V2Meow: Meowing to the Visual Beat via Video-to-Music Generation

Kun Su[*†3], **Judith Yue Li**[*1], **Qingqing Huang**[†4], **Dima Kuzmin**[1], **Joonseok Lee**[1,5], **Chris Donahue**[2,6], **Fei Sha**[1], **Aren Jansen**[1], **Yu Wang**[†7], **Mauro Verzetti**[2], **Timo Denk**[2]

[1]Google Research, [2]Google DeepMind, [3]University of Washington, [4]ByteDance,
[5]Seoul National University, [6]Carnegie Mellon University, [7]Spotify

Abstract

Video-to-music generation demands both a temporally localized high-quality listening experience and globally aligned video-acoustic signatures. While recent music generation models excel at the former through advanced audio codecs, the exploration of video-acoustic signatures has been confined to specific visual scenarios. In contrast, our research confronts the challenge of learning globally aligned signatures between video and music directly from paired music and videos, without explicitly modeling domain-specific rhythmic or semantic relationships. We propose V2Meow, a video-to-music generation system capable of producing high-quality music audio for a diverse range of video input types using a multi-stage autoregressive model. Trained on 5k hours of music audio clips paired with video frames mined from in-the-wild music videos, V2Meow is competitive with previous domain-specific models when evaluated in a zero-shot manner. It synthesizes high-fidelity music audio waveforms solely by conditioning on pre-trained general-purpose visual features extracted from video frames, with optional style control via text prompts. Through both qualitative and quantitative evaluations, we demonstrate that our model outperforms various existing music generation systems in terms of visual-audio correspondence and audio quality. Music samples are available at tinyurl.com/v2meow.

Introduction

Recent advancements in high-resolution neural audio codecs (Zeghidour et al. 2021a; D'efossez et al. 2022; Kumar et al. 2023) have introduced novel possibilities for directly generating high-quality music waveforms comparable to human-made music (Borsos et al. 2022; Agostinelli et al. 2023). Notably, AudioLM (Borsos et al. 2022) employs a multi-stage autoregressive modeling approach for audio generation. It utilizes a masked language model pre-trained on tokenized audio encodings (Chung et al. 2021) to capture long-term structures and the discrete neural audio codec (Zeghidour et al. 2021a) for high-quality synthesis. Music language models like MusicLM (Agostinelli et al. 2023) have further demonstrated that autoregressive models can generate music conditioned on text prompts (Agostinelli et al. 2023; Copet et al. 2023) or input vocals (Donahue et al. 2023). However, the challenge persists in video-conditioned music generation, where a gap exists between the coarser global video signature and the aforementioned high-resolution audio representations or audio codecs for waveform generation.

Most existing work on video-to-music generation focuses on modeling the audiovisual correspondence between domain specific video representation and symbolic representation of music. For example, dance-to-music literature (Zhu et al. 2022a,b; Yu et al. 2023) relies on modeling music rhythms, style from annotated human motion, other works focuses on generating natural sound that is faithful to the physical motion in the silent video (Su, Liu, and Shlizerman 2020a; Gan et al. 2020; Owens et al. 2016; Zhou et al. 2018; Chen et al. 2020; Su et al. 2023), for example, reconstructing instrumental music from silent instrument performance videos. As a result, the input video types are restricted to certain visual scenarios, and cannot be generalized to arbitrary video input types, e.g., a cat video or slideshows of images. In contrast, we propose to learn a general audiovisual correspondence directly from paired video frames and music waveform data. Specifically, we bridge the gap between the coarser video representation and the high-resolution audio representation through finding the video-audio aligned low-resolution representation space. This approach allows us to generalize to a wide range of video input types, and leverage the vast amount of parallel music and video data available on the internet for scaling.

We introduce *V2Meow*, a high-fidelity music audio waveform generator conditioned on diverse video inputs. Drawing inspiration from MusicLM (Agostinelli et al. 2023), we employ a multi-stage autoregressive language modeling approach. V2Meow takes video frames as input with optional style control through text prompts, treating video and text as a unified input stream fed into the Transformer with feature-specific adaptors. By not explicitly modeling domain-specific audiovisual correspondence, V2Meow exhibits zero-shot transfer capabilities, demonstrated in evaluations on AIST++ dance videos (Li et al. 2021). Quantitative and qualitative assessments of music-video correspon-

[*]These authors contributed equally.
[†]Work done while at Google.
Correspondence to: Judith Yue Li (judithyueli@google.com)

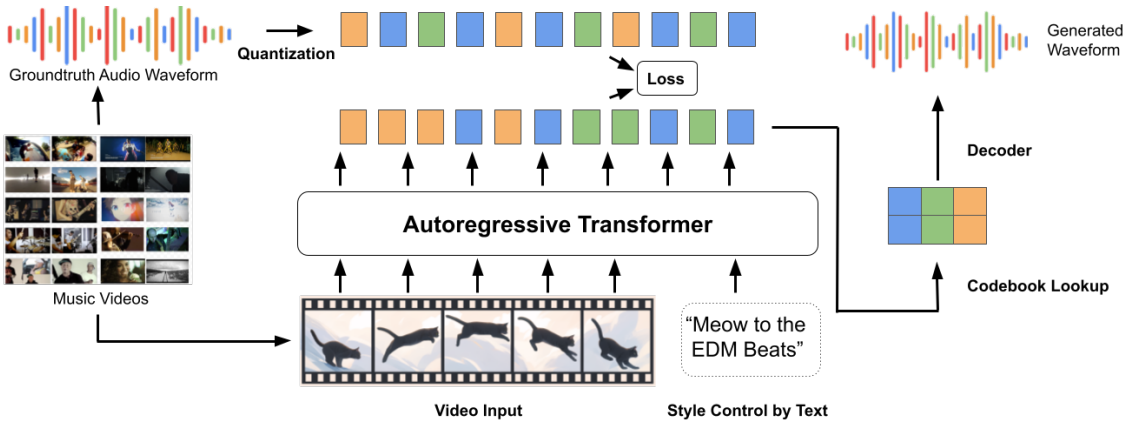

Figure 1: The video-to-music generation model V2Meow synthesizes high-fidelity music conditioned on video input and optionally text describing high-level style.

dence, along with an extensive ablation study, elucidate factors influencing generation quality. Numerical and human study results affirm that, compared to MIDI-based baselines, V2Meow aligns better with human music preferences. Additionally, including video input enhances V2Meow's visual relevance learning compared to text-only input.

Related Work

Video to Audio. Over the past few years, there have been advancements in deep-learning approaches to produce realistic sounds from silent videos. Owens et al. (2016) initially predicted impact sound features based on image features and then retrieved the closest sound sample from the dataset instead of directly generating the sound. Chen et al. (2017) investigated the utilization of conditional GAN for generating sound from images, although the experiments were constrained to music performances collected in a laboratory setting. Visual2Sound (Zhou et al. 2018) suggested generating audio waveform from videos captured in-the-wild, but the dataset is limited to only ten types of sounds. Subsequently, novel loss functions (Chen et al. 2018, 2020) are introduced to enhance the semantic alignment of generated audio from videos. The effectiveness of models for generating audio from videos is primarily constrained by the typically weak correspondence between the video and audio, as well as the limited scale of training data. Generating high-quality audio from a silent video poses a challenge without a robust audio generative model and adequate audio representations.

Video to Music. Apart from natural sounds, several studies have delved into the generation of music from videos. Initial efforts focused on generating symbolic music (MIDI) from videos depicting a musician playing the piano (Koepke et al. 2020; Su, Liu, and Shlizerman 2020a) or other instruments (Gan et al. 2020; Su, Liu, and Shlizerman 2020b). Later, RhythmicNet (Su, Liu, and Shlizerman 2021) showcased the potential to generate music soundtracks synchronized with arbitrary human body movements. Following studies have expanded the generation of symbolic music representations to encompass waveform generation (Zhu et al. 2022b,a; Yu et al. 2023). Nevertheless, the music generation systems proposed still depend on intricate motion extractors to explicitly capture visual rhythm from domain-specific data, such as dance videos. Apart from relying on visual cues from human motion, a music Transformer has been suggested for the generation of video background music (Di et al. 2021), albeit through MIDI generation. All these approaches face challenges due to their domain-specific modeling assumptions, such as focusing on music from particular instruments or relying heavily on visual cues like human body motion. Consequently, the generated samples are constrained to specific instrument types and visual scenarios. In contrast to the aforementioned studies, our approach utilizes music videos captured in real-world settings to establish a general mapping from visual input to audio waveforms.

Music Generation. A robust music representation such as MIDI has been widely employed in the modeling of music. Early studies transformed MIDI into piano-roll representation by employing GANs (Dong et al. 2018) or variational autoencoders (Roberts et al. 2018; Gillick et al. 2019) to generate novel music. Subsequently, there were proposals for event-based representations aimed at more efficient representation of MIDI (Oore et al. 2020; Huang et al. 2018; Hawthorne et al. 2018; Huang and Yang 2020). Control signals are additionally integrated into the music generative models based on MIDI (Engel et al. 2017; Choi et al. 2019; Lattner and Grachten 2019).

In terms of modeling music directly from raw audio without the need for transcripts or symbolic music representations, WaveNet (Oord et al. 2016) introduced autoregressive modeling to synthesize music audio with satisfactory quality. Jukebox (Dhariwal et al. 2020) adopted a hierarchical approach to generate tokens at different temporal resolutions, which were subsequently combined to reconstruct music. Recent work on high quality audio representation (Zeghidour et al. 2021a; D'efossez et al. 2022; Kumar et al. 2023) directly apply residual vector quantization on the raw waveform. Later, several recent works adopted such representation for text-to-audio generation using transformer-based autoregressive models, e.g., AudioLM (Borsos et al. 2022), Mubert (Mubert-Inc. 2022), Au-

dioGen (Kreuk et al. 2022) and MusicLM (Agostinelli et al. 2023) or non-autoregressive models (Copet et al. 2023; Garcia et al. 2023). Alternatively, Riffusion (Forsgren and Martiros 2022) and other recent work (Huang et al. 2023a; Liu et al. 2023; Huang et al. 2023b; Schneider et al. 2023) adopted a diffusion based approach.

The Proposed Method: V2Meow

In this section, we describe the feature representations and modeling pipeline of our proposed method, V2Meow, in detail.

Feature Representations

For audio waveforms, we follow MusicLM (Agostinelli et al. 2023), using semantic and acoustic tokens extracted from two different pre-trained self-supervised models. For visual inputs, we explore various types of visual features to find the most informative representation suitable for music generation task. Finally, we demonstrate how we represent the control signal without paired music-video-text examples.

Semantic Music Tokens. We extract semantic tokens using a pre-trained w2v-BERT model (Chung et al. 2021). w2v-BERT is a self-supervised model using masked language modeling (MLM) with contrastive loss, coarsely learning to represent audio, capturing both local dependencies such as local melody in music and global long-term structure such as harmony and rhythm. To obtain the semantic tokens, we extract embeddings from an intermediate layer of w2v-BERT. We then apply k-means algorithm with K_s clusters on these embeddings and use the centroid indices as semantic tokens. For each audio waveform, we obtain semantic tokens $\{S_t : t = 1, ..., T_s\}$, where T_s is the total number of tokens. While the coarse resolution of the semantic tokens enables us to model long-term dependencies, the audio reconstruction solely from these semantic tokens usually leads to poor quality.

Acoustic Music Tokens. To generate high-quality music audio, we additionally rely on acoustic tokens extracted from a pre-trained SoundStream (Zeghidour et al. 2021a) model. SoundStream is a universal neural audio codec that compresses arbitrary audio at low bit rates and reconstructs the audio back in a high quality. Specifically, a convolutional encoder embeds the input waveform, followed by residual vector quantization (RVQ) to discretize them. RVQ is a hierarchical quantization scheme composing a series of N vector quantizers, where the target signal is reconstructed as the sum of quantizer outputs. Each quantizer with a vocabulary size of K_a learns to quantize the embedding simultaneously during training. Thanks to the residual quantization, the acoustic tokens have a hierarchical structure such that tokens from the coarse quantizers recover acoustic properties like music recording conditions, while leaving only the fine acoustic details to the fine quantizer tokens. Since coarser levels are more important for high-fidelity reconstruction as illustrated in AudioLM, we first construct a mapping from music semantic tokens to coarse acoustic tokens and then learn a mapping from coarse acoustic tokens to fine-grained acoustic tokens in the later stage.

Model Conditioning

Video Frames Conditioning. Given a video as a sequence of T frames, $\left\{\mathbf{v}_t \in \mathbb{R}^{H\times W\times 3} : t = 1, ..., T\right\}$, we aim to extract useful visual features from existing pre-trained visual model. We explore various visual representations for this, among pure visual models, multimodal models, and quantized models, since it is unclear which kind of visual representation could provide sufficient information for music generation. In particular, we explored a combination of the following visual features as.

1. *Purely visual representations*: Video understanding models learn underlying patterns from the pixel distributions observed in a collection of images or videos, using CNNs (Tran et al. 2015; Carreira and Zisserman 2017; Feichtenhofer 2020) or Transformers (Arnab et al. 2021; Bertasius, Wang, and Torresani 2021), without access to additional modalities. As it is common that the visual changes in a video have correspondences to musical rhythm, we adopt the Inflated 3D (I3D), which explicitly considers the optical flow, which is known to be useful for analyzing motions. In our experiments, we denote the visual flow embeddings as $\left\{\mathbf{f}_t \in \mathbb{R}^{D_f} : t = 1, \ldots, T\right\}$, extracted from an I3D model pretrained on Kinetics (Carreira and Zisserman 2017), where D_f indicates the dimensionality of the I3D features.
2. *Multimodal embeddings*: The second type is an embedding learned from multimodal correspondence, in addition to the visual modality. Contrastive Language Image pre-training (CLIP) (Radford et al. 2021) is a popular image-text model, widely used in a variety of downstream tasks including generative applications. We expect its generality and robustness would be potentially useful to incorporate semantics of the video. We denote the CLIP embeddings as $\{\mathbf{c}_t \in \mathbb{R}^{D_c} : t = 1, \ldots, T\}$, where D_c is the dimensionality of the CLIP embedding.
3. *Visual tokens*: Since the semantic and acoustic music representations in our pipeline are both discrete tokens, we explore using a similar type of discrete tokens for visual inputs. To obtain discrete tokens for a video frame, we adopt ViT-VQGAN (Yu et al. 2021), the state-of-the-art self-supervised Vision Transformer (ViT) model that performs image quantization on each image to obtain a set of discretized latent codes and uses a Transformer to predict these image tokens autoregressively for image reconstruction. Given a video frame $\mathbf{v}_t \in \mathbb{R}^{H\times W\times 3}$, where H, W indicate the image height and width, respectively, the VIT-VQGAN encodes the image into $H/D_q \times W/D_q$ discretized latent codes, where D_q is the size of non-overlapping image patches mapped to one token. A video with T frames is represented as a set of tokens $\left\{Q_t \in \mathbb{Z}_+^{H/D_q\times W/D_q} : t = 1, \ldots, T\right\}$.

Text Conditioning. Since the music for a video could be highly dependent on personal preference, we allow users to optionally provide a music-related text description, in addition to the visual input, to control the generated music at a high-level. However, it is challenging to collect music-video-text pairs in the wild. To overcome this is-

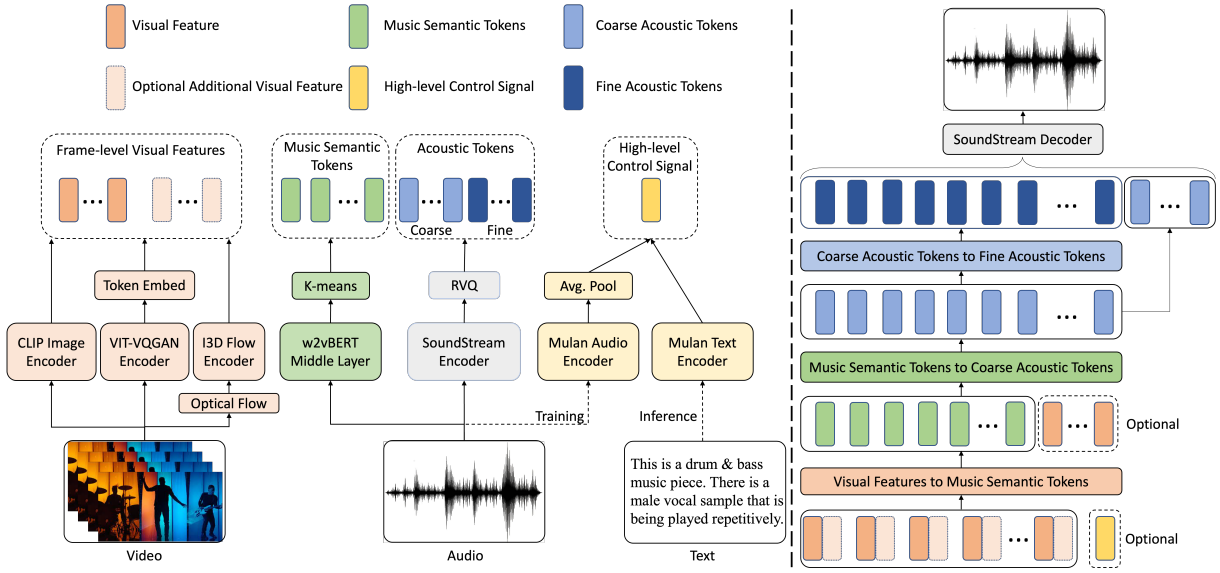

Figure 2: V2Meow Architecture Overview: (left) Feature extraction pipeline for video, audio and text representations. (right) Overview of multi-stage video to music modeling.

sue, we leverage a music-text joint embedding model, MuLan (Huang et al. 2022), which is trained on paired music-text data using a contrastive loss. For each video in the training set, V2Meow first extracts MuLan embeddings of all audio segments $\{\mathbf{m}_j \in \mathbb{R}^{D_m} : j = 1, \ldots, J\}$, where each segment is ten second long and J indicates the total number of audio segments in the video, $D_m = 128$ is the dimension of MuLan audio embedding. We then average the embeddings into a single video-level. It is worth noting that we extract a fixed-length segment at a random starting point from a music video at each training iteration. Although it would be ideal to perform inference on MuLan with the selected segment only, we avoid doing this aiming for an efficient experiment. Instead, we use a video-level embedding for the entire video and empirically verify that this is sufficient, probably because our goal is to have a high-level control on the style of generated music, instead of fine-grained control. At inference time, we may use a music-related text description to obtain a MuLan embedding and condition our V2Meow model on it.

Modeling Pipeline

We adapt the AudioLM pipeline to train the visual conditioned music generation model. There are three main stages of sequence-to-sequence modeling tasks.

Stage 1. Visual Features to Music Semantic Tokens. In the first stage, V2Meow learns a mapping from visual inputs to the music semantic tokens. Specifically, we use an encoder-decoder Transformer (Vaswani et al. 2017) where the encoder takes the visual features, and the decoder predicts the music semantic tokens autoregressively. It turns out that this stage is the most critical part of generating music which reflects the video well. On the one hand, it builds up the connection between visual and audio modalities, and models the semantic transformation from visual information to audio. On the other hand, this stage does not output high-quality fine-grained audio, allowing the model to focus on associating the two modalities with each other.

Stage 2. Music Semantic Tokens to Coarse Acoustic Tokens. In the second stage, we aim to convert the music semantic tokens to acoustic tokens for high-quality synthesis. We follow the AudioLM pipeline to split this stage into coarse and fine acoustic modeling. In the coarse acoustic modeling, we explore two different training strategies: 1) We follow AudioLM to train a decoder-only Transformer to map music semantic tokens to coarse acoustic tokens. Since such a training strategy does not require visual information, we could use the large-scale music data in the wild to pre-train a robust model. 2) We also explore to see whether adding visual conditioning at this stage improves the performance. Specifically, we train an encoder-decoder Transformer model where the encoder takes the visual features and music semantic tokens and the decoder generates the coarse acoustic tokens. While ideally the second approach should work better, the final results are not necessarily better since the amount of music-video pairs available is still incomparable to the amount of music-only data.

Stage 3. Coarse to Fine Acoustic Tokens and Audio Decoding. Once we have the coarse acoustic tokens, we follow AudioLM (Borsos et al. 2022) and perform coarse to fine acoustic tokens modeling. This stage maps the tokens in the first N_c levels of SoundStream RVQ to the tokens of the remaining N_f levels. Finally, all levels of tokens are passed to SoundStream Decoder to reconstruct the audio.

Adding Control. To incorporate the control signal into V2Meow during training, we simply feed the MuLan audio embedding as an additional input with a sequence length be one to the Transformer encoder along with the visual features in the first stage. Both Mulan audio embedding and visual features are projected to the same feature dimension.

At inference, we instead use the MuLan text embedding with the visual features to generate the semantic tokens.

Experiments

Experimental Settings

Training Datasets. Following (Surís et al. 2022), we filtered a public available video dataset (Abu-El-Haija et al. 2016) to 110k videos with the label Music Videos and refer to it as MV100K. The training and validation datasets were split into an 80:20 ratio. We trained the Stage 1 model and Stage 2 model on 5k hours of music videos. A version of Stage 2 model is trained on audio-only data for ablation study. For computing semantic and acoustic audio tokens, we adopt the SoundStream tokenizer and w2v-BERT tokenizer, both of which are pre-trained on 46k hours of music only audio data sampled at 16kHz sampling rate.

Evaluation Datasets. We evaluate our methods on three different datasets. For the task of video conditional music generation, we use the test partition of the MV100K. We select 13 genres of music videos to comprise a genre balanced subset with a total number of 4076 videos. For the task of video and text conditional music generation, we use the latest MusicCaps dataset (Agostinelli et al. 2023) which is a subset of AudioSet (Gemmeke et al. 2017). The MusicCaps has about 5.5k human annotated text captions, music, and video pairs. With the text caption, we can verify whether the generated music could be controllable and whether its performance is comparable with text-to-music generation models like MUBERT (Mubert-Inc. 2022) and Riffusion (Forsgren and Martiros 2022). For both tasks, we generate ten second audio for each video clip. For dance-to-music generation task, we evaluate temporal alignment on 20 dance videos in the test split of AIST++ (Li et al. 2021). The evaluation is in zero-shot fashion without any fine-tuning on the AIST++ train split, and only video frames are used for modeling while no motion data is involved. The reported metrics represent averages over 20 10-second audio segments and 86 2-second audio segments, with 5 inference examples per segment.

Implementation Details. For all visual features, we use a frame rate at 1 fps, following the standard on MV100K (Abu-El-Haija et al. 2016). We use the released ViT-L/14 model[1] to extract the CLIP embeddings, whose dimensionality is 768. For computing the I3D Flow embeddings, we use a model pre-trained on the Kinetics dataset, whose dimensionality is 1024. We use a pre-trained VIT-VQGAN encoder to obtain 1024 tokens for each image and the vocabulary size is 8192. For the visual feature to music semantic tokens modeling, we use encoder-decoder Transformer with 12 layers, 16 attention heads, an embedding dimension of 1024, feed-forward layers of dimensionality 4096, and relative positional embeddings. We use 10-second random crops of the music video for visual to music semantic tokens modeling and semantic tokens to coarse acoustic tokens modeling. The coarse to fine acoustic tokens modeling is trained on 3-second crops. During inference, we use temperature sampling for all stages, with temperatures $\{1.0, 0.95, 0.4\}$ for modeling stages 1, 2, and 3, respectively.

[1] huggingface.co/sentence-transformers/clip-ViT-L-14

Evaluation Metrics

Objective Metrics. We follow (Agostinelli et al. 2023) to use different quantitative metrics to automatically assess the fidelity, the semantic relevance of the generated samples and (Zhu et al. 2022a,b) to evaluate rhythmic alignment.

- **Audio Quality.** We use Fréchet Audio Distance (FAD) based on two audio embedding models to measure different aspects of the audio quality, both of which are publicly available (1) TRILL[2] (Shor et al. 2020), which is trained on speech data, and (2) VGGish[3] (Hershey et al. 2017), which is trained on the public audio event dataset (Abu-El-Haija et al. 2016; Lee et al. 2018).
- **Semantic Relevance.** KL Divergence (KLD) (Yang et al. 2022; Kreuk et al. 2022) and MuLan Cycle Consistency (MCC) (Agostinelli et al. 2023) is used to determine whether generated music is semantically relevant to the reference audio or text. We run a LEAF classifier (Zeghidour et al. 2021b) for multi-label classification on AudioSet, and use KLD over the predicted class probabilities between the original audio and the generated audio to evaluate if they share similar concept. For the video to music task, we use MuLan audio embedding of ground truth audio as reference to compute MCC as the average cosine similarity between the MuLan audio embedding of the generated music audio and reference. For the video and text to music task, we use the MuLan text embedding of the text description as reference instead to check text adherence.
- **Rhythmic Alignment.** (Zhu et al. 2022a,b) introduces Beats Coverage Scores (BCS) and Beats Hit Scores (BHS) to count the aligned rhythm points of synthesized music and ground-truth music. BCS refers to the fraction of generated musical beats by the ground truth musical beats, while BHS refers to the ratio of aligned beats to the ground truth beats. Here we adopt the adjusted BCS and BHS introduced in Yu et al. (2023) and compute F1 score in addition.

Subjective Metrics. Whether the video and background music match is subjective. The generated music can be a reasonable match to the video, even if it is not similar to the ground truth music that accompanies the original video. Thus we conduct a human study to measure visual relevance and music preferences. Specifically, we sampled 89 distinct video examples from the MV100K test set and 76 distinct video examples from the MusicCaps test set. We surveyed around 200 participants individually, and each participant was asked to evaluate a pair of videos with the same video but different background music. Each video pair is rated by 3 person. A total number of 3500 ratings are collected in the end.

- **Visual Relevance**. We asked human raters to conduct a side-by-side comparison of the music generated from the baseline models (CMT, Riffusion, or MUBERT) and the five different V2Meow model variants by answering the question: "Which music do you think goes best with the

[2] tfhub.dev/google/nonsemantic-speech-benchmark/trill/3

[3] tfhub.dev/google/vggish/1

Method	Visual	Text	Semantic / Semantic + Acoustic Modeling FAD TRILL ↓	FAD VGG ↓	KL Div. ↓	MCC ↑	Semantic + Acoustic Modeling Visual Relevance ↑	Music Preference ↑
Eval Dataset: MV100K								
CMT	✓	✗	N/A	N/A	N/A	N/A	20.6%	30.0%
Random Shuffle	✗	✗	-	-	0.67	0.268	N/A	N/A
V2Meow-CLIP	✓	✗	0.236/0.158	6.094/2.779	0.63/0.54	0.312/0.372	78.2%	67.6%
V2Meow-I3D	✓	✗	0.236/**0.151**	6.278/2.328	0.77/0.65	0.279/0.296	74.3%	65.8%
V2Meow-VIT	✓	✗	0.240/0.174	6.097/1.988	0.73/0.62	0.276/0.294	81.4%	71.5%
V2Meow-VIT+I3D	✓	✗	0.236/0.178	5.801/**1.945**	0.68/0.57	0.298/0.327	83.8%	76.8%
V2Meow-CLIP+I3D	✓	✗	0.235/0.165	6.126/2.003	0.64/**0.49**	0.343/**0.419**	79.2%	68.2%
Eval Dataset: MusicCaps								
CMT	✓	✗	N/A	N/A	N/A	N/A	19.7%	20.7%
Riffusion	✗	✓	0.760	13.4	1.19	0.34	38.6%	41.2%
MUBERT	✗	✓	0.450	9.6	1.58	0.32	43.3%	49.3%
V2Meow-CLIP	✓	✓	0.379/**0.328**	5.198/4.628	1.31/1.19	0.364/0.377	63.6%	58.5%
V2Meow-I3D	✓	✓	0.389/0.331	5.190/**4.623**	1.26/1.22	0.377/0.371	68.8%	68.0%
V2Meow-VIT	✓	✓	0.377/0.366	4.970/5.039	1.34/1.23	0.380/0.392	66.9%	60.3%
V2Meow-VIT+I3D	✓	✓	0.381/0.359	5.094/4.819	1.34/1.21	0.379/0.389	71.8%	67.1%
V2Meow-CLIP+I3D	✓	✓	0.391/0.349	5.385/4.948	1.27/**1.19**	0.369/**0.394**	67.4%	65.8%

Table 1: Quantitative evaluations on MV100K and MusicCaps for different models. For FAD and KL Divergence, lower is better. For MCC, higher is better. Bold font indicates the best value. The five V2Meow variants are named based on the video features used as input. Semantic Modeling indicates video conditioning is used only for semantic modeling, while Semantic + Acoustic Modeling indicates video conditioning is used for both semantic and acoustic modeling.

Model (Length)	Beat Coverage	Beat Hit	F1
GT	100	100	100
V2Meow CLIP+I3D (10s)	100.0 (0.00)	84.4 (25.1)	91.5
V2Meow CLIP+I3D (6s)	99.3 (8.64)	84.7 (25.7)	91.4
V2Meow CLIP+I3D (2s)	90.0 (30.0)	84.8 (32.1)	87.3
CDCD Step-Intra (6s)	87.9	83.2	85.5
D2M-GAN (2s)	88.2	84.7	86.4
CMT (2s)	85.5	83.5	84.5

Table 2: Zero-shot evaluation results on AIST++. For CMT and V2Meow only video frames are used as input, while CDCD Step-Intra and D2M-GAN requires additional motion annotation as inputs. For each video input we randomly generate 10 music samples and report the average score and standard deviation.

video?". They are asked to ignore the sound quality and only focus on how well the music matches the video.

- **Music Preference**. We asked human raters to choose which music they prefer to hear and ignore the video content. This task aim to study whether the generated music is aligned with human perceptual preference. Here we ask the listener to ignore sound quality and tell us which music they like.

Results

We initiate our evaluation of V2Meow's video-to-music generation capabilities by comparing it to the state-of-the-art model CMT (Di et al. 2021), which relies on video-driven symbolic music representations. This evaluation is conducted on the MV100K dataset. Subsequently, we extend our comparison to two text-to-music systems, Mubert (Mubert-Inc. 2022) and Riffusion (Forsgren and Martiros 2022), using the MusicCaps dataset augmented with videos. The objective here is to assess the impact of incorporating video frames as conditioning signals in additional to text. For the task of dance-to-music generation, we further compare V2Meow with baseline models D2M-GAN (Zhu et al. 2022a), CDCD Step-Intra (Zhu et al. 2022b), and CMT (Di et al. 2021) on the AIST++ test split, aiming to evaluate V2Meow's understanding of complex dance motion. Detailed results are presented in Table 1 and Table 2. The assessment concludes with an ablation study that dissects the significance of each stage in the modeling pipeline.

Video Conditional Music Generation

In the MV100K dataset, introducing video conditioning during the acoustic modeling stage notably enhances both audio quality-related metrics and semantic relevance, as illustrated in the second column of Table 1. Notably, when conditioned on specific visual embeddings, we find that clip embedding attains the highest MCC score, whereas I3D flow embedding exhibits superior performance in FAD metrics. This implies that different visual features capture distinct aspects within the video-music aligned subspace. The combination of Clip and I3D Flow embeddings achieves the highest MCC score across all models, with a corresponding enhancement in FAD VGGish compared to models with either Clip or I3D Flow embedding alone. While VIT-VQGAN tokens do not surpass others in individual metrics, the amalgamation of VIT-VQGAN tokens and I3D Flow embedding demonstrates improved performance compared to a single visual input. In terms of visual relevance and music preference,

V2Meow significantly outperforms CMT by a substantial margin, as indicated in the third column of Table 1.

Video and Text Conditional Music Generation

In the MusicCaps evaluation, our approach, enhanced by the inclusion of video frames as an additional control, demonstrates a 20-30% improvement in visual relevance compared to Riffusion and MUBERT that only condition on text. It's noteworthy that our approach also achieves lower FAD and higher MCC scores. Here the MCC is the similarity between generated music audio and text. This indicates that augmenting the conditioning with video frames not only enhances visual relevance but also contributes to improved audio quality and text adherence, despite utilizing a relatively small dataset of 5,000 hours of music videos. The combination of Clip and I3D Flow embeddings maintains the highest KLD and MCC scores, while the combination of VIT-VQGAN tokens and I3D Flow embeddings achieves the best visual relevance. Across all variations, V2Meow consistently outperforms the baseline in terms of audio quality, text adherence, visual relevance and music preference.

Dance to Music Generation

The zero-shot evaluation on dance videos in the AIST++ test split (Li et al. 2021) reveals that V2Meow can attain performances comparable to those of specialized dance-to-music generation baselines (Zhu et al. 2022a,b), measured by beat coverage and beat hit. This evaluation is conducted in a zero-shot fashion, without any fine-tuning on the AIST++ training split. Only video frames are used for modeling, and no human-annotated motion data is involved. The results suggest that our proposed framework can adeptly handle videos with significant rhythm changes, even when dance motion occurs below our 1fps sampling rate. It's important to note that D2M-GAN and CDCD Step-Intra utilize the AIST++ training split for fine-tuning, requiring a much higher sampling rate and additional motion annotation as input. In contrast, our model and CMT exclusively take video frames as input.

Ablation Study

Illustrated in Figure 3, we conducted additional ablation studies to assess the contribution of each stage, (a) on MV100K, measured using FAD VGGish score (↓), and (b) on MusicCaps, measured by MCC score (↑). The results in the figure suggest that video conditioning is crucial for both semantic and acoustic modeling. Furthermore, directly modeling acoustic tokens without semantic tokens proves suboptimal compared to the two-stage modeling in terms of audio quality and semantic consistency. Directly predicting acoustic tokens from CLIP visual features results in a degradation of the FAD score from 2.779 to 3.331 and a reduction in semantic alignment from 0.377 to 0.275. This novel two-stage design facilitates the implicit learning of both coarse-grained (style) and fine-grained (rhythm) semantics shared between music and video, enabling the generation of visually relevant soundtracks. Notably, this is achieved despite training solely on 5,000 hours of music videos, where video and audio are not perfectly semantically aligned, unlike in other tasks such as text-to-speech.

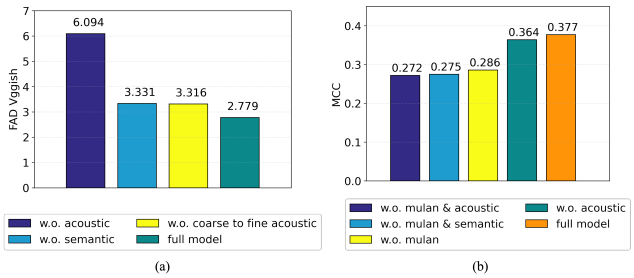

Figure 3: (a) Ablation study on the contribution of each component for MV100K dataset using FAD VGGish score, the lower the better. (b) Ablation study on the contribution of each component of for MusicCaps dataset using MCC score between text and generated audio, the higher the better.

Conclusion

We introduce V2Meow, a versatile soundtrack generation model capable of producing high-fidelity music for a diverse array of video inputs. To bridge the gap between coarser global video signatures and high-resolution audio representations designed for waveform reconstruction, our approach learns the general video-audio-aligned low-resolution representations implicitly, without modeling domain-specific audiovisual correspondence. Our hierarchical two-stage autoregressive modeling is crafted to achieve optimal music generation quality. The semantic modeling stage embeds video and audio in the same semantic space, while the acoustic modeling stage models the high-resolution audio representation space in a coarse-to-fine fashion. What sets this two-stage approach apart is its decoupling; only the semantic modeling stage needs to be trained on paired data with semantic video and audio features. The acoustic modeling stage can be trained on audio-only data and optionally fine-tuned on audio data from video to address domain shift. Zero-shot evaluation on dance videos suggests that the temporal and semantic correlation between video and music learned from 5,000 music videos can be transferred to unseen video input types. Experimental results demonstrate that, compared to MIDI-based generation models, V2Meow can generate music more aligned with visual content and human perception. Furthermore, compared to text-to-music generation models, additional video conditioning results in higher music-video correspondence and better audio quality. Ablation studies underscore the critical role of video conditioning in both semantic and acoustic modeling stages for generating high-fidelity sounds from video inputs, emphasizing that directly generating acoustic tokens without semantic tokens leads to a degradation in generation quality.

Ethical Statement

Large generative models learn to imitate patterns and biases inherent in the training set, and in our case the model might propagate the potential biases built in the video and music corpora used to train our models. Our introduction

of text control allows us to debias undesirable stereotypical video-to-music associations. These biases can originate from the video and music corpora used during training, leading to skewed genre distributions and unequal representation of gender, age, and ethnic groups within each genre. These concerns extend to learned visual-audio associations, which can result in stereotypical links between video content (e.g., people, body movements, dance styles, locations, or objects) and a limited set of musical genres. Additionally, derogatory associations may arise between video choreography and audio output (e.g., minstrelsy, parody, miming).

Acknowledgments

We are grateful for having the support from Jesse Engel, Ian Simon, Hexiang Hu, Christian Frank, Neil Zeghidour, Andrea Agostinelli, David Ross and authors of MusicLM project for sharing their research insights, tutorials and demos. Many thanks to Austin Tarango, Leo Lipsztein, Fernando Diaz, Renee Shelby, Rida Qadri and Cherish Molezion for reviewing the paper and supplementary materials and share valuable feedbacks regarding responsible AI practice. Thanks Sarvjeet Singh, John Anderson, Hugo Larochelle, Blake Cunningham, Jessica Colnago for supporting publication process. We owe thanks to Muqthar Mohammad, Rama Krishna Devulapalli, Sally Goldman, Yuri Vasilevski, Alena Butryna for supporting our human study. Special thanks to Joe Ng, Zheng Xu, Yu-Siang Wang, Ravi Ganti, Arun Chaganty, Megan Leszczynski, Li Yang for exchanging research ideas and sharing engineering best practice. Thanks Li Li, Jun Wang, Jeff Wang, Bruno Costa, Mukul Gupta for sharing early feedbacks to our demo. Joonseok Lee was partially supported by NRF (2021H1D3A2A03038607, 2022R1C1C1010627, RS-2023-00222663) and IITP (2022-0-00264).

References

Abu-El-Haija, S.; Kothari, N.; Lee, J.; Natsev, P.; Toderici, G.; Varadarajan, B.; and Vijayanarasimhan, S. 2016. Youtube-8M: A large-scale video classification benchmark. *arXiv:1609.08675*.

Agostinelli, A.; Denk, T. I.; Borsos, Z.; Engel, J.; Verzetti, M.; Caillon, A.; Huang, Q.; Jansen, A.; et al. 2023. MusicLM: Generating Music From Text. *arXiv:2301.11325*.

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. ViViT: A video vision transformer. In *ICCV*.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is space-time attention all you need for video understanding? In *ICML*.

Borsos, Z.; Marinier, R.; Vincent, D.; Kharitonov, E.; Pietquin, O.; Sharifi, M.; et al. 2022. AudioLM: a language modeling approach to audio generation. *arXiv:2209.03143*.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.

Chen, K.; Zhang, C.; Fang, C.; Wang, Z.; Bui, T.; and Nevatia, R. 2018. Visually indicated sound generation by perceptually optimized classification. In *ECCV Workshops*.

Chen, L.; Srivastava, S.; Duan, Z.; and Xu, C. 2017. Deep cross-modal audio-visual generation. In *Proc. of the on Thematic Workshops of ACM Multimedia*.

Chen, P.; Zhang, Y.; Tan, M.; Xiao, H.; Huang, D.; and Gan, C. 2020. Generating Visually Aligned Sound from Videos. *IEEE Transactions on Image Processing*, 29: 8292–8302.

Choi, K.; Hawthorne, C.; Simon, I.; Dinculescu, M.; and Engel, J. 2019. Encoding musical style with transformer autoencoders. *arXiv:1912.05537*.

Chung, Y.-A.; Zhang, Y.; Han, W.; Chiu, C.-C.; Qin, J.; Pang, R.; and Wu, Y. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE.

Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2023. Simple and Controllable Music Generation. *arXiv:2306.05284*.

D'efossez, A.; et al. 2022. High Fidelity Neural Audio Compression. *arxiv:2210.13438*.

Dhariwal, P.; Jun, H.; Payne, C.; Kim, J. W.; Radford, A.; and Sutskever, I. 2020. Jukebox: A generative model for music. *arXiv:2005.00341*.

Di, S.; Jiang, Z.; Liu, S.; Wang, Z.; Zhu, L.; He, Z.; Liu, H.; and Yan, S. 2021. Video background music generation with controllable music transformer. In *ACM MM*.

Donahue, C.; et al. 2023. SingSong: Generating musical accompaniments from singing. *arXiv:2301.12662*.

Dong, H.-W.; Hsiao, W.-Y.; Yang, L.-C.; and Yang, Y.-H. 2018. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *AAAI*, volume 32.

Engel, J.; et al. 2017. Neural audio synthesis of musical notes with wavenet autoencoders. In *ICML*.

Feichtenhofer, C. 2020. X3D: Expanding architectures for efficient video recognition. In *CVPR*.

Forsgren, S.; and Martiros, H. 2022. Riffusion - Stable diffusion for real-time music generation.

Gan, C.; Huang, D.; Chen, P.; Tenenbaum, J. B.; and Torralba, A. 2020. Foley Music: Learning to Generate Music from Videos. In *ECCV*.

Garcia, H. F.; et al. 2023. VampNet: Music Generation via Masked Acoustic Token Modeling. *arxiv:2307.04686*.

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*.

Gillick, J.; Roberts, A.; Engel, J.; Eck, D.; and Bamman, D. 2019. Learning to Groove with Inverse Sequence Transformations. In *ICML*.

Hawthorne, C.; et al. 2018. Enabling factorized piano music modeling and generation with the MAESTRO dataset. *arXiv:1810.12247*.

Hershey, S.; Chaudhuri, S.; Ellis, D. P.; Gemmeke, J. F.; Jansen, A.; et al. 2017. CNN architectures for large-scale audio classification. In *ICASSP*.

Huang, C.-Z.; et al. 2018. Music transformer. *arXiv:1809.04281*.

Huang, Q.; Jansen, A.; Lee, J.; Ganti, R.; Li, J. Y.; and Ellis, D. P. 2022. MuLan: A joint embedding of music audio and natural language. In *ISMIR*.

Huang, Q.; Park, D. S.; Wang, T.; Denk, T. I.; Ly, A.; Chen, N.; et al. 2023a. Noise2Music: Text-conditioned Music Generation with Diffusion Models. *arXiv:2302.03917*.

Huang, R.; et al. 2023b. Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models. *arxiv:2301.12661*.

Huang, Y.-S.; and Yang, Y.-H. 2020. Pop Music Transformer: Beat-based modeling and generation of expressive Pop piano compositions. In *ACM MM*.

Koepke, A. S.; Wiles, O.; Moses, Y.; and Zisserman, A. 2020. Sight to Sound: An End-to-End Approach for Visual Piano Transcription. In *ICASSP*.

Kreuk, F.; Synnaeve, G.; Polyak, A.; Singer, U.; Défossez, A.; Copet, J.; Parikh, D.; Taigman, Y.; and Adi, Y. 2022. Audiogen: Textually guided audio generation. *arXiv:2209.15352*.

Kumar, R.; et al. 2023. High-Fidelity Audio Compression with Improved RVQGAN. In *NeurIPS*.

Lattner, S.; and Grachten, M. 2019. High-level control of drum track generation using learned patterns of rhythmic interaction. In *WASPAA*.

Lee, J.; Reade, W.; Sukthankar, R.; Toderici, G.; et al. 2018. The 2nd youtube-8M large-scale video understanding challenge. In *ECCV*.

Li, R.; Yang, S.; Ross, D. A.; and Kanazawa, A. 2021. AI Choreographer: Music Conditioned 3D Dance Generation with AIST++. *arXiv:2101.08779*.

Liu, H.; Chen, Z.; Yuan, Y.; Mei, X.; Liu, X.; Mandic, D.; Wang, W.; and Plumbley, M. D. 2023. AudioLDM: Text-to-audio generation with latent diffusion models. *arXiv:2301.12503*.

Mubert-Inc. 2022. MuBERT. https://mubert. com/, https://github.com/MubertAI/ Mubert-Text-to-Music.

Oord, A. v. d.; et al. 2016. Wavenet: A generative model for raw audio. *arXiv:1609.03499*.

Oore, S.; Simon, I.; Dieleman, S.; Eck, D.; and Simonyan, K. 2020. This time with feeling: Learning expressive musical performance. *Neural Computing and Applications*, 32(4): 955–967.

Owens, A.; Isola, P.; McDermott, J.; Torralba, A.; Adelson, E. H.; and Freeman, W. T. 2016. Visually indicated sounds. In *CVPR*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Roberts, A.; Engel, J.; Raffel, C.; Hawthorne, C.; and Eck, D. 2018. A hierarchical latent vector model for learning long-term structure in music. In *ICML*.

Schneider, F.; et al. 2023. Moûsai: Text-to-Music Generation with Long-Context Latent Diffusion. *arXiv:2301.11757*.

Shor, J.; Jansen, A.; Maor, R.; Lang, O.; Tuval, O.; Quitry, F. d. C.; Tagliasacchi, M.; Shavitt, I.; Emanuel, D.; and Haviv, Y. 2020. Towards learning a universal non-semantic representation of speech. *arXiv:2002.12764*.

Su, K.; Liu, X.; and Shlizerman, E. 2020a. Audeo: Audio generation for a silent performance video. *NeurIPS*, 33.

Su, K.; Liu, X.; and Shlizerman, E. 2020b. Multi-Instrumentalist Net: Unsupervised Generation of Music from Body Movements. *arXiv:2012.03478*.

Su, K.; Liu, X.; and Shlizerman, E. 2021. How Does it Sound? *NeurIPS*, 34: 29258–29273.

Su, K.; Qian, K.; Shlizerman, E.; Torralba, A.; and Gan, C. 2023. Physics-Driven Diffusion Models for Impact Sound Synthesis from Videos. In *CVPR*.

Surís, D.; Vondrick, C.; Russell, B.; and Salamon, J. 2022. It's Time for Artistic Correspondence in Music and Video. In *CVPR*.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NIPS*, 30.

Yang, D.; Yu, J.; Wang, H.; Wang, W.; Weng, C.; Zou, Y.; and Yu, D. 2022. Diffsound: Discrete diffusion model for text-to-sound generation. *arXiv:2207.09983*.

Yu, J.; Li, X.; Koh, J. Y.; Zhang, H.; Pang, R.; Qin, J.; Ku, A.; Xu, Y.; Baldridge, J.; and Wu, Y. 2021. Vector-quantized image modeling with improved VQGAN. *arXiv:2110.04627*.

Yu, J.; Wang, Y.; Chen, X.; Sun, X.; and Qiao, Y. 2023. Long-Term Rhythmic Video Soundtracker. *arXiv:2305.01319*.

Zeghidour, N.; Luebs, A.; Omran, A.; Skoglund, J.; and Tagliasacchi, M. 2021a. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 495–507.

Zeghidour, N.; Teboul, O.; Quitry, F. d. C.; and Tagliasacchi, M. 2021b. LEAF: A learnable frontend for audio classification. *arXiv:2101.08596*.

Zhou, Y.; Wang, Z.; Fang, C.; Bui, T.; and Berg, T. L. 2018. Visual to sound: Generating natural sound for videos in the wild. In *CVPR*.

Zhu, Y.; Olszewski, K.; Wu, Y.; Achlioptas, P.; Chai, M.; Yan, Y.; and Tulyakov, S. 2022a. Quantized GAN for Complex Music Generation from Dance Videos. *arXiv:2204.00604*.

Zhu, Y.; Wu, Y.; Olszewski, K.; Ren, J.; Tulyakov, S.; and Yan, Y. 2022b. Discrete Contrastive Diffusion for Cross-Modal and Conditional Generation. *arXiv:2206.07771*.