Optical Flow Matching for Video-to-Music Retrieval

Scott Merrill
Department of Computer Science
University of North Carolina
Chapel Hill, NC 27514
smerrill@unc.edu

Abstract

This paper tackles the challenging task of video-to-music retrieval (VMR), which involves automatically selecting the most suitable music for a given video clip. Extending the VM-Net framework, this work addresses two key limitations: the use of static, time-agnostic feature representations and the assumption of a single ground-truth audio track per video. The core hypothesis is that high-motion video segments should correspond to high-energy, fast-paced music. To capture this dynamic, the paper introduces a novel segmentation method based on optical flow to detect "regime changes" in video motion. A new loss function is proposed to align video and audio segments according to optical flow-derived motion rankings. Temporal dependencies in both modalities are modeled using transformer encoders, replacing VM-Net's simpler fully connected architecture. Evaluations on the HIMV-50K and SymMV datasets show that while the proposed model (OF-VM-Net) lags behind VM-Net on standard retrieval metrics like Recall@k and FAD, it consistently outperforms on the AV-ALIGN metric—highlighting improved temporal synchronization. Notably, OF-VM-Net demonstrates greater robustness to distribution shifts, showing less performance drop on the out-of-distribution SymMV dataset. All code for this project is publicly available at: https://github.com/smerrillunc/VMR.

1 Introduction

Music plays a pivotal role in enhancing video content, providing a deeper sense of immersion and engagement. However, selecting an appropriate track is a complex challenge, as it requires aligning both the emotional tone and rhythmic structure of the music with those in the video. For a system to perform well at this task, it must possess an understanding of both video and music semantics, and, more crucially, how they relate over time.

In this work, we tackle the challenging problem of video-to-music retrieval (VMR), with the goal of automatically identifying the most fitting and engaging piece of music to complement a given video clip. Our approach builds upon the VM-Net framework proposed by Hong et al (1). While the simplicity of VM-Net is appealing, it makes unrealistic assumptions that oversimplify the relationship between video and music. For example, VM-Net models video and music features as timeless vectors, neglecting the inherent temporal structure of both modalities. Additionally, it assumes a one-to-one mapping between each video and a single ground truth audio track, overlooking the natural redundancy present within and across musical compositions. A song often consists of repetitive components, such as choruses, and much of Western music follows similar harmonic progressions and melodies, creating further redundancy between tracks. By prescribing a single ground truth for each video-music pair, VM-Net disregards these redundancies.

Our approach seeks to address these limitations. To exploit the structural similarities within songs, we focus on matching video and audio segments, rather than whole clips. Furthermore, instead of

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

assuming a single ground truth, we hypothesize that fast-paced video scenes should be paired with relatively faster-paced sections of the audio track. With this assumption in mind, we aim to align optical flow rankings between video and audio segments. Lastly, we employ transformer encoders to model the temporal dependencies in both video and audio, ensuring a more dynamic and accurate representation of the relationships between the two.

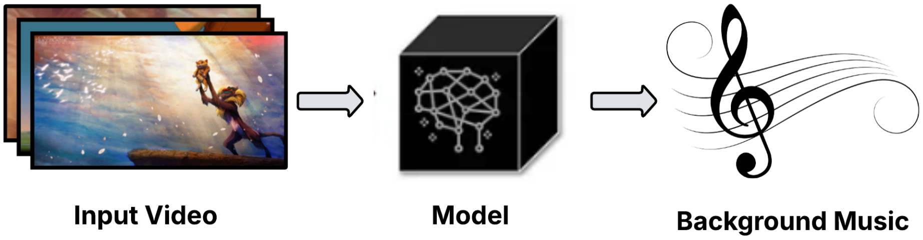

Figure 1: **Video to Music Retrieval:** Given, an input video, the goal of VMR is to identify the corresponding background music that corresponds to that clip.

2 Related Work

2.1 Optical Flow

Optical flow-based features have gained significant attention in computer vision due to their effectiveness in motion detection and gesture recognition tasks. By analyzing pixel-wise motion between consecutive frames, optical flow can capture the dynamic behavior of moving objects. These features have proven valuable in a variety of motion-centric applications, including human activity recognition (2; 3), human detection (4), hand gesture recognition (5; 6), and even dance movement recognition (7). We believe incorporating motion modeling into the VMR task can enhance temporal synchronization between visual and auditory modalities by enabling the model to more effectively align high-tempo visual sequences with their corresponding fast-paced musical motifs.

2.2 Video-to-Music Retrieval

VMR has been approached through both supervised and self-supervised learning paradigms, each offering distinct advantages and limitations. Supervised methods typically rely on annotated data—such as mood tags, emotional labels, or semantic descriptors—to establish explicit correspondences between video and audio content. For example, Shah et al. (8) leverage heuristic rankings and fusion models to generate personalized soundtracks, while Hsia et al. (9) train a convolutional neural network (CNN) to associate images with lyrics-derived keywords. Similarly, (10) proposes learning a latent emotional space using emotional tags to drive VMR through an architecture akin to VM-Net. While these supervised methods can enhance training efficiency and semantic alignment, they are often limited by the scope of the available annotations and the additional effort required for data labeling.

In contrast, self-supervised approaches aim to learn audiovisual correspondences from large-scale unlabeled data, circumventing the need for manual annotations. A prominent example is VM-Net, proposed by Hong et al. (1), which utilizes an extended triplet loss and pre-trained feature extractors to align video and music content efficiently. Seg-VM-Net (11) extended the architecture by incorporating semantic segmentation of both modalities to better capture structural content. This segmentation-based strategy proved critical for improving retrieval performance. Additional efforts have focused on refining loss functions to enhance robustness; for instance, (12) introduce a loss function designed to account for the many-to-one relationship between video and suitable musical accompaniments, reducing the impact of false negatives. Inspired by this, our work integrates optical flow information to further regularize such noise.

2.3 Video-to-Music Generation

A closely related task to VMR is Video-to-Music Generation. Rather than identifying the most suitable background music, these methods aim to generate musical pieces that complement the video. While similar tasks, the approaches to solving them are often quite different. Many video-to-music generation tasks rely on autoregressive music generation models conditioned on extracted video features (13; 14; 15). Lin et al. (14) further utilize optical flow to detect scene changes and generate musical beats corresponding to these transitions. We adopt a similar strategy for the VMR task.

3 VM-Net

The VM-Net, introduced by Hong et al. (1), is a two-branch neural network designed to map pre-processed video and music features into a shared embedding space. Music is encoded as a 1,140-dimensional vector, derived from handcrafted audio features such as spectral centroid and chroma features, with statistical aggregation. Video features are extracted from using pre-trained ImageNet models and represented as a 1,024-dimensional vector. These feature vectors are then processed through a series of fully connected layers and projected into a common embedding space, where the final music and video representations, denoted as e_m and e_v, are 512-dimensional vectors.

To train these networks, VM-Net employs a self-supervised triplet loss that encourages related music and video content to be projected closer together in the embedding space. Specifically, the triplet loss function ensures that an anchor sample (e.g., a video clip) is closer to its corresponding positive sample (e.g., its associated music), while simultaneously pushing it away from unrelated negative samples. Formally, the loss function is defined as:

$$L(a,p,n) = \max\left(||f(a) - f(p)||^2 - ||f(a) - f(n)||^2 + \alpha, 0\right) \tag{1}$$

where α is a margin parameter that controls the separation between negative pairs. Unlike traditional triplet loss approaches, VM-Net optimizes two separate embedding functions, f_m for music and f_v for video. The model is trained bidirectionally, enabling both video and music embeddings to act as the anchor during training, thus improving cross-modal alignment.

4 Methods

VM-Net introduced an innovative, lightweight network capable of performing VMR tasks. However, the original model has limitations, as it assumes video and audio features timeless vectors and assigns a single ground truth to each video-audio pair. To address these issues, we propose segmenting video and audio clips, using transformers to model their temporal relationships, and updating the loss function to better encourage alignment between video and audio. Our full model is shown in Figure 2

4.1 Video Segmentation

Instead of processing entire music and video clips, we segment videos into distinct segments. Following the approach of Seg-VM-Net, we divide a video and audio pair into N distinct time intervals (formally defined as $\{t_1, t_2, ..., t_N\}$, where each t_i represents a segment). Our choice of segmentation strategy differs from any of the choices proposed by Seg-VM-Net. While Seg-VM-Net suggests using audio features (e.g., beat and tempo) to determine segment boundaries, this is impractical in the VMR task where audio is unavailable at inference. Thus, we propose using the video optical flow to construct these boundaries.

Our optical flow segmentation strategy attempts to identify *regime changes* defined as intervals in which the optical flow exhibits abnormally high values followed by abnormally low values. We leverage common techniques in signal processing to detect change points (16). Our segments are created by minimizing a cost function designed to identify changes in both the mean and the covariance of the opical flow. Specifically, given a signal y_t defined over an interval I, the cost function is defined as:

$$c(y_I) = |I| logdet(\hat{\Sigma}_I + \epsilon Id)$$

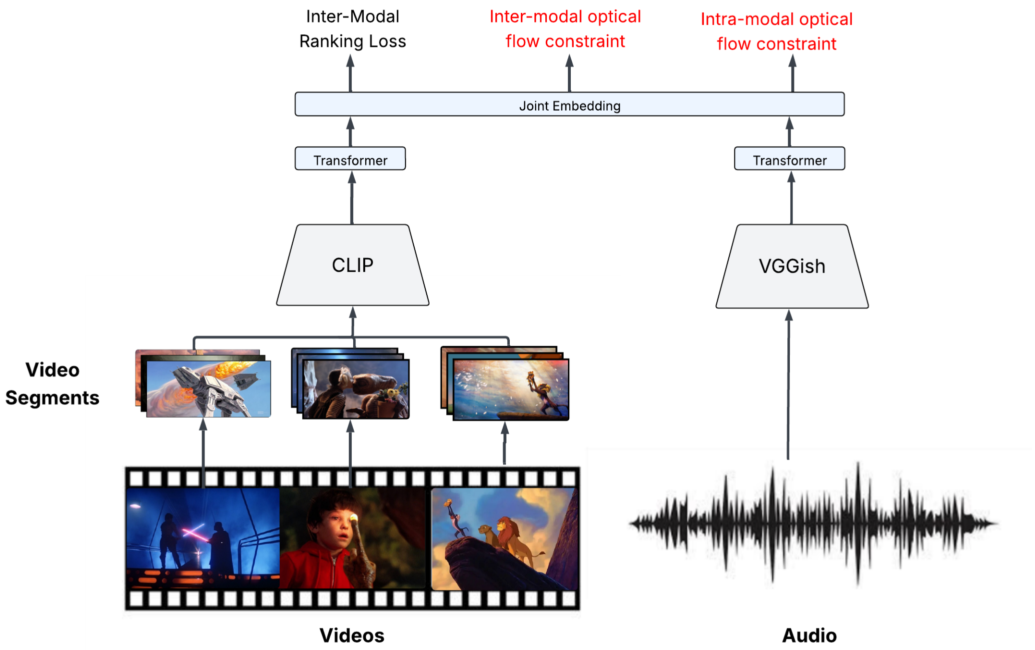

Figure 2: **Our Proposed OF-VM-Net:** An overview of our optical-flow-based video-to-music retrieval framework. Given a video input, we first apply an optical flow-based segmentation strategy to divide it into semantically meaningful segments. These segments are then encoded using a transformer to capture temporal dependencies. Parallel audio segments are processed similarly. The joint embedding space is optimized using a novel loss function that aligns video and audio segments both semantically and by motion intensity (optical flow), improving retrieval performance.

where $\hat{\Sigma}_I$ is the empirical covariance matrix of the sub-signal $\{y_t\}_{t \in I}$ and $\epsilon > 0$ is added for numerical stability. In contrast to other cost functions that detect shifts in the median or mean of a signal, our cost function also captures the variance of the optical flow, providing additional context on how the flow is changing in a given interval. We use a dynamic programming approach to minimize this cost function exactly with the only requirement that each segment have at least 10 frames (10 seconds of video).

4.2 Transformer-based Modeling

Instead of relying on fully connected layers to project audio and video features, we adopt transformer encoders. Transformers are better suited for modeling temporal dependencies, as they can attend to past video and audio frames, unlike fully connected layers that ignore such temporal relations. We hypothesize that leveraging transformers to model these dependencies will result in more meaningful embeddings.

4.3 Optical Flow Ranking Loss

In VM-NET, the bi-directional inter-modal triplet loss function attempts to match positive and negative video-audio pairs. This approach is naive as it assumes a single ground truth audio for each video. Given that that the optimal background music for a particular video is subjective, this is an aggressive assumption. Their approach is thus heavily dependent on access to a large and well-aligned dataset. To address this issue, we propose an alternative assumption that "engaging" video scenes should align with "engaging" music segments. To encourage this behavior, we incorporate an optimal flow term into the loss function. Our updated loss not only seeks to align video and audio pairs, but also synchronizes fast-paced video and music segments.

To encourage alignment based on motion dynamics, we divide each video into S distinct, non-overlapping clips as previously described. For both video and audio streams, we compute the mean optical flow for each segment, which serves as a proxy for motion intensity. These segments are then ranked within their modality: the segment with the highest mean optical flow receives a rank of 1, and the lowest is assigned rank S. We then define a ranking-based triplet loss to align segments of similar motion characteristics. Specifically, we mine triplets where the anchor is the video segment with the highest optical flow (rank 1), the positive example is any audio segment that also ranks 1, and the negative is an audio segment with the lowest optical flow (rank S). By applying this ranking-based loss symmetrically in both directions—video-to-audio and audio-to-video—we encourage high-motion video scenes to be paired with equally dynamic music segments. Figure 3 provides a visual overview of this proposed optical flow ranking mechanism.

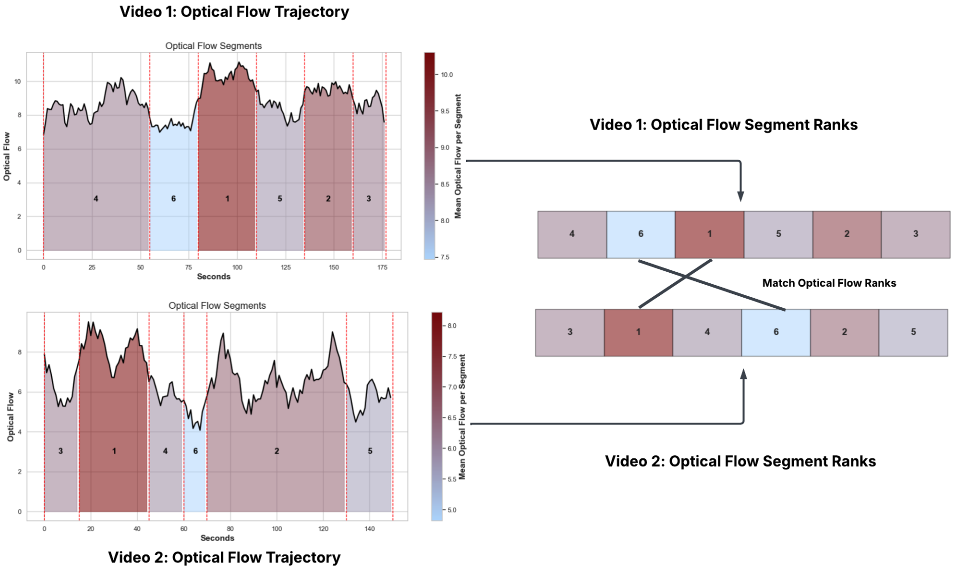

Figure 3: **Optical Flow Ranking Loss:** Segments are ranked by motion intensity using optical flow. Within each segment we calculte the mean optical flow to compute a ranking vector. Triplets are then mined such that high-motion video clips are paired with high-energy music segments (rank 1), while low-motion clips are discouraged from matching with such segments.

4.4 Full Loss Function

To formally define our loss function, first consider a mini-batch comprising N music videos, each of which is further divided into S distinct segments or clips. This setup allows us to generate a total of $N \times S$ pairs of embedded features, represented mathematically as (v_{ij}, m_{ij}). Here, v_{ij} denotes the video features extracted from the j-th segment of the i-th video, while m_{ij} represents the corresponding music features for that same segment.

To effectively align the ground truth video with its corresponding audio, we construct two sets of triplets: the first set consists of (v_{ij}, m_{ij}, m_{kl}), and the second set includes (m_{ij}, v_{ij}, v_{kl}). In these formulations, we ensure that at least one of the indices *k* or *l* is different from *i* or *j*, respectively. In simpler terms, this means that a positive pair consists of a specific video segment and its matched audio segment, while a negative pair can be any other segment, potentially sourced from the same video or entirely different videos.

In addition to matching ground of truth labels, we also focus on matching the optical flow rankings. Let $v_{i,high}$ and $v_{i,low}$ denote the highest and lowest optical flow segment from video *i* respectively. To match optical, optical flow rankings, we mine an additional four triplets: $(v_{i,high}, m_{j,high}, m_{j,low})$, $(m_{i,high}, v_{j,high}, v_{j,low})$, $(v_{i,high}, v_{j,high}, v_{k,low})$ and $(m_{i,high}, m_{j,high}, m_{k,low})$. The first two operate inter-modally, aligning rankings between vision and audio, while the latter two are intra-modal triplets.

Collecting all of these triplets, we can write our loss function as:

$$
\begin{aligned}
\mathcal{L} = \lambda_1 \sum \max\big(0,\ v_{ij}^{\top} m_{kl} - v_{ij}^{\top} m_{lj}\big) + \\
\lambda_2 \sum \max\big(0,\ m_{ij}^{\top} v_{kl} - m_{ij}^{\top} v_{lj}\big) + \\
\lambda_3 \sum \max\big(0,\ v_{i,high}^{\top} m_{j,low} - v_{i,high}^{\top} m_{j,high}\big) + \\
\lambda_4 \sum \max\big(0,\ m_{i,high}^{\top} v_{j,low} - m_{i,high}^{\top} v_{j,high}\big) + \\
\lambda_5 \sum \max\big(0,\ v_{i,high}^{\top} v_{k,low} - v_{i,high}^{\top} v_{j,high}\big) + \\
\lambda_6 \sum \max\big(0,\ m_{i,high}^{\top} m_{k,low} - m_{i,high}^{\top} m_{j,high}\big)
\end{aligned}
$$

where λ_1 and λ_2 balance the impact of aligning music and video with the ground truth. λ_3 - λ_6 weigh the influence of aligning the optical flow. It is also important to note that because the features of video and music are L2 normalized, the dot product serves as a valid metric for similarity, while the negative dot product is an appropriate metric for distance.

5 Dataset

To train and evaluate our model, we require datasets consisting of videos with background music. To acquire these pairings, we leverage publicly available datasets based on YouTube videos. This approach allows us to create a unified pipeline that processes all datasets in a consistent manner. Specifically, we rely on the following two datasets:

- **HIMV-50K** (17): A subset of the YouTube-8M dataset (17), HIMV-50K contains 50,000 training videos and 1,000 validation videos, all of which are music videos. While this dataset is large and diverse, the video quality within the audio-visual clips varies significantly. Due to this noise, it'd be inappropriate to evaluate our model on solely on this dataset. However, due to it's scale it's well suited for unsupervised training.

- **SymMV Dataset** (18): This dataset is much smaller scale, consisting of only 1,141 manually curated music video clips. While small in size, the these clips are of professional quality making them particularly suitable for evaluation.

After downloading the videos from these datasets, we extract image-based features using ResNet (19), CLIP (20), and I3D (21). ResNet was trained on large image classification datasets and is particularly good at capturing high-level spatial features, making it effective for object recognition and general feature extraction from video frames. It generates a 2048-dimensional feature vector representing rich spatial content. CLIP, trained on a large dataset of 400 million image-text pairs, excels in aligning visual and textual data in a shared embedding space. This enables CLIP to understand the semantic relationship between images and textual descriptions, making it well-suited for tasks that require contextual interpretation of visual data. The model produces a 512-dimensional feature vector, capturing a more semantic representation of the visual content in relation to text. I3D, trained on action recognition datasets like Kinetics-400, specializes in capturing both spatial and temporal information from video clips. By inflating traditional 2D convolutions into 3D, I3D models temporal dynamics in videos, making it particularly effective for understanding motion and actions. It generates two separate 1024-dimensional feature vectors, one for RGB frames and one for optical flow, which are concatenated into a single 2048-dimensional vector. This feature extraction approach makes I3D ideal for video-based tasks requiring motion and action recognition. For audio processing, we use VGGish features (22), which convert each second of audio into a 128-dimensional vector.

6 Experiments

6.1 Implementation Details

We divided each video into 10 non-overlapping segments, ensuring that each segment contains between 10 and 200 frames. Segments exceeding 200 frames were discarded, while shorter ones were zero-padded to exactly 200 frames. As a result, both the video and audio transformers operate on fixed-length sequences of 200 tokens. These transformers embed the video and audio features into a shared latent space of dimension 32.

To reduce computational overhead during training, we avoid calculating the loss over all possible triplets. Instead, we identify and use the top 10 most violated triplets within each mini-batch. A violated triplet is defined as one where the similarity between mismatched embeddings v_{ij} and m_{kl} is high despite despite $i \neq k$ or $j \neq l$. We apply the same strategy for selecting the top 10 most violated optical flow triplets. Training was conducted with a batch size of 32 over 100 epochs.

6.2 Evaluation Metrics

Evaluating our model poses certain challenges, particularly because it is not solely trained to match the ground-truth. Consequently, we do not expect our model to achieve high scores on traditional retrieval metrics such as Recall@k. While we believe that human evaluation is the gold standard for assessing the quality and relevance of music-video retrieval, the scale and cost of such evaluations make them impractical within the constraints of a single semester. As an alternative, we rely on a combination of established quantitative metrics from the literature to evaluate our model's performance.

One key metric is the AV-Align score (23), which measures synchronization by comparing the peaks in the mean magnitude of optical flow across video frames with the corresponding peaks in the audio waveform amplitude. A higher AV-Align score indicates stronger temporal alignment between the video and audio signals. Another metric we use is the Fréchet Audio Distance (FAD) (24), which quantifies the similarity between the retrieved and real music samples based on their distributions in VGGish (22) space. While we do not expect our retrieved audio to closely match the ground-of-truth tracks, there still might be some alignment in the deep feature space.

Together, these metrics provide a multifaceted evaluation framework that captures both the semantic alignment and perceptual quality of retrieved music, offering a feasible and informative alternative to human judgment.

6.3 Results

We evaluated our proposed model, OF-VM-Net, against the original VM-Net using a subset of 20,000 samples from the HIMV-50k dataset, reserving 1,600 samples for validation. After training, we also assessed both models on the 1,141 samples of the SymMV dataset.

Our quantitative results are summarized in Table 1, with the best performance in each row boldfaced. VM-Net significantly outperforms our model on traditional retrieval metrics such as recall and FAD. This outcome is expected, as our model is not optimized solely for retrieval accuracy. Instead, by focusing on improving optical flow-based alignment between video and audio, OF-VM-Net achieves consistently higher AV-ALIGN scores. Notably, our model not only surpasses VM-Net in AV-ALIGN but, in some cases, even outperforms the alignment observed in the ground-truth pairs—suggesting that the labeled ground truth may not always represent the most tightly aligned audio-visual examples.

Note that the results reported in Table 1 differ to those presented in class. After some debugging, we found that by modulating the weights, λ_1 and λ_2, we can greatly improve performance on retrieval metrics with a consequentially hit on alignment scores. This makes sense as assigning higher values to these parameters gives more weight to matching the ground of truth and less weight to optical flow rank matching. In future work, we plan to tune these parameters and try to strike the optimal balance between optimizing for retrieval and optimizing for audio-video alignment.

In Table 2 we show the results on the SymMV dataset. VM-Net suffers a substantial performance degradation with recall scores dropping sharply and FAD increasing nearly fourfold. This suggests that VM-Net may be overfitting to the HIMV-50k distribution. In contrast, OF-VM-Net demonstrates greater robustness, and is able to outperform VM-Net on both retrieval and alignment metrics on this alternative data distribution. This result highlights the potential of training for alignment (rather than retrieval alone) to improve generalization across diverse audio-visual domains.

Finally, our experiments reveal that the choice of feature extractor has a significant impact on performance. For VM-Net, CLIP and ResNet features yield the best results on the HIMV-50k dataset. However, when evaluated on out-of-distribution data, the influence of the feature extractor becomes less pronounced. Notably, VM-Net performs poorly with I3D features, which are designed to capture motion. This suggests that VM-Net does not effectively utilize motion information, and incorporating such features may serve to confuse the model. In contrast, OF-VM-Net performs best when using I3D features, which is consistent with our goal of explicitly modeling motion. This further suggests that motion-based features complement OF-VM-Net and enhance the model's ability to learn more accurate audio-video alignments.

Model	**Vision Features**	**AV-ALIGN (GOT) ↑**	**AV-ALIGN ↑**	**FAD ↓**	**R@1 ↑**	**R@5 ↑**	**R@10 ↑**
VM-Net	Resnet	0.0621	0.0621	0.5897	0.8979	0.8998	0.9023
VM-Net	CLIP	0.0621	0.0600	**0.5609**	0.8979	**0.9041**	**0.9085**
VM-Net	I3D	0.0621	0.0621	0.5515	**0.8274**	0.8318	0.8350
OF-VM-Net	Resnet	0.0621	0.0675	1.1378	0.2551	0.3427	0.4223
OF-VM-Net	CLIP	0.0621	0.0700	1.0409	0.2624	0.3566	0.4511
OF-VM-Net	I3D	0.0621	**0.0723**	1.0224	0.2884	0.4269	0.5031

Table 1: Performance comparison on the HIMV-50k validation set. OF-VM-Net demonstrates superior alignment capabilities (AV-ALIGN), while VM-Net achieves higher scores on conventional retrieval metrics.

Model	**Vision Features**	**AV-ALIGN (GOT) ↑**	**AV-ALIGN ↑**	**FAD ↓**	**R@1 ↑**	**R@5 ↑**	**R@10 ↑**
VM-Net	Resnet	0.0740	0.0655	2.0092	0.2200	0.2342	0.2414
VM-Net	CLIP	0.0740	0.0667	1.8925	0.2295	0.2556	0.2759
VM-Net	I3D	0.0740	0.0700	1.9267	0.2497	0.2640	0.2759
OF-VM-Net	Resnet	0.0740	0.0691	1.5543	0.2318	0.3040	0.3526
OF-VM-Net	CLIP	0.0740	0.0699	1.4600	0.2492	0.3267	0.3591
OF-VM-Net	I3D	0.0740	**0.0738**	**1.3550**	**0.2831**	**0.3640**	**0.3701**

Table 2: Performance on the out-of-distribution SymMV dataset. VM-Net exhibits significant degradation across all metrics, particularly in Recall and FAD, suggesting sensitivity to distributional shifts. In contrast, OF-VM-Net maintains relatively stable, indicating enhanced generalization and robustness to domain variation.

7 Conclusion

While OF-VM-Net was unable to outperform VM-Net on the in-distribution dataset with respect to traditional retrieval metrics, this outcome is not unexpected given that OF-VM-Net was not trained solely for retrieval accuracy. It was trained with an additional element designed improve audio-video alignment. In this respect, OF-VM-Net excels. As measured by the AV-ALIGN score, it not only outperforms VM-Net but, in some cases, even surpasses the ground-truth alignment. This suggests that OF-VM-Net is capable of identifying music tracks that, according to the alignment metric, better match the visual content than the originally paired audio.

Excitingly, we also observe that optimizing for alignment leads to models that are more robust to shifts in data distribution. When evaluated on a different dataset (SymMV) from the one it was trained on (HIMV-50k), VM-Net's performance degrades substantially—Recall drops by up to 75%, and FAD increases by a factor of four. In contrast, OF-VM-Net maintains relatively and is able to outperform VM-Net on this data distribution.

Our work begs the question: *What music should a VMR model retrieve?* Is it sufficient to mimic the training distribution, or should retrieval prioritize musical selections with higher semantic alignment to the video content? Given the subjective nature of VMR assigning a single ground of truth seems naive. In contrast, alignment serves as a practical and quantifiable metric for evaluating the semantic fit between music and video. Models trained to optimize alignment not only generalize better across varied inputs but also support richer, more contextually appropriate and diverse retrievals—which are key to real creative applications to VMR.

Looking ahead, several promising directions emerge for future work. One critical avenue is the development of richer alignment metrics that capture deeper semantic, temporal, and emotional correspondences between music and video. Improved quantitative metrics could provide more reliable supervision signals and reduce dependence on noisy or inconsistent training labels. Another important step is incorporating human evaluations. While we show that OF-VM-Net improves AV-ALIGN scores, it remains to be seen whether these improvements correlate with human judgments of musical fit. Establishing this link would provide strong validation for alignment-based optimization.

References

[1] S. Hong, W. Im, and H. S. Yang, "Content-based video-music retrieval using soft intra-modal structure constraint," 2017.

[2] A. Ladjailia, I. Bouchrika, H. F. Merouani, N. Harrati, and Z. Mahfouf, "Human activity recognition via optical flow: decomposing activities into basic actions," *Neural Computing and Applications*, vol. 32, pp. 16387 – 16400, 2019.

[3] S. S. Kumar and M. John, "Human activity recognition using optical flow based feature set," *International Carnahan Conference on Security Technology*, pp. 1 – 5, 2016.

[4] S. Hoshino and K. Niimura, "Optical flow for real-time human detection and action recognition based on cnn classifiers," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 23, pp. 735 – 742, 2019.

[5] D. Sarma, V. Kavyasree, and M. K. Bhuyan, "Two-stream fusion model for dynamic hand gesture recognition using 3d-cnn and 2d-cnn optical flow guided motion template," *arXiv: Computer Vision and Pattern Recognition*, 2020.

[6] M. U. Rehman, T. Ilyas, L. Seneviratne, and I. Hussain, "Enhanced gesture recognition through graph-based multimodal fusion," pp. 1 – 5, 2024.

[7] Z. jie Lin, "Dance movement recognition method based on convolutional neural network," pp. 255 – 258, 2023.

[8] R. R. Shah, Y. Yu, and R. Zimmermann, "Advisor: Personalized video soundtrack recommendation by late fusion with heuristic rankings," in *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, (New York, NY, USA), p. 607–616, Association for Computing Machinery, 2014.

[9] C. Hsia, K. Lai, Y. Chen, C. Wang, and M. Tsai, "Representation learning for image-based music recommendation," *CoRR*, vol. abs/1808.09198, 2018.

[10] B. Li and A. Kumar, "Query by video: Cross-modal music retrieval," in *International Society for Music Information Retrieval Conference*, 2019.

[11] L. Prétet, G. Richard, C. Souchier, and G. Peeters, "Video-to-music recommendation using temporal alignment of segments," *IEEE Transactions on Multimedia*, vol. 25, p. 2898–2911, 2023.

[12] Z. Chen, P. Zhang, K. Ye, W. Dong, X. Feng, and Y. Zhang, "Start from video-music retrieval: An inter-intra modal loss for cross modal retrieval," 2024.

[13] S. Li, B. Yang, C. Yin, C. Sun, Y. Zhang, W. Dong, and C. Li, "Vidmusician: Video-to-music generation with semantic-rhythmic alignment via hierarchical visual features," 2024.

[14] Y.-B. Lin, Y. Tian, L. Yang, G. Bertasius, and H. Wang, "Vmas: Video-to-music generation via semantic alignment in web music videos," *arXiv preprint arXiv:2409.07450*, 2024.

[15] Z. Tian, Z. Liu, R. Yuan, J. Pan, Q. Liu, X. Tan, Q. Chen, W. Xue, and Y. Guo, "Vidmuse: A simple video-to-music generation framework with long-short-term modeling," 2024.

[16] M. Lavielle and G. Teyssière, "Detection of multiple change-points in multivariate time series," *Lithuanian Mathematical Journal*, vol. 46, no. 3, pp. 287–306, 2006.

[17] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," 2016.

[18] L. Zhuo, Z. Wang, B. Wang, Y. Liao, C. Bao, S. Peng, S. Han, A. Zhang, F. Fang, and S. Liu, "Video background music generation: Dataset, method and evaluation," 2023.

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2015.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” 2021.

[21] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” 2018.

[22] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, “Cnn architectures for large-scale audio classification,” 2017.

[23] G. Yariv, I. Gat, S. Benaim, L. Wolf, I. Schwartz, and Y. Adi, “Diverse and aligned audio-to-video generation via text-to-video model adaptation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 6639–6647, 2024.

[24] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, “Fréchet audio distance: A metric for evaluating music enhancement algorithms,” 2019.