CoT-VTM: Visual-to-Music Generation with Chain-of-Thought Reasoning

Xikang Guan[1] **Zheng Gu**[2] **Jing Huo**[1*] **Tianyu Ding**[3] **Yang Gao**[1]
[1]Nanjing University [2]Shenzhen University [3]Microsoft Corporation
xikangguan@smail.nju.edu.cn guzheng@szu.edu.cn
tianyudingu@microsoft.com {huojing,gaoy}@nju.edu.cn

Abstract

The application of visual-to-music generation (VTM) is rapidly growing. However, current VTM methods struggle with capturing the relationship between visuals and music in open-domain settings, mainly due to two challenges: the lack of large-scale, high-quality visual-music paired datasets and the absence of direct semantic correspondence between visuals and music. In this work, we propose CoT-VTM, a framework that distills Chain-of-Thought (CoT) reasoning to enable visual-to-music generation without paired data, while efficiently producing music aligned with visual content in open-domain settings. We first bridge the gap between visual, music, and text data using appropriate foundation models. Next, we identify key elements of the visual-music relationship and design a CoT prompt for visual-to-music mapping. To fully distill the reasoning of CoT, we incorporate latent information from intermediate reasoning steps as supervisory signals alongside visual and music supervision. Finally, we design a two-stage mapping distillation training process: the first stage uses discriminative MLP modules, while the second uses a generative embedding diffusion model (EmbedDiff). Our model achieves optimal performance on both image-to-music and video-to-music tasks. Project page: https://xxkkxxx.github.io/cot-vtm/

1 Introduction

Synesthesia, the phenomenon of associating music with specific visuals (Wang et al., 2023), has practical applications across various fields. In the film industry, appropriate background music enhances viewer immersion, a technique now widely used in advertising, animation, and social media content. However, manually selecting or composing music is costly and poses copyright challenges. Thus, this study focuses on generating instrumental music from visual content within an open visual domain.

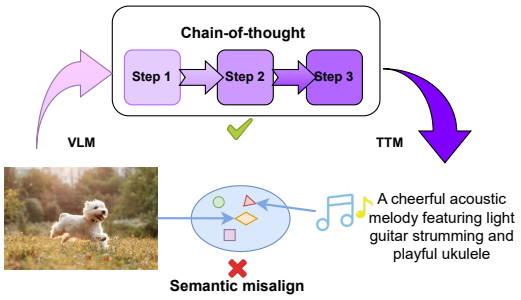

Figure 1: In the VTM task, there is no direct semantic correspondence between visuals and music, with LLM-based chain-of-thought reasoning capturing the implicit correspondence between them.

The task of generating music based on visual content is referred to as Visual-to-Music (VTM). Some existing VTM works are limited to specific scenarios, such as (Zhu et al., 2022b) and (Yu et al., 2023) that generate music from dance videos or other motion-rich video inputs. However, VTM in an open visual domain is far more challenging. The potential limitation for current methods in open-domain music generation lies in the need for large-scale, high-quality video-music paired datasets. This is similar to other cross-modal generation tasks like Text-to-Visual (TTV) (Rombach et al., 2022; Nan et al., 2024), where success is based on large-scale, high-quality paired text-visual datasets. However, most current video-music datasets(Hong et al., 2017; Zhu et al., 2022a), sourced from platforms like YouTube or TikTok, suffer from issues such as strong personal preferences and poor music quality. Some music video (MV) datasets (Kang et al., 2024; Zhuo et al., 2023) are mixed with human vocals, which limits the generation quality. Moreover, these datasets are substantially smaller than those used in TTV tasks.

Additionally, compared to other cross-modal generation tasks like text-to-video (TTV) or video-

*Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2025, pages 12493–12510

to-audio (VTA), visual-to-music (VTM) presents a unique challenge in capturing the mapping between the two modalities. In TTV tasks, the relationship between text and visuals is semantically clear, and in VTA tasks, visual content often corresponds directly to natural sounds (e.g., a barking dog or running water). In contrast, the relationship between visuals and music in VTM is ambiguous and subjective (Wang et al., 2023). Several approaches (Wang et al., 2023; Xiong et al., 2022) have used emotions or lyrics as intermediaries to bridge the gap between visuals and music, although this imposes stricter annotation requirements for training datasets.

Although there is no explicit semantic correspondence between visuals and music, certain visual factors(Ullah and Choi, 2024; Wu et al., 2016) implicitly influence corresponding elements in music, such as emotion, instruments, and rhythm. As a result, directly training a mapping model on paired data to capture the complex implicit relationships between visual and music proves challenging and prone to local optima. Inspired by the success of CoT prompting in Large Language Models (LLMs)(Kojima et al., 2022), we aim to leverage its reasoning power to uncover these complex mappings as illustrated in Figure 1. However, applying CoT to LLMs generates large volumes of intermediate outputs, which slows the inference speeds. We argue that for the specific task of visual-to-music mapping, such large-scale models are unnecessary.

We introduce the CoT-VTM framework, which leverages CoT reasoning to eliminate the reliance on paired video-music data, enabling efficient generation of high-quality music that aligns with visual content in open-domain. The CoT-VTM architecture is depicted in Figure 2. Initially, we identify implicit relationships between visual elements and musical elements, drawing on existing research (Ullah and Choi, 2024; Wu et al., 2016), and design a CoT prompt tailored for visual-to-music mapping, transforming visual captions into musical descriptions. To fully optimize the reasoning power of CoT, we integrate latent information, derived during the reasoning process but absent from the final output, as supervisory signals in the distillation training. We then encode textual supervisory signals into continuous embeddings using a suitable text encoder, forming supervision data for the mapping process. Given that latent information primarily pertains to visual analysis, we establish a one-to-one mapping between visual and latent information, while the relationship between visual elements and music is one-to-many. This leads to a two-stage approach: the first stage employs a multilayer perceptron (MLP) to map visual data to latent information, and the second stage uses a diffusion-based model, the embedding diffusion model (EmbedDiff), to map latent information to music. We utilize CLIP (Radford et al., 2021) to bridge the gap between visual data and visual captions, and musicGen (Copet et al., 2023) to bridge the gap between music audio and musical descriptions. CLIP (Radford et al., 2021) aligns visual and textual descriptions in a shared space, while musicGen (Copet et al., 2023) decodes music descriptions into actual audio.

The main contributions of our work can be summarized as follows:

- This is the first exploration of applying LLM-driven Chain-of-Thought reasoning to the VTM task.
- The CoT-VTM framework efficiently generates music based on visual content in open-domain scenarios without the need for paired visual-music data.
- Our theoretical and experimental results validate that CoT-VTM effectively utilizes CoT reasoning to produce high-quality and efficient visual-to-music mappings.
- CoT-VTM provides a novel paradigm for distilling LLM chain-of-thought reasoning.

2 Related Work

Music generation The music generation model has rapidly advanced (Dhariwal et al., 2020; Evans et al.; Forsgren and Martiros, 2022; Huang et al., 2023; Lu et al., 2023; Chen et al., 2024; Agostinelli et al., 2023; Schneider et al., 2024; Tang et al., 2024, 2023; Ziv et al., 2024). Recent models (Agostinelli et al., 2023; Copet et al., 2023) first convert continuous audio signals into discrete tokens, then train a text-conditioned music generation model from text prompts. Additionally, recent diffusion-based music generation methods (Schneider et al., 2024; Huang et al., 2023; Chen et al., 2024) use diffusion networks to predict music audio based on text prompts.

Visual-to-Music generation Current V2M methods can be divided into two types: Image-to-Music and Video-to-Music. In I2M work, BGT (Xiong et al., 2022) uses lyrics as an intermediary to generate music from images. Works such as (Santos

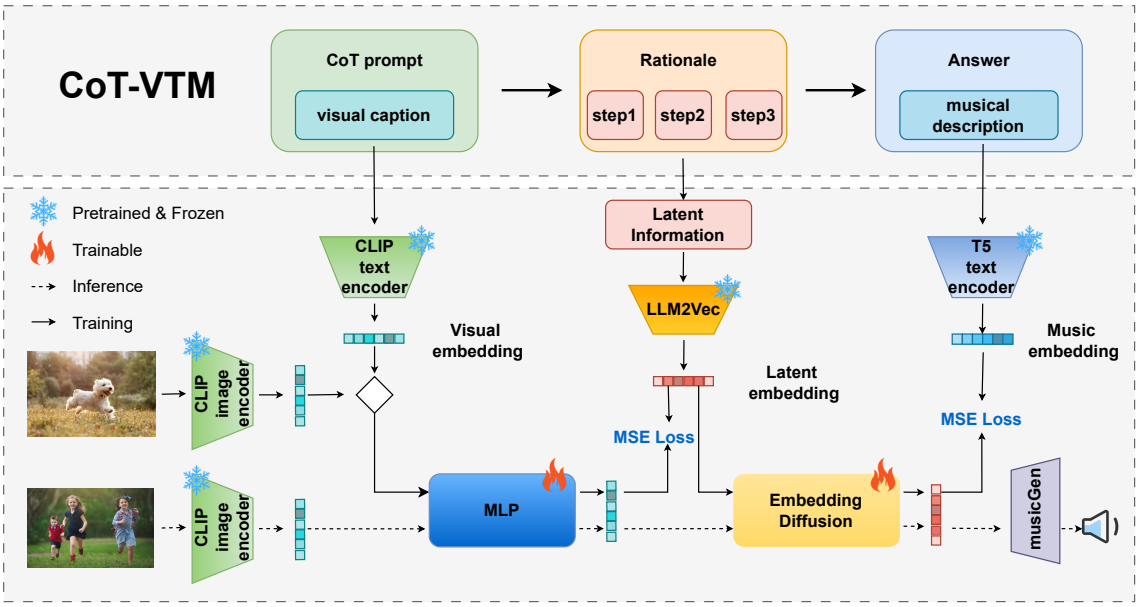

Figure 2: Overview of the CoT-VTM for video-to-music generation. The upper part shows the process of Chain-of-Thought reasoning by the LLM to generate supervised data. The lower part illustrates the distillation training and model inference process, where visual data X_v and visual text descriptions X_{vt} are used during training, and raw visual data is used during inference.

et al.; Saito et al., 2021) employ low-level image features as a bridge between images and music, while (Wang et al., 2023; Tian et al., 2025; Hisariya et al., 2024) use emotion as an intermediary to connect images with music. In V2M work, (Zhu et al., 2022a; Zhu et al.; Yu et al., 2023) focus on generating music conditioned on motion. In the open visual domain, (Di et al., 2021) uses rule-based methods to link video and music. Recent works such as (Kang et al., 2024; Li et al., 2024; Su et al., 2024; Lin et al., 2024; Tian et al., 2024) heavily rely on the scale and quality of visual-music paired data. Another line of work, such as (Haseeb et al., 2024; Liu et al., 2023), avoids dependence on large-scale video-music pairs by utilizing the powerful reasoning capabilities of LLMs. However, this leads to a significant increase in model parameters, which in turn causes slower inference times.

CoT distillation While direct prompting enables large language models (LLMs) to perform complex reasoning through Chain-of-Thought (CoT), smaller language models (SLMs) struggle due to limited capacity(Stolfo et al., 2023). Knowledge distillation (KD) provides an effective framework for transferring the reasoning capabilities of teacher models to SLMs(Xu et al., 2024). A simple yet effective approach is using a teacher-student paradigm, where teacher-generated CoT steps guide the SLMs, addressing their limitations and enhancing performance on reasoning-intensive tasks. Traditional approaches(Hsieh et al., 2023; Ho et al., 2023; Magister et al., 2023) typically train smaller student models to mimic the step-by-step outputs from larger teacher LLMs. (Ranaldi and Freitas, 2024) help students generate structured reasoning, improving performance in tasks such as question answering and mathematics, while (Zhuang et al., 2025) propose a unified structured CoT distillation framework for effective knowledge transfer. In contrast, our approach introduces a new architecture where, instead of using a transformer-based student model that performs autoregressive modeling, we leverage a suitable pre-trained encoder to encode discrete tokens into high-density continuous embeddings. We then use diffusion for joint probability modeling and conduct phase-wise distillation based on the characteristics of the CoT data.

3 CoT-VTM

3.1 Task Definition

In general, we model the visual-to-music generation task as the conditional probability distribution of generating music X_m given a visual condition X_v:

$$X_m \sim P(X_m \mid X_v) \tag{1}$$

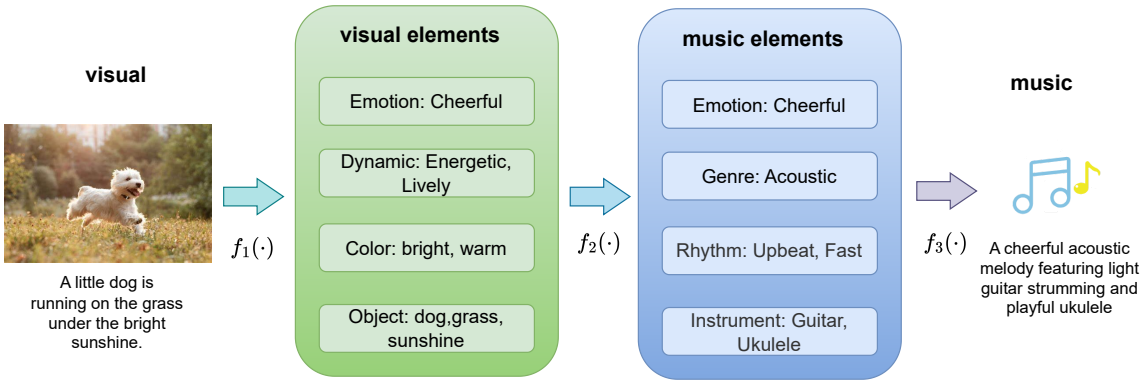

Figure 3: The three steps of Chain-of-Thought reasoning by the LLM: The first stage $f_1(\cdot)$ maps visual captions to visual elements; the second stage $f_2(\cdot)$ maps visual elements to music elements; the third stage $f_3(\cdot)$ maps music elements to a comprehensive music description.

Considering that the text-to-music task has already achieved significant success, in this work, we aim to generate the textual description of music based on the visual condition. We then use a pre-trained musicGen (Copet et al., 2023) to generate the corresponding music from the generated textual description. Thus, we model the task as generating the music textual description X_{mt} given the visual condition X_v:

$$X_{mt} \sim P(X_{mt} \mid X_v) \tag{2}$$

3.2 Reasoning Process

Due to the lack of direct semantic correspondence between visual and music, directly mapping the visual space to the music space is highly challenging. Therefore, we first explore the implicit associations between visual elements and music elements, and then divide the reasoning process into three stages, as shown in Figure 3.

In the first stage, we focus on analyzing the captions of visual data and extracting the various elements within the visual content. We do not extract all the elements present in the visual data, but rather focus on those visual elements that influence the music generation. Based on previous approaches (Ullah and Choi, 2024; Wu et al., 2016), we identify the key visual elements V_e that influence music generation, including object V_{obj}, emotion V_{emo}, dynamic V_{dyn}, and color V_{col}. The LLM maps the visual caption X_{vt} to the visual elements V_e in the first stage :

$$V_e = (V_{obj}, V_{emo}, V_{dyn}, V_{col}) = f_1(X_{vt}) \tag{3}$$

In the second stage, to set the required music elements, we analyze the text instructions used in musicGen (Copet et al., 2023) for text-to-music generation tasks. Specifically, we randomly select 50 text instruction samples from the training dataset. Using an LLM, we analyze the key musical elements in music text instructions. Based on this, we also refer to previous works (Ullah and Choi, 2024; Wu et al., 2016) to investigate how the visual elements identified in the first stage can influence the music elements extracted in this stage. Finally, we extract four key music elements M_e: genre M_{gen}, instrument M_{ins}, rhythm M_{rhy}, and emotion M_{emo}. The LLM maps the visual elements V_e obtained in the first stage to the music elements M_e in the second stage :

$$M_e = (M_{gen}, M_{ins}, M_{rhy}, M_{emo}) = f_2(V_e) \tag{4}$$

In the third stage, the LLM performs the final synthesis and coordination of the different music elements to produce a comprehensive music description X_{mt} :

$$X_{mt} = f_3(M) \tag{5}$$

3.3 CoT Prompt Engineering

Following previous approaches (*e.g.*, few-shot and CoT prompting), we carefully design the CoT prompt based on the analysis in Section 3.2 to ensure that the LLM can complete the three-stage reasoning process of visual-to-music mapping and generate high-quality music descriptions. Our designed CoT prompt P_{CoT} consists of five parts: the expert role section P_{expert}, the task objective section P_{task}, the step-by-step reasoning section P_{steps}, the output example section $P_{example}$, and the visual caption section X_{vt}. We present a de-

tailed CoT prompt in Appendix A:

$$P_{CoT} = [P_{expert}][P_{task}][P_{steps}][P_{example}][X_{vt}] \tag{6}$$

By using the designed CoT prompt to guide the LLM through the three-stage reasoning process mentioned in Section 3.2:

$$X_m = f(P_{CoT}(X_{vt})) \tag{7}$$

3.4 Data Preparation

We first construct the VTD dataset, which centers on visual text descriptions and comprises three parts: (1) paired text-visual data used in TTV tasks, (2) emotion-labeled image datasets annotated by image labeling models, and (3) diverse visual scene descriptions generated by carefully designed prompts guiding a LLM. Detailed information can be found in Appendix B and Section 4.1. Next, we extract the visual text descriptions X_{vt} from the VTD dataset, embed them into the CoT prompt, and use the LLM to generate the reasoning process $Rationale$ and final answer $Answer$. From both $Rationale$ and $Answer$, we extract the latent information X_{rt} and the music text description X_{mt}, respectively. This process is applied to all X_{vt} in the dataset, yielding the supervised data required for training the visual-to-music mapping module, denoted as $\mathcal{D} = \{(x_{vt}, x_{rt}, x_{mt})\}_{i=1}^{N}$.

To achieve more efficient training, we select appropriate pre-trained encoders to encode (X_{vt}, X_{rt}, X_{mt}) into low-dimensional embedding data. Considering that the CLIP (Radford et al., 2021) maps visual and textual data into a shared feature space and is trained on 400M paired text-image data, offering strong generalization capability, we first choose CLIP (Radford et al., 2021) as the visual encoder. Using visual and text encoders of CLIP(Radford et al., 2021), we encode the raw visual data X_v and visual text data X_{vt} into visual embeddings $E_v \in R^C$ and $E_{vt} \in R^C$. Given that the latent information is more complex, and to minimize information loss while ensuring comprehensive representation, we select the state-of-the-art text encoding model LLM2Vec (BehnamGhader et al., 2024) as the encoder for the latent information, encoding X_{rt} into latent embeddings $E_r \in R^D$. For music description data, since we are using the music generation model MusicGen (Copet et al., 2023) , which first encodes music description text into embeddings via the T5 (Raffel et al., 2020) text encoder, we select T5 (Raffel et al., 2020) as the encoder for music text description data X_{mt}, encoding X_{mt} into music embeddings $E_t \in R^S$.

3.5 Visual-to-Music Mapping

Through the work in this section, we encode the discrete textual supervisory signals (X_{vt}, X_{rt}, X_{mt}) into continuous embedding data (E_v, E_r, E_m). This means that our subsequent training process only needs to operate on the embedding data. To make the visual-to-music mapping results more stable and accurate, and to avoid the model from falling into local optima, we do not directly map E_v to E_m. Instead, we emulate the reasoning process of CoT by using the intermediate reasoning process to trigger the model's reasoning ability. Next, I prove the theoretical feasibility of this approach starting from the definition of the VTM task. The detailed proof process can be found in Appendix C. Based on the original VTM definition, we transform the problem into generating music E_m given the visual condition E_v:

$$E_m \sim P\left(E_m \mid E_v\right) \tag{8}$$

We adapt the CoT methodology by introducing E_r into the video-to-music mapping. Here, E_v corresponds to question and prompt in the CoT reasoning, E_m corresponds to the answer, and E_r corresponds to the rationale:

$$P(E_m, E_r \mid E_v) = P(E_m \mid E_r, E_v) \cdot P(E_r \mid E_v) \tag{9}$$

The original VTM modeling is therefore transformed as follows:

$$E_m \sim P(E_m \mid E_r, E_v) \cdot P(E_r \mid E_v) \tag{10}$$

Through the above theoretical analysis and modeling, we address the core issue of applying the powerful reasoning capability of CoT to the visual-to-music mapping task. Based on this, we design two mapping modules to sequentially achieve the mappings $P(E_r \mid E_v)$ from E_v to E_r, and $P(E_m \mid E_r, E_v)$ from E_v and E_r to E_m. For the first mapping $P(E_r \mid E_v)$, since the primary content of the latent information involves a discriminative analysis of visual factors, we treat the relationship between E_v and E_r as a one-to-one discriminative correspondence. Both the raw visual data and the latent information are encoded into dense continuous feature representations, enabling a lightweight MLP to effectively learn this mapping.Hence, we design a discriminative mapping

module (MLP) to perform the mapping from E_v to E_r based on multilayer perceptrons.

$$E_r^{'} = MLP(E_v) \tag{11}$$

We use Mean Square Error loss to guide the training. The training process can be formulated as follows:

$$L = E_{i\sim[1,K]}\left[\left\|E_r^{(i)} - E_r^{(i)'}\right\|^2\right] \tag{12}$$

where K is the batch size and $E_r^{(i)'}$ is from $MLP(E_v^{(i)})$.

Inspired by DALLE2's prior model (Ramesh et al., 2022), we treat the projection process as a conditional generation task, which models a one-to-many mapping to ensure the diversity and generalization of the target music distribution. Given that visual and music data are not one-to-one correspondences, we design an embedding conditional diffusion mapping module, abbreviated as EmbedDiff, based on the generative model diffusion to model $P(E_m \mid E_r, E_v)$. We first use a mapping layer $f_{mlp}(\cdot)$ to map $E_r \in R^D$ to the same size as $E_v \in R^C$, obtaining $E_r^{'} \in R^C$. We concatenate $E_v \in R^C$ and $E_r^{'} \in R^C$ as conditions to generate $E_m \in R^S$. In the forward process, the original music embedding distribution transforms into a standard Gaussian distribution by gradually adding noise with a fixed schedule $\alpha_1, \ldots, \alpha_T$, where T is the total number of timesteps, and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$:

$$q(E_m^t \mid E_m^{t-1}) = \mathcal{N}(E_m^t; \sqrt{\alpha_t}E_m^{t-1}, (1-\alpha_t)I) \tag{13}$$

$$q(E_m^t \mid E_m^0) = \mathcal{N}(E_m^t; \sqrt{\alpha_t}E_m^0, (1-\alpha_t)I) \tag{14}$$

The goal of EmbedDiff is to mirror score matching by optimizing the denoising objective:

$$\mathcal{L}_{EmbedDiff} = E_{E_m^0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(E_m^t, t, E_v, E_r)\|_2^2\right] \tag{15}$$

After EmbedDiff is trained, we generate the music embedding E_m by sampling through the reverse process with $E_m^T \sim \mathcal{N}(0, I)$, conditioned on the given visual embedding E_v and rationale E_r, using the following reverse dynamics:

$$p_\theta(E_m^{t-1} \mid E_m^t) = \mathcal{N}(E_m^{t-1}; \mu_\theta(E_m^t, t, E_v, E_r), \sigma_t^2 I) \tag{16}$$

Finally, the music audio $\hat{x}_m$ is obtained by decoding the generated E_m^0 with a music decoder D, where $\hat{x}_m = D(E_m^0)$. In the CoT-VTM framework, musicGen (Copet et al., 2023) is used as the music decoder. It employs a single-stage transformer model that processes compressed discrete music representations, eliminating the need for multiple cascading models or upsampling. This efficient design enables musicGen (Copet et al., 2023) to generate high-quality mono and stereo music samples at a faster speed.

4 Experiments

4.1 Dataset

To enable the model to generate music corresponding with data in an open visual domain, we constructed a training dataset, VDM, consisting of three parts. The first part includes text-image and text-video paired data (Srinivasan et al., 2021; Wang et al.), with 10k samples each. The second part is based on the Emoset (Yang et al., 2023) image sentiment classification dataset, which results in 11k text-image pairs generated using the Qwen2-vl (Wang et al., 2024) for image annotation. The third part consists of 10k textual descriptions of visual scenes, generated using carefully designed prompts with the GPT-4o API. The rationale for selecting these datasets is explained in Appendix B. In total, 41K data points were used for model training. To effectively evaluate the model's performance in generating music in the open visual domain, we have constructed two test sets for the I2M and V2M tasks based on existing visual-music paired datasets (Hong et al., 2017; Zhu et al., 2022a; Li et al., 2018, 2021, 2024; Verma et al., 2019). Further details are provided in Appendix B.

4.2 Implementation Details

We employ GPT-4o for chain-of-thought reasoning. For the CLIP (Radford et al., 2021) , LLM2Vec (BehnamGhader et al., 2024), T5 (Raffel et al., 2020) , and MusicGen (Copet et al., 2023) models, we use the "clip-vit-base-patch32*", "LLM2Vec-Meta-Llama-3-8B-Instruct-mntp†", "T5-base‡", and "musicgen-large§" versions, respectively. The MLP model is trained for 65 epochs using the AdamW optimizer with a constant learning rate of 5e-4, a batch size of 256 embedding pairs, and a dropout rate of 0.2 for regularization. For the EmbedDiff, we ap-

*https://huggingface.co/openai/clip-vit-base-patch32

†https://huggingface.co/McGill-NLP/LLM2Vec-Meta-Llama-3-8B-Instruct-mntp

‡https://huggingface.co/google-t5/t5-base

§https://huggingface.co/facebook/musicgen-large

Method	FD↓	FAD↓	KL↓	IBS↑	EMS↓	P@10↑	OQL↑	COV↑
GT	0.000	0.000	0.000	0.217	2.587	/	71.43	69.11
I-caption2music	61.834	6.447	2.041	0.132	4.679	20.84	39.85	41.46
CoDI (Tang et al., 2024)	156.744	18.763	5.854	0.114	4.785	15.37	29.95	31.12
M2UGen (Liu et al., 2023)	61.445	7.213	1.987	0.179	3.814	24.61	55.90	54.41
CoT-VTM	**43.946**	**4.606**	**1.610**	**0.194**	**2.473**	**29.84**	**60.09**	**58.49**

Table 1: Comparison of objective and subjective evaluation metrics for different models in the I2M task. Lower values for FD, FAD, KL and EMS are better, while higher values for P@10, OQL, and COV are better.

ply a cosine noise schedule with 1000 diffusion steps during training and 200 steps during inference. The EmbedDiff converges after 72 epochs, using the AdamW optimizer with a learning rate of 1e-4, a batch size of 256 embedding pairs, and a dropout rate of 0.1 for classifier-free guidance. Both MLP and EmbedDiff models are trained on a single NVIDIA RTX A6000 GPU.

4.3 Evaluation

Objective Evaluation We evaluate performance based on two aspects: music quality and its relevance to the visual data. For music quality, we employ Fréchet Distance (FD), Fréchet Audio Distance (FAD) (Kilgour et al., 2018), and Kullback-Leibler Divergence (KL) to assess both the overall quality and variability of the generated audio. For music-visual alignment, we use the ImageBind Score (IBS) (Girdhar et al., 2023) to evaluate the correspondence between video and generated music. However, we acknowledge that IBS has limitations, as it was not specifically trained on music data. To enhance our assessment of the music-visual connection, we also consider the emotional matching metric emotion matching score (EMS) and the music retrieval metric P@10, with further details provided in Appendix D.

Subjective Evaluation Subjective evaluation is conducted via a questionnaire distributed to 16 non-experts and 12 graduate students specializing in music. To minimize bias, we select images and videos from a variety of scenes and generate music for each visual using different models. The order of music corresponding to each visual is randomized. Participants are asked to rate the generated music based on overall quality (OQL) and relevance to the corresponding visual content (COV), using a scale from 1 to 100. The questionnaire includes 18 samples and takes approximately 10 minutes to complete. To assess the reliability of human evaluations, we measured inter-annotator agreement using *Fleiss' Kappa*, a standard metric for evaluating consistency among multiple raters. Given the subjective nature of our evaluation criteria—*Overall Quality* of the generated music (OQL) and *Consistency* with the visual content (COV)—we discretized the 1–100 rating scale into ten bins (e.g., [1–10], [11–20], ..., [91–100]) before computing agreement scores.

Models To assess the effectiveness of our model in capturing the intricate relationships between visuals and music, as well as generating high-quality music, we conduct experiments for both image-to-music (I2M) and video-to-music (V2M) tasks. For the I2M task, there is currently no direct method for generating music from images. CoDi (Tang et al., 2024) is an any-to-any generation model, and M2UGen (Liu et al., 2023)leverages a LLM to bridge vision and language. Both CoDi (Tang et al., 2024) and M2UGen (Liu et al., 2023) serve as baselines for the I2M and V2M tasks. For the V2M task, strong baselines include (Lin et al., 2024; Su et al., 2024). However, these models have not released their training or inference code or datasets. Other approaches, such as (Di et al., 2021; Kang et al., 2024; Zhu et al.; Yu et al., 2023) are excluded from comparison due to differences in scope (e.g., symbolic music generation or dance-to-music), which would make the comparison unfair. VidMuse (Tian et al., 2024), currently the state-of-the-art (SOTA) in open-source video-to-music, is included as a strong baseline. Additionally, we introduce two simple baselines for I2M and V2M: I-Caption2music and V-Caption2music. I-Caption2music uses BLIP (Li et al., 2022) to extract image captions and generates music by feeding them into MusicGen (Copet et al., 2023). V-Caption2music uses SpaceTimeGPT[¶] to extract video captions and generates music similarly by feeding them into MusicGen.

[¶] https://huggingface.co/Neleac/SpaceTimeGPT

Method	FD↓	FAD↓	KL↓	IBS↑	P@10↑	OQL↑	COV↑
GT	0.000	0.000	0.000	0.235	/	76.43	74.88
V-Caption2music	55.332	5.638	1.333	0.182	22.46	43.12	50.10
CoDI (Tang et al., 2024)	120.942	14.340	3.218	0.129	17.53	31.76	34.22
M2UGen (Liu et al., 2023)	60.435	6.811	1.462	0.183	26.43	57.84	54.47
VidMuse (Tian et al., 2024)	42.832	4.481	**1.194**	0.199	30.87	56.35	**61.10**
CoT-VTM	**34.004**	**4.037**	1.212	**0.210**	**31.46**	**63.44**	60.89

Table 2: Comparison of objective and subjective evaluation metrics for different models in the V2M task.

Group	OQL	COV
Non-experts	0.29	0.32
Music specialists	0.45	0.41

Table 3: Fleiss' Kappa scores for inter-rater agreement across different annotator groups.

4.4 Results

To account for domain expertise, we report Fleiss' Kappa scores separately for non-experts and graduate students specializing in music. As shown in Table 3, music specialists achieved notably higher agreement than non-experts, suggesting that domain knowledge leads to more consistent and reliable evaluations. These findings are consistent with prior observations in video-music research, where annotations from user-generated platforms such as YouTube or TikTok often exhibit high variance due to personal preference.

Table 1 and Table 2 display the performance of various models on the I2M and V2M tasks, respectively. For the I2M task, the proposed method outperforms all other models in both objective and subjective metrics. For the V2M task, although our method slightly lags behind VidMuse(Tian et al., 2024) in certain metrics, it achieves the best overall performance across the baselines. Specifically, in terms of music generation quality, our method yields slightly higher KL scores than VidMuse(Tian et al., 2024), but achieves better results on objective metrics such as FD and FAD. To assess the significance of this improvement, we conducted a paired t-test on FD between our method and VidMuse(Tian et al., 2024). The test yielded a t-statistic of -9.57 and a p-value of 5.16×10^{-6}, which is well below the standard threshold of 0.05, confirming that the improvement is statistically significant. Regarding music-video relevance, our method performs slightly worse than VidMuse(Tian et al., 2024) in the subjective metric COV. We attribute this to the fact that our model emphasizes global relationships between visuals (including both images and videos) and music, whereas VidMuse(Tian et al., 2024) relies more heavily on temporal visual transformations, which limits its generalizability to I2M tasks. Overall, the results demonstrate that our method not only generates higher-quality music but also exhibits stronger global visual-to-music correspondence.

4.5 Ablation

First, we replaced the EmbedDiff model with an MLP model within the original CoT-VTM framework to evaluate whether the EmbedDiff model offers a better mapping. Next, we trained an EmbedDiff model using only visual and music embeddings as supervision to map the visual embeddings to music embeddings. In other words, we abandoned the latent embedding in this experiment to assess whether the intermediate reasoning latent information enhances the mapping process.

As shown in Table 4, replacing the EmbedDiff model with the MLP model resulted in a significant decline in music generation quality, although the ImageBind score (which evaluates relevance) improved. We believe this is due to the MLP model's discriminative nature, which imposes stricter constraints on music generation. This improves relevance but limits music quality and diversity. Notably, other relevance metrics did not improve; in fact, the COV metric decreased substantially. We attribute this to the decline in music quality, which negatively affects human auditory perception, thereby reducing the observed relevance. Consequently, EmbedDiff, as a generative model, is better suited for the V2M task. Abandoning latent information resulted in performance degradation across all metrics, highlighting the critical role of latent information in learning the visual-to-music mapping. This also provides preliminary evidence that our model effectively utilizes CoT's reasoning capabilities.

Method	FD↓	FAD↓	KL↓	IBS↑	P@10↑	OQL↑	COV↑
ve→mlp→le→mlp→me	44.896	4.887	1.310	0.231	31.44	45.33	55.64
ve→EmbedDiff→me	36.966	4.264	1.222	0.199	24.62	60.34	54.82
ve→mlp→le→EmbedDiff→me	34.004	4.037	1.212	0.210	31.46	63.44	60.89

Table 4: `ve→mlp→le→mlp→me` is used to validate the advantages of the generative mapping module EmbedDiff; `ve→EmbedDiff→me` verifies the facilitative effect of latent information on the visual-to-music mapping.

Method	FD↓	FAD↓	KL↓	IBS↑	P@10↑	OQL↑	COV↑	Inference time↓	Memory↓
LLM2music	36.463	4.293	1.294	0.202	27.55	64.00	59.83	14.04s + (4.99s)	145.87G + (8.07G)
CoT2music	32.426	3.941	1.213	0.219	32.68	65.46	64.46	81.34s + (4.99s)	145.87G + (8.07G)
CoT-VTM	33.876	4.069	1.212	0.209	32.14	64.51	61.14	2.33s + (4.99s)	1.73G + (8.07G)

Table 5: Comparison of visual-to-music generation methods. CoT2music verifies the role of Chain-of-Thought reasoning in the mapping process, while CoT-VTM demonstrates its distilled reasoning capability with substantially lower inference cost.

To further investigate whether our model fully capitalizes on chain-of-thought reasoning, we conducted two experiments. In the first experiment, SpaceTimeGPT generates captions for videos, followed by a non-CoT prompt guiding a LLM to generate music descriptions based on the captions. These descriptions are then used to prompt MusicGen (Copet et al., 2023) to generate audio. The second experiment is similar, but replaces the non-CoT prompt with our carefully designed CoT prompt. For simplicity, we refer to these experiments as LLM2Music and CoT2Music, respectively.

From the results in Table 5, we observe that CoT2Music substantially outperforms LLM2Music in overall performance, demonstrating the importance of utilizing CoT's reasoning abilities for mapping visual to music. CoT-VTM achieves comparable performance to CoT2Music while substantially reducing inference time and GPU memory usage. Since our distillation focuses solely on the high-level mapping from visual representations to music representations, excluding the audio decoding stage, we report efficiency metrics for this component. The additional cost of downstream music generation using MusicGen is reported separately. These results demonstrate that CoT-VTM successfully distills CoT reasoning into a compact and efficient model, delivering significant computational savings with minimal impact on performance.

5 Conclusion

In this work, we introduced CoT-VTM, a framework that leverages CoT reasoning to address open-domain VTM without relying on large-scale paired datasets. By integrating latent information from CoT as supervisory signals, our two-stage approach—using MLP for visual-to-latent mapping and EmbedDiff for latent-to-music generation—efficiently produces high quality and semantically aligned music. Experimental results show that CoT-VTM outperforms existing models in both image-to-music and video-to-music tasks, providing a scalable and data-efficient solution for VTM. This work demonstrates the potential of CoT reasoning in bridging visual and musical domains, advancing cross-modal generation tasks.

Limitations

The current limitations of the CoT-VTM framework are two fold. First, since CoT-VTM relies on pre-trained text-to-music models, the quality of the generated music is constrained by the performance of these models. Second, our current focus is on visual-to-music mapping (including both images and videos), without addressing the temporal consistency between visuals and music, especially for video data.

In future work, we will explore how to apply the Chain-of-Thought reasoning ability of LLMs to visual-to-music mapping without relying on pre-trained text-to-music models. Additionally, we aim to investigate how to leverage this reasoning ability to generate music that aligns with the temporal changes in video content.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts,

Marco Tagliasacchi, and 1 others. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*.

Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2024. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1206–1210. IEEE.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36:47704–47720.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*.

Shangzhe Di, Zeren Jiang, Si Liu, Zhaokai Wang, Leyan Zhu, Zexin He, Hongming Liu, and Shuicheng Yan. 2021. Video background music generation with controllable music transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2037–2045.

Zach Evans, CJ Carr, Josiah Taylor, Scott H Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. In *Forty-first International Conference on Machine Learning*.

Seth Forsgren and Hayk Martiros. 2022. Riffusion-stable diffusion for real-time music generation. *URL https://riffusion. com*.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190.

Muhammad Taimoor Haseeb, Ahmad Hammoudeh, and Gus Xia. 2024. Gpt-4 driven cinematic music generation through text processing. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6995–6999. IEEE.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, and 1 others. 2017. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.

Tanisha Hisariya, Huan Zhang, and Jinhua Liang. 2024. Bridging paintings and music–exploring emotion based music generation through paintings. *arXiv preprint arXiv:2409.07827*.

Nam Ho, Lukas Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882.

Sungeun Hong, Woobin Im, and Hyun S. Yang. 2017. Content-based video-music retrieval using soft intra-modal structure constraint. *Preprint*, arXiv:1704.06761.

Cheng-Yu Hsieh, Chen-Lin Li, Chia-Hao Yeh, and 1 others. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.

Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, and 1 others. 2023. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*.

Jaeyong Kang, Soujanya Poria, and Dorien Herremans. 2024. Video2music: Suitable music generation from videos using an affective multimodal transformer model. *Expert Systems with Applications*, 249:123640.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2018. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. 2018. Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia*, 21(2):522–535.

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

Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412.

Sizhe Li, Yiming Qin, Minghang Zheng, Xin Jin, and Yang Liu. 2024. Diff-bgm: A diffusion model for video background music generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27348–27357.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, and 1 others. 2023. Mert: Acoustic music understanding model with large-scale self-supervised training. *arXiv preprint arXiv:2306.00107*.

Yan-Bo Lin, Yu Tian, Linjie Yang, Gedas Bertasius, and Heng Wang. 2024. Vmas: Video-to-music generation via semantic alignment in web music videos. *arXiv preprint arXiv:2409.07450*.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. 2023. M Θ{2} ugen: Multi-modal music understanding and generation with the power of large language models. *arXiv preprint arXiv:2311.11255*.

Peiling Lu, Xin Xu, Chenfei Kang, Botao Yu, Chengyi Xing, Xu Tan, and Jiang Bian. 2023. Musecoco: Generating symbolic music from text. *arXiv preprint arXiv:2306.00110*.

Linus C Magister, Jonathan Mallinson, Jaroslav David Adamek, and 1 others. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. 2024. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: A survey. *arXiv preprint arXiv:2212.09597*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Lorenzo Ranaldi and André Freitas. 2024. Aligning large and small language models via chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1812–1827.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Yasuyuki Saito, Honoka Fujii, and Shigeki Sagayama. 2021. Semi-automatic music piece creation based on impression words extracted from object and background in color image. In *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, pages 268–272. IEEE.

A Santos, H Pinto, R Pereira Jorge, and Nuno Correia. Musyfi: music synthesis from images. In *Proceedings of the 12th International Conference on Computational Creativity*, pages 103–112.

Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. 2024. Moûsai: Efficient text-to-music diffusion models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8050–8068, Bangkok, Thailand. Association for Computational Linguistics.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449.

Alessandro Stolfo, Ziniu Jin, Kumar Shridhar, and 1 others. 2023. A causal framework to quantify the robustness of mathematical reasoning with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 545–561.

Kun Su, Judith Yue Li, Qingqing Huang, Dima Kuzmin, Joonseok Lee, Chris Donahue, Fei Sha, Aren Jansen, Yu Wang, Mauro Verzetti, and 1 others. 2024. V2meow: Meowing to the visual beat via video-to-music generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4952–4960.

Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. 2024. Codi-2: In-context interleaved and interactive any-to-any generation. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 27425–27434.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. 2023. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36:16083–16099.

Sida Tian, Can Zhang, Wei Yuan, Wei Tan, and Wenjie Zhu. 2025. Xmusic: Towards a generalized and controllable symbolic music generation framework. *arXiv preprint arXiv:2501.08809*.

Zeyue Tian, Zhaoyang Liu, Ruibin Yuan, Jiahao Pan, Qifeng Liu, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. 2024. Vidmuse: A simple video-to-music generation framework with long-short-term modeling. *arXiv preprint arXiv:2406.04321*.

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

Gaurav Verma, Eeshan Gunesh Dhekane, and Tanaya Guha. 2019. Learning affective correspondence between music and image. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3975–3979. IEEE.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yajie Wang, Mulin Chen, and Xuelong Li. 2023. Continuous emotion-based image-to-music generation. *IEEE Transactions on Multimedia*.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, and 1 others. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*.

Xixuan Wu, Yu Qiao, Xiaogang Wang, and Xiaoou Tang. 2016. Bridging music and image via cross-modal ranking analysis. *IEEE Transactions on Multimedia*, 18(7):1305–1318.

Zeyu Xiong, Pei-Chun Lin, and Amin Farjudian. 2022. Retaining semantics in image to music conversion. In *2022 IEEE International Symposium on Multimedia (ISM)*, pages 228–235. IEEE.

Xiaojie Xu, Ming Li, Chongyang Tao, and 1 others. 2024. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Danny Cohen-Or, and Hui Huang. 2023. Emoset: A large-scale visual emotion dataset with rich attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20383–20394.

Jiashuo Yu, Yaohui Wang, Xinyuan Chen, Xiao Sun, and Yu Qiao. 2023. Long-term rhythmic video soundtracker. In *International Conference on Machine Learning*, pages 40339–40353. PMLR.

Ye Zhu, Kyle Olszewski, Yu Wu, Panos Achlioptas, Menglei Chai, Yan Yan, and Sergey Tulyakov. 2022a. Quantized gan for complex music generation from dance videos. In *European Conference on Computer Vision*, pages 182–199. Springer.

Ye Zhu, Yu Wu, Kyle Olszewski, Jian Ren, Sergey Tulyakov, and Yan Yan. Discrete contrastive diffusion for cross-modal music and image generation. In *The Eleventh International Conference on Learning Representations*.

Ye Zhu, Yu Wu, Kyle Olszewski, Jian Ren, Sergey Tulyakov, and Yan Yan. 2022b. Discrete contrastive diffusion for cross-modal music and image generation. *arXiv preprint arXiv:2206.07771*.

Xin Zhuang, Zhaoliang Zhu, Zhen Wang, and 1 others. 2025. Unicott: A unified framework for structural chain-of-thought distillation. In *The Thirteenth International Conference on Learning Representations*.

Le Zhuo, Zhaokai Wang, Baisen Wang, Yue Liao, Chenxi Bao, Stanley Peng, Songhao Han, Aixi Zhang, Fei Fang, and Si Liu. 2023. Video background music generation: Dataset, method and evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15637–15647.

Alon Ziv, Itai Gat, Gael Le Lan, Tal Remez, Felix Kreuk, Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2024. Masked audio generation using a single non-autoregressive transformer. *arXiv preprint arXiv:2401.04577*.

A More Details about CoT prompt

As shown in Figure 1, the overall structure and components of the CoT prompt are illustrated. Below shows the detailed contents of the different sections within the CoT prompt.

A.1 The content of CoT prompt

Your task is to generate a textual description of background music that matches the scene depicted in the image. Since there is no direct correspondence between the visual and music domains, the task must be broken down into steps for structured output. Follow these instructions:

Step 1: Based on the scene description, analyze the following visual elements:

- **Object:** Identify the objects present in the scene that might influence the music (e.g., animals, people, trees, etc.).

- **Emotion:** What emotions does the scene evoke (e.g., tranquil, joyful, energetic)?
- **Dynamic:** What is the overall dynamic of the scene (e.g., calm, energetic, fast-paced)?
- **Color:** What is the dominant color or color scheme (e.g., warm tones, cool tones)?

Step 2: Based on the visual elements extracted in Step 1, analyze the corresponding music elements:

- **Genre:** Suggest an appropriate music genre (e.g., melodic, classical, upbeat, electronic).
- **Instruments:** List the instruments that suit the atmosphere (e.g., piano, guitar, strings, electronic sounds).
- **Emotion:** What emotion should the music evoke (e.g., peaceful, joyful, energetic)?
- **Rhythm:** Describe the tempo and pacing of the music (e.g., slow, fast, syncopated, steady).

Step 3: Based on the musical elements extracted in Step 2, synthesize and generate the final music description.

Output Format: Provide the final output in JSON format as follows:

```
{
 "image_description": "<input description>",
  "step1 reasoning": "<reason>",
  "visual elements": {
    "object": "<object description>",
    "emotion": "<emotion description>",
    "dynamic": "<dynamic description>",
    "color": "<color description>"
  },
 "visual description": "<visual description>",
  "step2 reasoning": "<reason>",
  "music elements": {
    "genre": "<genre>",
   "instruments": ["<instrument1>", ...],
    "emotion": "<emotion>",
    "rhythm": "<rhythm>"
  },
  "step3 reasoning": "<reason>",
 "music description": "<music description>"
}
```

A.2 The role of each section in the CoT prompt

The expert role section P_{expert} defines the model's professional identity and task scope, specifying that the model's role is to analyze images and create background music, guiding it to perform precise reasoning and generation within this specific domain. The task objective section P_{task} sets the core goal of the task to generate a matching background music description based on the visual description, providing the model with a clear target and guidance. The process decomposition section P_{steps} breaks down the overall task into three reasoning processes, with each step specifying the operations to be performed, ensuring the model can progressively analyze the image and generate the corresponding music description, enhancing task efficiency and accuracy. The output example section $P_{example}$ provides an input-output instance in the form of a tuple $Question, Rationale, Answer >$, giving the model a clear reasoning process and output format, ensuring that the generated result meets the expected format, is practically applicable, and maintains consistency and standardization for subsequent information extraction. Finally, the visual caption section X_{vt} serving as the basis for the model's reasoning.

A.3 CoT sample

The following content includes CoT-generated samples: **Image Description:** *A child holds a bright red balloon, standing at the edge of a calm lake. The balloon sways gently in the breeze, reflecting the afternoon sun.*

Step 1 – Visual Reasoning: The visual scene evokes a peaceful, serene atmosphere. The presence of the child with the bright red balloon suggests innocence and joy, while the calm lake and gentle breeze further add to the tranquility of the moment.

Extracted Visual Elements:

- **Objects:** A child, a bright red balloon, calm lake
- **Emotions:** peaceful, joyful, serene
- **Dynamics:** calm, gentle
- **Colors:** warm tones (afternoon sun), bright red balloon

Visual Scene Description: The scene depicts a child standing by the lake, holding a bright red balloon. The lake is calm, and the soft breeze makes the balloon sway gently. The warm afternoon sun adds a golden glow to the surroundings, creating a peaceful and joyful atmosphere.

Step 2 – Music Reasoning: The serene and joyful atmosphere of the visual scene suggests music that is peaceful yet uplifting. The gentle movement of the balloon and the calmness of the lake call for slow-paced, melodic music. The bright red balloon indicates a sense of lightheartedness and innocence, which can be reflected in the choice of instruments.

Predicted Music Elements:

- **Genre:** melodic, acoustic
- **Instruments:** piano, strings, acoustic guitar
- **Emotions:** peaceful, joyful
- **Rhythm:** slow, steady, gentle

Step 3 – Music Reasoning Refinement: The combination of a peaceful, calm scene with a sense of joy and innocence calls for music that complements the tranquility of the lake while still capturing the lighthearted joy of the child with the balloon. The melody should be soothing, with gentle instrumentation to reflect the overall dynamic of the scene.

Generated Music Description: A soft piano melody with light string harmonies, accompanied by gentle acoustic guitar strumming, creating a peaceful and warm atmosphere with a slow, steady rhythm.

B More Details about dataset

B.1 The rationale for selecting these datasets

The first part includes text-image and text-video paired data (Srinivasan et al., 2021; Wang et al.), with 10k samples each. The second part is based on the Emoset (Yang et al., 2023) image sentiment classification dataset, which results in 11k text-image pairs generated using the Qwen2-vl (Wang et al., 2024) for image annotation. The third part consists of 10k textual descriptions of visual scenes, generated using carefully designed prompts with the GPT-4o API. We chose the first part, text-image and text-video paired data (Srinivasan et al., 2021; Wang et al.), to provide high-quality visual-text pairings for the dataset. Although the CLIP model effectively encodes both visual and textual information into a shared space, a gap remains between the visual and textual modalities. By incorporating this data, we aim to mitigate the errors introduced by the modality gap during the training process. We chose the second part of the data (Yang et al., 2023) based on the fact that one of the most prominent features linking visual content to music is emotion. This choice allows the model to better capture the emotional connection between visuals and music. The third part of the data was chosen to address a potential limitation in the visual-text annotations of the first two datasets, where the information may be insufficient. Simple semantic annotations could leave too much room for interpretation, making it difficult for the model to converge. Therefore, it is essential to leverage large language models to generate diverse, visually rich scene descriptions that provide more detailed and comprehensive information.

B.2 Data generation prompt

To ensure the generated data has diversity, we combine different variables by generating reasonable prompts, which are then used to create visual text descriptions. Below are some template instructions and variables.

Variable Templates

Single Variable Templates

- Describe an image of a `natural_scenarios`.
- Imagine a `urban_scenarios` scene.
- Create a detailed description of an `indoor_scenarios` setting.
- Describe a scene that takes place during `time_periods`.
- Illustrate an image showcasing the `seasons` season.
- Imagine an image set in a `holidays` celebration.
- Describe a vivid picture with `excited_behaviors`.
- Imagine a melancholy scene with `sad_behaviors`.
- Create an image focusing on `angry_behaviors`.

- Describe a scene conveying `lonely_behaviors`.
- Imagine an image where someone is displaying `confident_behaviors`.
- Describe an artwork using `tonalities` tones.
- Create a picture that conveys a `atmospheres` mood.

Double Variable Templates

- Describe a `seasons` landscape with `animals` moving through it.
- Imagine a serene scene with `calm_behaviors` and `outdoor_natural_objects`.
- Describe a vivid image featuring `confident_behaviors` and `tonalities` shades.
- Illustrate a `urban_scenarios` environment during `holidays`.
- Create a peaceful image showing `natural_scenarios` and `atmospheres`.
- Describe a bustling scene featuring `urban_scenarios` and `character_actions`.
- Imagine an emotional setting with `character_emotions` and `sad_behaviors`.

Variable

- **Natural Scenarios:**
  - forest, tropical rainforest, mountain range, desert, beach, river, lake, waterfall, grassland, canyon,
  - wilderness, plateau, glacier, snowy mountains, hot springs, swamp, cave, volcano, cliff, sand dunes,
  - woodland, deep canyon, island in lake, sinkhole, flower field, orchard, rice field, terraced fields, bamboo forest, hills,
  - polar regions, aurora zone, seabed, coral reef, hot springs pool, stonehenge, beach rocks, mangrove forest, tundra, great plains,
  - valley, subtropical forest, vast wasteland, wetland, riverbank, mountain stream, primitive forest, beach palm trees, wilderness under the stars, path by the stream
- **Urban Scenarios:**
  - city street, square, shopping mall, park, subway station, train station, airport, residential area, commercial street, pedestrian street,
  - skyscraper, office building, restaurant, cafe, bar, museum, theater, cinema, school, library,
  - hospital, clinic, police station, stadium, amusement park, swimming pool, bus stop, bridge, dock, harbor,
  - vegetable market, night market, ancient city wall, open-air market, sculpture square, parking lot, night city lights, bus cabin, residential balcony, city skyline,
  - interchange bridge, clock tower, city tunnel, scenic street, commercial building, station waiting room, smart city corner, city rooftop garden, music fountain square, riverbank at night
- **Indoor Scenarios:**
  - bedroom, living room, kitchen, bathroom, study, dining room, office, conference room, laboratory, classroom,
  - library, theater, cinema, gym, swimming pool room, music room, dance room, yoga studio, cafe, restaurant,
  - game room, children's playroom, mall, gallery, museum, pet store, storage room, basement, attic, clinic,
  - waiting room, hospital ward, pharmacy, beauty salon, barbershop, hot spring room, garage, archive room, bookstore, hotel lobby,
  - hotel room, private cinema, photography studio, recording studio, greenhouse, exhibition hall, pet clinic, operating room, BBQ room, banquet hall

To ensure that the generated visual text descriptions contain rich visual details, we provide sample visual descriptions.

Example

- A broad, open field extends to the horizon, dotted with patches of wildflowers. A gentle breeze moves through the tall grasses, creating ripples across the landscape.

- A vast mountain range stretches across the horizon, its jagged peaks touching the sky. The landscape is bathed in the warm glow of the setting sun, casting long shadows over the valleys below.
- A fluffy gray cat's paws rest gently on the windowsill, its claws just visible. The rain outside drizzles down the window, with tiny droplets clinging to the glass.
- A white ceramic coffee cup sits on a wooden table, steam rising from the dark liquid inside. The sunlight reflects off the surface, casting a soft glow on the cup's edge.
- The vast ocean stretches out as far as the eye can see, crashing against the rocky cliffs below. The waves shimmer under the midday sun, and seagulls circle high above.
- A hand grips a fountain pen, the ink flowing steadily onto a piece of paper. The fingers, slightly aged, press gently against the smooth surface, leaving a dark trail behind.

Result

- An empty wooden bench sits beneath a tall oak tree, surrounded by fallen leaves. The sunlight filters through the branches, casting dappled shadows on the ground.
- A child holds a bright red balloon, standing at the edge of a calm lake. The balloon sways gently in the breeze, reflecting the afternoon sun.
- Snowflakes fall gently on a quiet street, covering parked cars and sidewalks in a soft white blanket. The streetlights glow faintly in the misty evening air.
- A lighthouse stands tall on a rocky shore, its beam of light cutting through the dark night. The waves crash below, sending sprays of water high into the air.
- A couple walks along a cobblestone path, lined with lanterns glowing softly in the night. The path winds through an ancient village, its buildings covered in ivy.
- The glowing embers of a campfire flicker against the dark forest. Shadows dance on the nearby trees as sparks rise into the cool night air.

B.3 The Details of evaluation dataset

To effectively evaluate the performance of different models in generating music based on visual content in an open visual domain, we considered potential biases in existing datasets. For instance, the URMP (Li et al., 2018) dataset is biased towards performance videos, SymMV (Zhuo et al., 2023) has a strong MV style, and TikTok (Zhu et al., 2022a) predominantly feature human subjects. Therefore, we extracted a portion of data from various datasets to create a combined evaluation set that better aligns with the open visual domain setting.

Specifically, for the V2M task, we selected 44 samples from the URMP dataset (Li et al., 2018), 100 samples each from the BGM909 (Li et al., 2024), SymMV(Zhuo et al., 2023), AIST++(Li et al., 2021), and TikTok (Zhu et al., 2022a) datasets, and then cropped the selected samples into 10-second segments. The final dataset contains a total of 2,600 samples.

For the I2M task, we applied a similar method, selecting samples from these datasets and randomly choosing keyframes. Given the potential weak correlation between images and music, we also included 1,000 image-music pairs from the IMAC (Verma et al., 2019) dataset, where emotion serves as the linking factor, to enhance the evaluation.Below is a summary of the datasets used:

- **URMP**(Li et al., 2018): A dataset for multimodal analysis of music performances, containing 44 classical music pieces. Each piece is assembled from separately recorded tracks of different instruments that are temporally aligned.
- **SymMV**(Zhuo et al., 2023): This dataset includes 1,140 video-music pairs, where the music consists of piano covers of popular songs and the videos are the corresponding official music videos. It spans over 10 music genres and is carefully curated from online sources.
- **BGM909**(Li et al., 2024): A dataset containing 909 piano pieces paired with well-aligned videos. All video-music pairs are manually edited to ensure perfect temporal and semantic alignment.
- **AIST++**(Li et al., 2021): A large-scale dataset focusing on street dance videos paired with copyright-cleared dance music. It is designed

to support research in dance information processing and is the first of its kind.

- **TikTok**(Zhu et al., 2022a): A collection of short (10–15 second) dance videos sourced from TikTok. It includes a wide variety of dance styles, clothing, expressions, and performer demographics. We manually curated over 300 single-person dance clips from monthly challenge compilations.
- **IMAC**(Verma et al., 2019): A labeled dataset capturing affective correspondence between images and music, useful for evaluating cross-modal emotion alignment.

C More Details of the theoretical proof of feasibility

Referring to (Qiao et al., 2022), we define the following notations: question Q, prompt T, probabilistic language model p_{LM}, and answer A. In the standard one-shot prompt setup, the prompt T_{sp} consists of an instruction I and a single question-answer pair. The large language model (LLM) takes the question Q and the prompt T as input and generates the answer A, as shown in Equation:

$$\mathcal{T}_{SP} = \{I, (x_1, y_1)\} \tag{17}$$

$$p(A \mid \mathcal{T}, \mathcal{Q}) = \prod_{i=1}^{|A|} p_{LM}(a_i \mid \mathcal{T}, \mathcal{Q}, a_{<i}) \tag{18}$$

In a one-shot CoT prompt setting, the prompt T_{CoT} includes instructions, questions, answers, and rationales. Instead of directly generating the answer, the model first produces a step-by-step reasoning process R, followed by the final answer A, as shown in the following equation:

$$\mathcal{T}_{CoT} = \{I, (x_1, e_1, y_1)\} \tag{19}$$

$$p(A, R \mid \mathcal{T}, \mathcal{Q}) = p(A \mid \mathcal{T}, \mathcal{Q}, R) \cdot p(R \mid \mathcal{T}, \mathcal{Q}) \tag{20}$$

$$p(R \mid \mathcal{T}, \mathcal{Q}) = \prod_{i=1}^{|R|} p_{LM}(r_i \mid \mathcal{T}, \mathcal{Q}, r_{<i}) \tag{21}$$

$$p(A \mid \mathcal{T}, \mathcal{Q}, R) = \prod_{j=1}^{|A|} p_{LM}(a_j \mid \mathcal{T}, \mathcal{Q}, R, a_{<j}) \tag{22}$$

We adapt the CoT methodology by introducing E_r into the video-to-music mapping. Here, E_v corresponds to $\mathcal{T}$ and $\mathcal{Q}$ in the CoT reasoning, E_m corresponds to the answer A, and E_r corresponds to the reasoning R:

$$P(E_m, E_r \mid E_v) = P(E_m \mid E_r, E_v) \cdot P(E_r \mid E_v) \tag{23}$$

The original VTM modeling is therefore transformed as follows:

$$E_m \sim P(E_m \mid E_r, E_v) \cdot P(E_r \mid E_v) \tag{24}$$

D More Evaluation Details

D.1 Explanation of Evaluation Metrics

FD measures the distance between the embedding distributions of synthesized and real samples. While FD uses PANNs (Kong et al., 2020) for embedding extraction, FAD utilizes VGGish (Hershey et al., 2017). KL calculates the divergence between the class outputs of generated and real music.

For the EMS metric, similar to the approach in (Wang et al., 2023), we first use the IMEMENT dataset with continuous VA emotional annotations. We begin by encoding the data into embedding representations using a pre-trained MERT (Li et al., 2023) model and the CLIP (Radford et al., 2021) model. Then, two separate multilayer perceptrons are trained to predict the VA values for music and video. The EMS metric is subsequently calculated based on these predictions:

$$\text{sim}(x, y) = \sqrt{((v_x - v_y)^2 + (a_x - a_y)^2)} \tag{25}$$

For the P@10 metric, we adopt the approach described in [ ]. Given a piece of generated music and its corresponding ground-truth music for the condition video , we randomly select 69 music pieces. The MERT model (Li et al., 2023) is employed to extract the music features for each generated sample. If the ground-truth music ranks within the top-10, we consider it a successful retrieval. The final precision score, P@10, is calculated by evaluating the successful retrieval rate across all generated samples.

E More Details of the experiments

Distilling CoT into M2UGen. We fine-tuned M2UGen with our CoT-generated supervision for comparison. The results, shown in Table 7, indicate that the fine-tuned M2UGen achieves performance close to CoT-VTM across most metrics,

demonstrating the effectiveness of our distilled CoT knowledge. However, due to M2UGen's reliance on a large LLM-based pipeline, its inference time and memory consumption are substantially higher than those of CoT-VTM, limiting its practicality for efficient deployment.

Replacing CLIP with ImageBind. We investigated replacing the CLIP text encoder with ImageBind, a more powerful cross-modal encoder with longer input tolerance and stronger multi-modal alignment. As Table 6 shows, ImageBind improves both generation quality and alignment metrics compared to CLIP. This result validates that our framework supports flexible encoder substitution and benefits from stronger visual-text grounding.

Encoder	FD↓	FAD↓	KL↓	IBS↑	P@10↑
ImageBind	33.764	3.979	1.208	0.223	33.78
CLIP	34.004	4.037	1.212	0.210	31.46

Table 6: Comparison between CLIP and ImageBind as visual encoders in our framework. ImageBind shows consistent improvements across all metrics.

Method	FD↓	FAD↓	KL↓	P@10↑	Inference
M2UGen	34.764	4.179	1.218	29.78	9.74s + (4.99s)
CoT-VTM	34.004	4.037	1.212	31.46	2.33s + (4.99s)

Table 7: Comparison between M2UGen with CoT supervision and CoT-VTM. CoT-VTM achieves comparable performance with substantially lower inference cost.