via prompting.
- **Video-Salmonn2**: A recent unified audio-visual language model capable of jointly processing video and audio streams for multimodal reasoning and generation.
- **Qwen-2.5-Omni** & **Gemini-2.5-Pro**: The underlying multimodal foundation model used by VMEval and the pseudo-supervised annotator of VMEval.

Evaluation Metrics. For objective evaluation, we focus on video-to-audio (V2A) retrieval, where a video is used to retrieve its corresponding music. We randomly sample 1,000 video-music pairs from each test set using a fixed random seed (40 in a batch) and evaluate the retrieval rankings. Following [39], we report Recall@K, Median Rank (MedR) and Average Rank (AvgR) metrics. Recall@K

Table 1: Results of V2A retrieval objective evaluation, where MedR and AvgR represent median rank and average rank. Each number of recall rate is a percentage.

Model	HarmonySet						EmoMV					
	MedR↓	AvgR↓	R@1↑	R@5↑	R@10↑	R@20↑	MedR	AvgR	R@1	R@5	R@10	R@20
GVMGen-Eval	19.5	19.68	3.3	14.4	27.5	52.6	19	19.75	2.8	13.8	27.5	53.0
VideoLLaMA2	21	20.58	**7.5**	16.4	27.2	49.5	20.5	20.50	2.5	12.5	25.0	50.0
VideoLLaMA3	21	20.78	2.4	11.7	23.5	48.7	20	20.10	**4.1**	14.4	25.3	50.6
Salmonn2	21	20.40	2.7	14.3	27.6	49.6	20	19.78	2.6	13.1	26.5	52.4
Qwen-2.5-Omni	18	18.40	3.5	15.9	30.1	56.0	18	18.55	3.3	15.1	29.2	57.6
Gemini-2.5-pro	22	20.90	2.1	12.4	25.2	48.2	21	20.94	2.4	11.5	23.7	48.2
VMEval	**15**	**16.82**	5.5	**21.8**	**39.4**	**63.6**	**16**	**17.76**	**4.1**	**17.4**	**32.7**	**59.8**

Table 2: Results of A2V retrieval supplementary experiment in HarmonySet.

Model	MedR↓	AvgR↓	R@1↑	R@5↑	R@10↑	R@20↑
GVMGen-Eval	18	18.91	2.9	14.5	29.2	55.7
VideoLLaMA2	20	20.27	2.7	13.5	25.7	51.0
VideoLLaMA3	21	20.34	3.5	12.4	24.9	48.2
Salmonn2	20	20.16	3.6	13.5	26.9	50.7
Qwen-2.5-Omni	**18**	**18.33**	3.4	15.4	**29.8**	**58.1**
Gemini-2.5-Pro	21	20.83	1.8	10.4	24.4	50.0
VMEval	**18**	18.45	**3.9**	**17.0**	**29.8**	57.0

measures whether the ground-truth music appears among the top-K retrieved candidates, while MedR and AvgR measure the median and average rank of the ground-truth pair. Furthermore, although the practical goal of video-music matching is primarily to select appropriate music for a given video rather than selecting video for music, we treat A2V retrieval as a supplementary analysis to examine whether the learned correspondence is also effective in the reverse direction.

Since there is currently no widely accepted objective metric that directly measures video-music harmony, we use these retrieval metrics as a practical and commonly adopted quantitative proxy. High-quality harmony-aware representations should assign higher similarity to compatible video-music pairs and therefore facilitate their retrieval. However, retrieval performance alone cannot fully characterize the perceptual quality or multi-dimensional reasoning ability of a harmony evaluator. Therefore, we also conduct complementary subjective and LLM-based evaluation.

For quantitative evaluation, we randomly select 80 HarmonySet video-music pairs, with each sample annotated by 10 human raters. The average rating is used as the reference score, and we report the MAE between the model prediction and the human judgment. For qualitative evaluation, we select 40 HarmonySet test samples and ask all models to generate harmony analyses using the same prompts. 20 participants rank the generated descriptions according to correctness and reasonableness, with MedR and AvgR are reported. Following Zhou et al. (2025), we additionally use GPT-5.1 (Singh et al., 2025) to score 90 generated descriptions from each model on a 10-point scale evaluation.

4.3 Experimental Results

Objective Evaluation. From Table 1, we can find that: 1) VMEval achieves the best rank performance and high-recall retrieval performance on both datasets. On HarmonySet, it reduces MedR from 18.0 to 15.0 and AvgR from 18.40 to 16.82 compared with the strongest baseline, while improving R@5/R@10/R@20 to 21.8%/39.4%/63.6%. Similar gains are observed on EmoMV; 2) The improvements consistently hold across both datasets and remain substantial over both Qwen-2.5-Omni and Gemini-2.5-Pro, indicating that the gains are not merely attributable to the underlying backbone or direct prompting of general-purpose MLLMs, but benefit from VMEval's structured multi-dimensional harmony modeling and adaptive weighting. The consistent gains on EmoMV further demonstrate that the learned correspondence generalizes beyond the training benchmark; 3) As shown in Table 2, the supplementary A2V experiment on HarmonySet also shows competitive performance, with VMEval achieving the best R@1/R@5 and tied-best R@10, although the gains are smaller than those observed in the V2A setting. Overall, these results support the effectiveness of VMEval on retrieval-based objective evaluation, providing initial evidence that VMEval learns representations useful for video-music harmony evaluation.

Table 3: Results of ablation study, where FD stands for feature decomposition.

Model	V2A						A2V					
	MedR↓	AvgR↓	R@1↑	R@5↑	R@10↑	R@20↑	MedR	AvgR	R@1	R@5	R@10	R@20
w.o all FD	**15**	17.32	4.7	19.8	36.0	62.6	18	19.08	3.0	14.0	27.6	56.0
w.o video FD	16	17.85	4.2	19.0	33.6	59.2	18	18.83	3.7	16.2	29.0	56.6
w.o music FD	16	17.70	4.6	19.6	34.6	61.5	18	18.91	3.5	14.7	29.2	55.7
w.o $\mathcal{L}_{sparse}$	16	17.49	5.1	19.3	33.9	61.9	19	19.13	2.1	13.7	28.8	55.8
w.o $\mathcal{L}_{ortho}$	16	17.47	4.2	19.5	35.6	60.0	18	18.70	3.5	15.1	**30.4**	55.8
avg weight	17	17.87	4.4	17.6	33.6	60.1	18	18.86	2.7	15.2	28.1	55.9
w.o style	20	20.38	2.7	12.9	25.2	50.5	20	20.44	2.4	12.7	25.3	50.2
VMEval	**15**	**16.82**	**5.5**	**21.8**	**39.4**	**63.6**	**18**	**18.45**	**3.9**	**17.0**	<u>29.8</u>	**57.0**

Table 4: Results of subjective/LLM evaluation, where Rhy., Syn., Con., Emo., and Cul. stand for Rhythm, Synchronization, Content, Emotion and Culture. GVMGen cannot provide textual analysis.

Model	MAE↓	MedR↓	AvgR↓	GPT Score			
				Rhy.& Syn.↑	Theme&Con.↑	Emo.↑	Cul.↑
GVMGen-Eval	4.55±0.62	-	-	-	-	-	-
VideoLLaMA2	2.96±0.44	3	2.61	4.90±0.56	5.17±0.43	5.16±0.43	6.11±0.58
VideoLLaMA3	3.13±0.49	4	3.31	5.79±0.52	4.33±0.46	5.03±0.44	6.84±0.60
Salmonn2	2.94±0.48	**2**	**1.91**	6.68±0.46	7.17±0.40	**6.83±0.36**	7.77±0.46
VMEval	**1.90±0.33**	**2**	<u>2.17</u>	**7.19±0.41**	**7.31±0.41**	<u>6.56±0.43</u>	**8.26±0.39**

Subjective/LLM Evaluation. The results of subjective and LLM-based evaluation are presented in Table 4, which shows that: 1) VMEval achieves the lowest MAE (1.90), significantly outperforming all baselines. This indicates that our model produces harmony scores that are more consistent with human judgments; 2) In the ranking-based evaluation, VMEval achieves competitive performance comparable to Salmonn2, while showing a small gap in AvgR. This indicates that the generated explanations are generally consistent with human preferences, although fine-grained preference remains challenging; 3) In LLM evaluation, VMEval achieves the highest scores in most fine-grained features while remains competitive in emotion, which suggest that our method can still demonstrate balanced performance across multiple aspects of video-music harmony. Overall, VMEval provides strong quantitative agreement with human judgments while maintaining competitive explanation quality across multiple harmony dimensions.

Ablation Study. We conduct ablation studies to analyze the contribution of each component in VMEval, as summarized in Table 3: 1) Removing video or music feature decomposition (w.o. FD) consistently degrades ranking and recall performance, confirming the benefit of explicitly modeling structured harmony features; 2) Replacing adaptive weighting with uniform averaging (avg weight) leads to consistent degradation, highlighting the importance of modeling pair-specific feature contributions. Removing the sparsity or orthogonality constraints (w.o. L_{sparse} / w.o. L_{ortho}) also reduces performance, supporting the role of structured feature weighting and non-redundant feature representations; 3) Removing global style regularization (w.o. style) causes a substantial performance drop across metrics (e.g., V2A R@10 decreases from 39.4% to 25.2%), demonstrating that coarse-grained style compatibility provides complementary information beyond feature-level correspondence. As a supplementary analysis, the A2V results show similar component-level trends, further supporting the robustness of the proposed design in the reverse retrieval direction.

5 Conclusion

In this paper, we present **VMEval**, a multi-dimensional and interpretable framework for video-music harmony evaluation that jointly supports quantitative scoring and qualitative explanation. By introducing pseudo-supervised annotations, feature decomposition and alignment, adaptive weighting, and global style regularization, VMEval explicitly models the structured reasoning process. Extensive experiments demonstrate that our method generally achieves competitive or superior performance over existing approaches in both quantitative harmony evaluation and qualitative explanation generation, with strong alignment to human aesthetic preferences.

AI USE STATEMENT

Gemini-2.5-Pro was used as a multimodal pseudo-annotator to produce the intermediate feature descriptions and numerical pseudo-labels described in Section 3.2. Qwen-2.5-Omni was used as the pretrained multimodal backbone and its decoder was adapted with LoRA as part of the proposed method. GPT-5.1 was used as an evaluator to judge whether the generated texts match the ground truth. We also used GPT-5.6 to aid and polish our writing. Generative AI tools were not used to replace human judgments in the subjective study. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Brandon James Carone, Iran R Roman, and Pablo Ripollés. Llms can read music, but struggle to hear it: An evaluation of core music perception tasks. In *1st International Workshop on Emerging AI Technologies for Music*, 2026.

Linxia Chen. Chinese mythology and the reconstruction of contemporary chinese film aesthetics. *Journal of Chinese Film Studies*, 5(2):261–280, 2025.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Eelco den Heijer. Evolving art using measures for symmetry, compositional balance and liveliness. *IJCCI*, 12:52–61, 2012.

Seungheon Doh, Minhee Lee, Sangmoon Lee, Ben Sangbae Chon, and Juhan Nam. Multimodal video-to-music recommendation via semantic retrieval and temporal reranking. *arXiv preprint arXiv:2607.05971*, 2026.

Zhikang Dong, Xiulong Liu, Bin Chen, Pawel Polak, and Peng Zhang. Musechat: A conversational music recommendation system for videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12775–12785. IEEE, 2024.

Xin Gu, Yinghua Shen, and Chaohui Lv. A dual-path cross-modal network for video-music retrieval. *Sensors*, 23(2):805, 2023.

Alan Hanjalic. Extracting moods from pictures and sounds: Towards truly personalized tv. *IEEE Signal Processing Magazine*, 23(2):90–100, 2006.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z Wang, Jia Li, and Jiebo Luo. Aesthetics and emotions in images. *IEEE Signal Processing Magazine*, 28(5):94–115, 2011.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *Science China Information Sciences*, 68(10):200102, 2025.

Sizhe Li, Yiming Qin, Minghang Zheng, Xin Jin, and Yang Liu. Diff-bgm: A diffusion model for video background music generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27338–27347. IEEE, 2024.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*, 2023.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

Thomas Allen Nelson. Film styles and film meanings. *Film Criticism*, 3(3):2–17, 1979.

Laure Prétet, Gael Richard, and Geoffroy Peeters. Cross-modal music-video recommendation: A study of design choices. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9. IEEE, 2021.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Zhibin Su, Yiming Feng, Jinyu Liu, Jing Peng, Wei Jiang, and Jingyu Liu. An audiovisual correlation matching method based on fine-grained emotion and feature fusion. *Sensors*, 24(17):5681, 2024.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*, 2024.

Dídac Surís, Carl Vondrick, Bryan Russell, and Justin Salamon. It's time for artistic correspondence in music and video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10554–10564. IEEE, 2022.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025.

Ha Thi Phuong Thao, Gemma Roig, and Dorien Herremans. Emomv: Affective music-video correspondence learning datasets for classification and retrieval. *Information Fusion*, 91:64–79, 2023.

Zeyue Tian, Zhaoyang Liu, Ruibin Yuan, Jiahao Pan, Qifeng Liu, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. Vidmuse: A simple video-to-music generation framework with long-short-term modeling. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18782–18793. IEEE, 2025.

Ubaid Ullah and Hyun-Chul Choi. Muim: Analyzing music–image correlations from an artistic perspective. *Applied Sciences*, 14(23):11470, 2024.

Nuno Vasconcelos and Andrew Lippman. Towards semantically meaningful feature spaces for the characterization of video content. In *Proceedings of International Conference on Image Processing*, volume 1, pp. 25–28. IEEE, 1997.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Hee Lin Wang and Loong-Fah Cheong. Affective understanding in film. *IEEE Transactions on circuits and systems for video technology*, 16(6):689–704, 2006.

Junxian Wu, Weitao You, Heda Zuo, Dengming Zhang, Pei Chen, and Lingyun Sun. Controllable video-to-music generation with multiple time-varying conditions. In *Proceedings of the 33rd ACM international conference on multimedia*, pp. 10427–10436, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

Kaixing Yang, Xukun Zhou, Xulong Tang, Ran Diao, Hongyan Liu, Jun He, and Zhaoxin Fan. Beatdance: A beat-based model-agnostic contrastive learning framework for music-dance retrieval. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pp. 11–19, 2024.

Weitao You, Heda Zuo, Junxian Wu, Dengming Zhang, Zhibin Zhou, and Lingyun Sun. Spatial-temporal decomposition and alignment in controllable video-to-music generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10278–10286, 2025.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Zitang Zhou, Ke Mei, Yu Lu, Tianyi Wang, and Fengyun Rao. Harmonyset: A comprehensive dataset for understanding video-music semantic alignment and temporal synchronization. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3152–3162. IEEE, 2025.

Heda Zuo, Weitao You, Junxian Wu, Shihong Ren, Pei Chen, Mingxu Zhou, Yujia Lu, and Lingyun Sun. Gvmgen: A general video-to-music generation model with hierarchical attentions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 23099–23107, 2025.

A AI Prompts and Evaluation Protocols

This appendix records the prompts used to construct pseudo-labels, obtain model explanations, and evaluate explanation quality. The prompts are reported to make the data-generation and evaluation pipeline reproducible. Placeholder fields are shown literally as they were passed to the evaluator.

A.1 Feature decomposition prompt

Role: You are an expert multimodal analyst. You are provided with a video file that contains both visual footage and an integrated audio track.

Task: Perform a "Single-Modality" analysis by decoupling the visual and auditory streams. You must provide two distinct sets of descriptions (one for Video Content and one for Background Music) across 6 orthogonal dimensions each.

Strict Orthogonality Requirement: - When analyzing Video, focus exclusively on visual cues as if the sound were muted.

When analyzing Audio, focus exclusively on musical/sonic properties as if the screen were blank.

Do not cross-reference for 6 dimensions. For example, do not mention "the dancer's steps" in the Audio-Rhythm section; instead, describe the "percussive transients" or "BPM".

PART 1: VISUAL MODALITY ONLY (Muted Analysis) Emotion: Affective state (Valence, Arousal, Dominance) based strictly on facial expressions, lighting, and visual mood.

Spatiotemporal: accurate timing of cuts, storyboard transitions, and the physical amplitude/energy of visible motion.

Pacing: Narrative speed and visual storytelling density, independent of the music's tempo.

Atmosphere: Visual tension, color weight (heavy/light), color temperature (warm/cold), and stylistic affinity.

Structure: Visual composition, framing, rule of thirds, and geometric arrangement of elements.

Content: Thematic elements, specific objects, actors, and the literal narrative plot.

PART 2: AUDIO MODALITY ONLY (Blank-Screen Analysis) Emotion: Affective state (Valence, Arousal) conveyed solely by harmony, mode (major/minor), and melodic contour.

Spatiotemporal: Directional sense of the melody (ascending/descending) and the dynamic range/fluctuation over time (accurate to milliseconds).

Rhythm: Musical meter, tempo (BPM), syncopation, and the "groove" of the rhythmic arrangement.

Atmosphere: Timbre (warm/cold, bright/dark), textural density, and the perceived acoustic space (reverb/dryness).

Structure: Harmonic richness, instrumentation layers, and the formal arrangement (e.g., introduction, climax).

Content: Primary melodic motifs, sound constituents (synthesizers vs. acoustic), and the arrangement's core elements.

Output Format Requirements:

First, output the 6 video dimensions.

Then, output the 6 audio dimensions.

Each line must follow: "Dimension Name: [Detailed Description]".

Use a single newline (Enter) between each dimension.

Total 12 lines of output. No conversational filler.

Example Output: Video-Emotion: ... Video-Spatiotemporal: ... ... (all 6 video dims) Audio-Emotion: ... Audio-Spatiotemporal: ... ... (all 6 audio dims)

A.2 Pseudo-label scoring and weighting prompt

Task: Based on the provided video and its expert textual analysis, quantify the video-music alignment across four dimensions in the text.

Core Objective: Translate the qualitative descriptions in the text into quantitative scores and importance weights.

Scoring & Weighting Rules:

Dimension Scores (1-10): An integer representing alignment quality. (1: Poor/Mismatched; 10: Perfect/Harmonious). Ensure the score strictly reflects the tone of the provided analysis text.

Weights (0.0-1.0): A one-decimal float representing the dimension's importance to the overall coherence of this specific video and textual summary.

High Variance Required: Significantly higher weights for dominant/critical dimensions; weights can be 0.0 if a dimension is irrelevant (e.g., Neither the video nor the music seem to have strong ties to any specific cultural elements).

Constraint: The sum of all four weights must equal exactly 1.0.

Dimensions: Rhythm and Synchronization, Theme and Content, Emotion, Culture Relevance

Output Format: (Strictly 5 lines) Line 1-4 Format: "Dimension: [Weight] [Score]" -> Rhythm: [W] [S] Theme: [W] [S] Emotion: [W] [S] Culture: [W] [S] Line 5 Format: "Overall: [Score]".

A.3 HARMONY ANALYSIS GENERATION PROMPT

Role: You are a senior Audio-Visual Critic.

Task: Analyze the provided video and audio. Evaluate their congruence across four dimensions. Keep each dimension's analysis concise and precise.

Analysis Dimensions (brief):

1. Rhythm & Synchronization: Do visual cuts/movements align with audio beats/phrasing? Note successes or failures (use timestamps briefly).

2. Theme & Content: Does the music's genre/style/lyrics match the visual narrative? Any conceptual mismatch?

3. Emotion: Do the emotional tones of visuals and audio reinforce or conflict with each other?

4. Cultural Relevance: Any meaningful cultural interplay between audio and visual markers? Or is it culturally neutral?

Output Requirements: - Professional, critical, concise and brief. Use the four headings with no more than two sentences for each one. - Support claims with brief visual descriptions or timestamps.

A.4 LLM SCORING EVALUATION PROMPT

Please evaluate the similarity between the following response and the correct answer in terms of their semantic meaning or the true intent they express.

Rate these 4 aspects respectively on a scale of 1 to 10, where a higher score indicates better semantic similarity: 1. Rhythm and Synchronization 2. Thematic Content 3. Emotional Impact 4. Cultural Relevance

Please output a single line containing only four values, separated by commas, in the order above. Do not output any words other than the four scores. Example output: 8,9,7, 6

User Response: {response}

Correct Answer: {correct_answer}

B DISTRIBUTIONAL ANALYSIS OF EVALUATION EXPERIMENTS

We further visualize the distribution of evaluation score used in the two subjective/LLM evaluation studies.

B.1 LLM-ASSIGNED FOUR-DIMENSIONAL SCORES

Figure 3 reports the score distributions for Rhythm, Theme, Emotion, and Culture in LLM evaluation. The distributions also show that VMEval avoids the very low central scores seen for VideoLLaMA2 and VideoLLaMA3, although its long tails indicate difficult clips remain. The similar distribution with a higher average score also shows that VMEval keeps superior or competitive performance compared with Salmonn2.

B.2 MAE TO HUMAN MEAN SCORES

Figure 4 shows the distribution of subjective MAE evaluation. VMEval has the lowest mean MAE (1.90) and median MAE (1.60), compared with 2.94–3.13 for the three MLLM baselines and 4.55 for GVMGen. Its narrower distribution (population standard deviation 1.45) indicates that the improvement is not caused by a small number of easy clips alone.

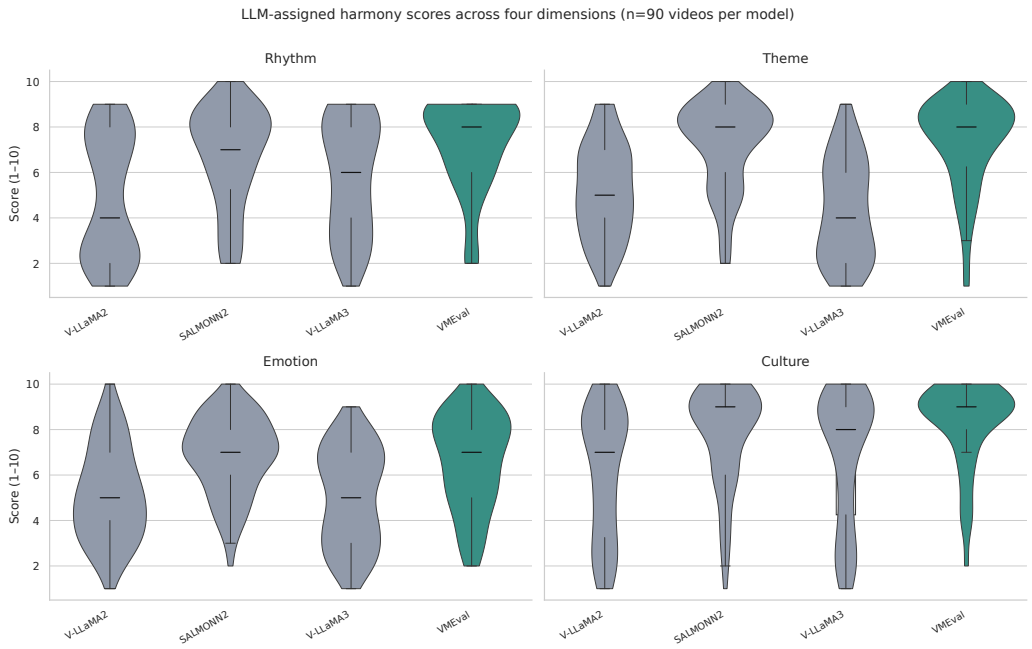

Figure 3: Per-video LLM-assigned scores for the four harmony dimensions. Violin shapes show the empirical distribution and the inset box marks the interquartile range and median.

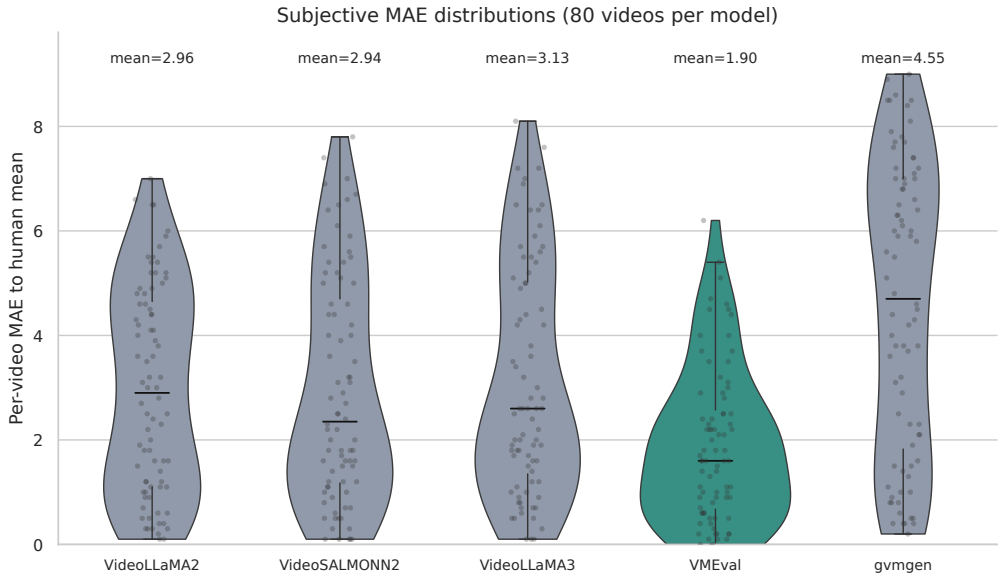

Figure 4: Per-video MAE distributions against the human mean score. Lower is better; points show individual videos and annotations show the mean.

C Additional Ablation Study

Although we introduce a global style evaluator to categorize video and music into five prototypical expressive style archetypes according to (Nelson, 1979; Chen, 2025), we also conduct an ablation study of the number of styles for validation, since the process is unsupervised. As shown in 5, VMEval with 5-style global style regularization reaches highest retrieval rates on most metrics, which confirm the effectiveness of this inspiration.

Table 5: Results of ablation study of global style regularization.

Model	V2A						A2V					
	MedR↓	AvgR↓	R@1↑	R@5↑	R@10↑	R@20↑	MedR	AvgR	R@1	R@5	R@10	R@20
3-style	**15**	16.9	4.7	19.9	36.8	63.4	**18**	18.58	**3.9**	16.6	**30.5**	56.2
4-style	17	17.52	4.1	18.8	35.6	60.9	**18**	18.76	2.9	15.2	29.9	56.1
6-style	**15**	17.14	3.3	19.4	36.5	61.5	18.5	18.68	3.4	16.1	**30.5**	55.8
7-style	16	17.65	4.6	19.8	34.4	59.7	**18**	18.70	3.1	16.0	**30.5**	56.8
VMEval	**15**	**16.82**	**5.5**	**21.8**	**39.4**	**63.6**	**18**	**18.45**	**3.9**	**17.0**	29.8	**57.0**

D LIMITATION

Despite the effectiveness, VMEval relies on LLM-generated pseudo-supervised annotations for intermediate process, which may still introduce annotation noise or bias despite the human validation conducted in our study. In addition, while VMEval achieves strong quantitative agreement with human judgments, its generated explanations remain slightly behind the Salmonn2 baseline in some subjective or LLM-based evaluations. Therefore, while VMEval improves descriptive interpretability and quantitative scoring, further enhancing the expressiveness of generated explanations and exploring ways to mitigate pseudo-annotation bias remain an important direction. Finally, V2A retrieval provides useful objective evidence for harmony-aware representation learning, but does not fully capture perceptual harmony; developing more direct and standardized objective metrics for video-music harmony remains an important direction for future work.