

# KARMA-MV: A BENCHMARK FOR CAUSAL QUESTION ANSWERING ON MUSIC VIDEOS

Archishman Ghosh   Abhinaba Roy   Dorien Herremans

AMAAI Lab, Singapore University of Technology and Design

archishman\_ghosh@mymail.sutd.edu.sg, {abhinaba\_roy, dorien\_herremans}@sutd.edu.sg

## ABSTRACT

While significant progress has been made in Video Question Answering and cross-modal understanding, causal reasoning about how visual dynamics drive musical structure in music videos remains under-explored. We introduce KARMA-MV, a large-scale multiple-choice QA dataset derived from 2,682 YouTube music videos, designed to test models’ ability to integrate temporal audio-visual cues and reason about visual-to-musical influence across reasoning, prediction, and counterfactual questions. Unlike traditional datasets requiring manual annotation, KARMA-MV leverages LLM reasoning for scalable generation and validation, yielding 37,737 MCQs.

We propose a causal knowledge graph (CKG) approach that augments vision-language models (VLMs) with structured retrieval of cross-modal dependencies. Experiments on state-of-the-art VLMs and LLMs show consistent gains from CKG grounding—especially for smaller models—establishing the value of explicit causal structure for music-video reasoning. KARMA-MV provides a new benchmark for advancing causal audio-visual understanding beyond correlation.

## 1. INTRODUCTION

Audio and video are two dominant perceptual channels through which human experience the world, and understanding how these modalities co-evolve is fundamental to a broad class of multimedia reasoning problems. When watching a music video by famous rock bands or a choreographed dance sequence from a film, it is natural to observe tight coordination between what unfolds visually and how the music responds. Scene transitions align with chord progression, intensity of action correlates with a change in loudness, etc. Investigating the precise nature of these causal dependencies, specifically how changes in visual features drive changes in musical features, is both scientifically interesting and practically important for building systems that truly understand audio-visual content. To that end, we present a structural framework for characterizing the causal relationship between video and music. Since we are primarily concerned with music videos, we focus on the musical features that change in response to shifts in video features. Throughout this work, “causal” refers to the directional visual-to-audio framing, i.e., how visual changes systematically precede and predict musical changes, rather than interventional causality in the formal

sense. In this paper, we introduce **KARMA-MV**, a Causal Music-Video Question Answering MCQ dataset based on the scenes from music videos, targeting the causal effect of visual changes on musical and audio features.

The main focus of KARMA-MV is on causal influence of video on music, *i.e.*, how musical properties change when video features change. To construct the dataset, we extract audio and visual features independently and feed them into the state-of-the-art Large Language Model Qwen2.5-7B-Instruct [1,2] to generate semantically grounded question-answer pairs. The questions are designed to be interpretable by domain experts and to require genuine cross-modal reasoning rather than unimodal shortcuts.

Following Li et al. [3], KARMA-MV evaluates whether models can understand what is there in the video (description), explain the intentions of actions or procedures to certain targets (explanation), predict what may happen in the future (prediction), and imagine the scenarios in different conditions (counterfactual). However, unlike Li et al. [3], we focus solely on the visual content; our benchmark is explicitly grounded in how visual changes causally influence musical properties. We validate our work against two Vision Language Models capable of jointly processing text, video and audio.

**Contributions.** The primary contributions of this work are as follows:

1. We introduce a large-scale, automatically generated MCQ dataset for causal reasoning in music videos, covering description, explanation, prediction, and counterfactual question types across diverse audio-visual scenes.
2. We propose an LLM-driven pipeline for dataset construction that eliminates the need for manual annotation while preserving semantic complexity, enabling straightforward extension to new video sources.
3. We design an architecture that integrates a Causal Knowledge Graph with a Vision Language Model, enabling structured cross-modal reasoning over audio-visual dependencies.
4. We provide a comprehensive evaluation of state-of-the-art VLM baselines on KARMA-MV and demonstrate that explicit causal modeling via knowledge graphs yields consistent improvements, establishing a strong baseline for future work on causal music-video understanding.

In the remainder of this paper, we describe how a Causal

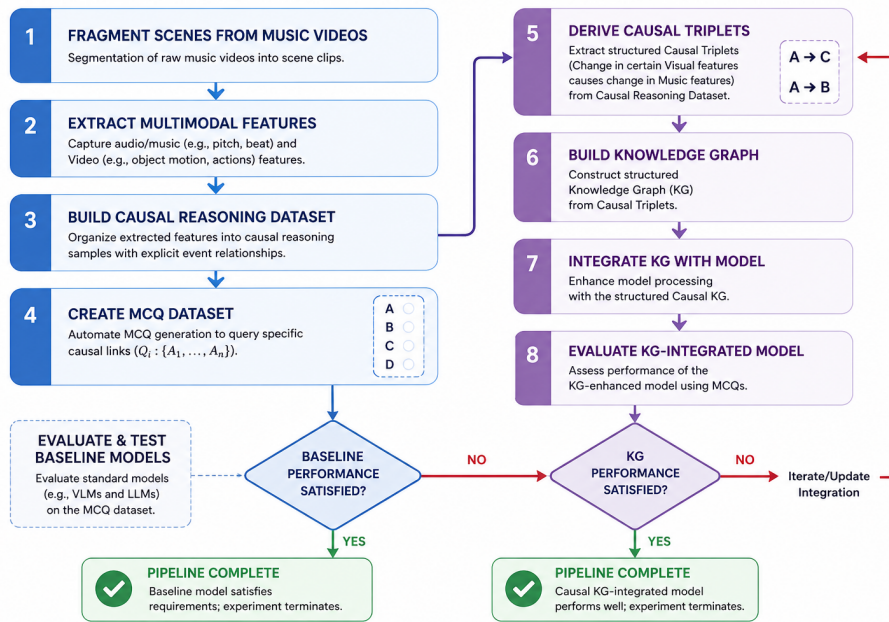


Figure 1: Overview of the proposed KARMA-MV approach.

Knowledge Graph based VLM architecture can be constructed for Music-Video Question Answering task. We examine in detail how the causal knowledge helps the Visual language Model or reasoning model to improve its efficiency in question answering task related to music videos.

## 2. RELATED WORK

### 2.1 Multimodal and Audio-Visual Understanding

KARMA-MV operates across three modalities, hence is at the intersection of multimodal representation learning and audio-visual understanding. Sun et al. in [4] proposed Interaction Canonical Correlation Network (ICCN), for multimodal sentiment analysis and emotion recognition by jointly modelling text, audio, and video features.

### 2.2 Video Question Answering and Causal Reasoning

A broad range of video QA datasets has been developed for video understanding [5], spanning story comprehension [6], TV show understanding [7], and general web videos [8]. Among these, the datasets most relevant to our work are those that target causal or temporal reasoning rather than simple factual retrieval.

NExT-QA [9] provides MCQs for causal and temporal action reasoning in video, establishing an important precedent for structuring causal inference as a multiple-choice task. Most directly related is Causal-VidQA [3], which introduced a large-scale benchmark of 107,600 QA pairs designed to advance video understanding into causal and commonsense reasoning. The dataset defines four question types—description, explanation, prediction, and counterfactual—which we adopt in KARMA-MV. However, Causal-VidQA focuses entirely on visual content, whereas KARMA-MV is specifically concerned with how visual changes causally drive changes in music and audio. The Audio-Visual Question Answering dataset

(AVQA) [10] addresses questions about sounds, visual objects such as instruments, and their associations in video, making it the closest prior work in the audio-visual QA space to our own. Our benchmark extends this direction by explicitly targeting the causal direction from video to music, introducing counterfactual and prediction question types absent from AVQA.

## 3. METHODOLOGY

The work is broadly divided into two parts: **I. Dataset Creation and testing on baseline VLMs.** **II. Building Knowledge Graph VLM architecture.** The dataset creation pipeline comprises four steps: 1) Feature extraction, 2) Generation of causal reasoning dataset 3) Generation of MCQ dataset 4) Validating the dataset. Our approach is very simple. Unlike many other datasets, KARMA-MV does not rely on human annotators for either creation or validation. Instead, we leverage an LLM for automated dataset creation and VLMs for baseline evaluation, which ensures scalability without sacrificing semantic richness. After validation, we proceed to build a Causal Knowledge Graph augmented VLM architecture to answer the MCQ questions in the KARMA-MV dataset. For the dataset construction, we assemble a corpus of music videos available on YouTube under Creative Commons licenses, totalling of 2,682 music-videos. These are segmented into clips using PySceneDetect [11] for scene transition analysis in the following steps.

### 3.1 Feature Extraction

Two broad categories of features are extracted: 1) **Music and Audio Features** and 2) **Visual Features**. After segmenting music videos, we filter the resulting clips for quality and relevance, yielding a set of 22,011 music-video scenes. For each scene, we extract a comprehensive set

of features including loudness, tempo, key, chord, mood, genre, instruments used, and voice detection using multiple specialized libraries. Similarly, visual features such as object count and visual intensity are extracted using separate libraries.

**Music and Audio Features.** For loudness extraction we use Librosa [12], a standard python library for analyzing audio and music signals. For mood extraction we use Music2Emo [13]. For tempo, genre, chord, instrument detection and voice detection, we follow [14]. For key, genre and instrument extraction we rely on the Essentia Library [15].

**Visual Features.** To quantify the visual characteristics of the video stimuli, we extract four key metrics: Motion, Brightness, Contrast, and Saturation, by sampling every fifth frame to balance granularity with computational efficiency. Motion is calculated via frame-differencing after applying a Gaussian blur to reduce noise, while the remaining aesthetic features are derived from the HSV (Hue, Saturation, Value) color space. These metrics are then aggregated into a Total Visual Intensity (TVI) score using a weighted linear combination:  $TVI = 0.5M + 0.3C + 0.2S$ . This composite metric represents the overall dynamic energy of each scene, providing a standardized basis for subsequent statistical analysis. Beyond visual characteristics, we use YOLOv8-Medium (YOLOv8m) pre-trained model [16] for number of types of objects. For each video segment, inference is performed on the middle frame to obtain a representative snapshot of the scene’s content. After feature extraction, we merge all the audio/music features and video features into per-video feature files, where each row corresponds to one scene. These raw features serve as the input for dataset preparation and causal analysis.

### 3.2 Generating Causal Reasoning Dataset

We leverage Qwen2.5-7B-Instruct [1, 2] to construct a causal reasoning dataset over scene transitions in the music video. Our causal reasoning is grounded in how the visual delta i.e. change in visual features affect the music or audio spectrum. Causal reasoning statements are generated in natural language by feeding the raw extracted features of each pair of consecutive scenes directly to the LLM. Each entry in the causal reasoning dataset captures three components required for downstream MCQ generation:

- **Audio/Music Delta:** the change in the audio and music properties occurring across the transition between two consecutive scenes.
- **Visual Delta:** change in visual features occurring across the transition between two consecutive scenes.
- **Causal Reasoning:** the central component our analysis, which focuses on the relation between visual delta and the resulting musical change. This gives us a very natural-language explanation about how the change in the visual cues affects the music.

Counterfactual questions in the MCQ dataset are subsequently framed around hypothetical scenarios derived from these causal links, with the LLM constructing hypothetical premise from the causal reasoning provided as input.

### 3.3 Generating the KARMA-MV MCQ Dataset

The causal reasoning generated in the previous step serves as a reference data for constructing the MCQ dataset. Questions are formulated to be concise, understandable and interpretable to human actors. The qwen-2.5-7B-Instruct LLM is prompted to generate questions grounded in reasoning statements from the previous step. Following Li et al. [3], each scene transition yields 3 MCQs, each having 4 options, of which only one is correct. The question types are:

- **Evidence Reasoning:** questions that probe the deeper causal relationship between changes in the visual domain and their effects on music, requiring the model to identify the correct causal mechanism.
- **Prediction:** questions that ask how the music will change across a scene transition given the observed change in visual content.
- **Counterfactual:** questions that posit a hypothetical scenario in the context of the scenes and ask what musical outcome would follow, testing the model’s ability to reason about alternative causal chains.

The resulting KARMA-MV dataset contains 12,579 MCQ questions with music video for each category, resulting in a total of 37,737 MCQs.

## 4. CAUSAL MODEL

### 4.1 Causal Knowledge Graph Construction

To model causality explicitly and improve downstream model performance, we augment our VLM baselines with Causal Knowledge information encoded as structured graph representations. A Causal Knowledge Graph (CKG) is a natural representation for this purpose, as it enables efficient search, retrieval, and reuse of causal facts across experiments. Following Xu et al. [17], we construct our causal knowledge based model from the causal reasoning dataset. In this graph, the nodes represent Music feature states and directed edges represent changes in visual features that cause transitions in music feature states. We prompt the same base LLM, Qwen2.5-7B-Instruct, to generate causal triplets from the causal reasoning dataset: for each transition pair, a triplet is produced comprising a *Source* (Change in Visual cues), *Relation* (brief reasoning on how change in video affects music/audio), and *Target* (change in Music/Audio features). These causal triplets serves as ground truth from which final Causal Knowledge Graph is constructed using NetworkX [18]. An extract of the resulting graph is provided in Figure 2.

The constructed causal graph comprises 9,258 nodes and 13,101 edges. While the low graph density (0.00015) and modest average degree (1.42) are consistent with the principle of causal parsimony [19], these metrics specifically define a music state transition network where edges reflect directional, temporally ordered shifts in genre, mood, and instrumentation [cite: 259, 260]. Our topological analysis reveals that nearly half of these music states (48.57%) function as causal initiators or root nodes with

zero in-degree, suggesting they act as spontaneous origins of musical sequences rather than derivatives of prior states.

Within this structure, “Genre: Pop” and “Instruments: Changed” emerge as the primary catalysts for musical evolution, initiating the highest number of downstream transitions (129 and 119, respectively). Furthermore, the decomposition of the graph into 1,357 distinct causal communities suggests the existence of self-contained musical sub-systems with limited cross-community flow. The largest of these clusters encompasses 3,380 nodes, or 36.5% of the total graph, indicating a dominant interconnected structure likely corresponding to mainstream popular music styles, which exists alongside a long tail of smaller, niche stylistic communities.

## 4.2 VLM grounding with CKG

Having constructed the Causal Knowledge Graph, we integrate it with Baseline VLMs to evaluate whether access to structured causal knowledge improves question answering performance. The core idea is to retrieve the most relevant causal facts from the graph for each MCQ at inference time and inject them into the model’s prompt as additional context. To enable efficient retrieval, the raw graph is serialized into column-oriented Parquet files and decomposed into distinct Entity and Relationship dataframes using pandas and NetworkX [18]. We adopt the 1-hop retrieval strategy operating over a NetworkX directed graph. We first extract semantic keywords from the question text and perform a topological search to identify matching anchor nodes within the graph. From each anchor node, we traverse only outgoing edges, which by construction ensures that retrieval captures forward-looking causal effects rather than backward-looking historical causes. We score each retrieved relationship using a bipartite heuristic scoring function: the base weight is derived from the anchor node’s degree centrality; naturally prioritizing highly connected, authoritative musical concepts; while a 1.5x transition boost is applied if the edge description contains kinetic terminology (e.g., increase, shift). Finally, we rank the candidate facts by this composite score and truncate to a top- $k$  list, where  $k$  can be tuned per model to manage the information bottleneck. In our experiment with set  $k$  to 25 for VLMs and 3 for thinking LLMs. We append the resultant causal context to the VLM prompts when answering MCQs in KARMA-MV dataset.

## 4.3 CKG Integration with LLM

For integrating our CKG facts into a thinking and reasoning Language Model, we use structured Zero-shot Chain of Thought (CoT) prompting technique. First we use NetworkX to develop a directed graph, from which causal facts are retrieved using keywords extracted from each MCQ for the given transition pair. A well-defined feature set is then assembled and provided to the LLM as a structured prompt. Specifically, the architecture retrieves relevant information from the graph and feeds them to the LLM alongside the ground truth feature-evidences for the

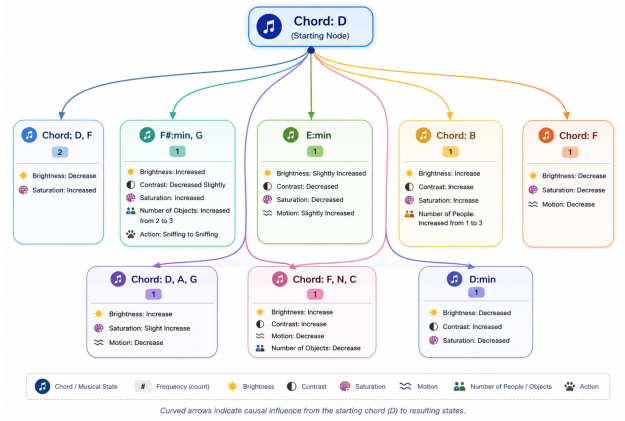


Figure 2: Excerpt from the causal knowledge graph.

transition pairs, and the LLM is then asked to answer the questions based on these information.

## 5. EXPERIMENTAL SETUP

We evaluate the baseline performance of current open VLMs and thinking LLM’s on KARMA-MV MCQs, and subsequently ground them with the developed Causal Knowledge Graph. We include two state-of-the-art VLM models in our experiment: Qwen-2.5-Omni-7B [20] and MiniCPM-o 4.5 [21]. During evaluation, the raw music-video transition scene pairs are directly fed to the VLM and subsequently the questions related to this particular transition pair are asked in the prompt. For prediction-type questions, where the model is asked to predict the music in the second scene, we mute the audio/music track of the second scene using ffmpeg to prevent information leak.

Finally, we use the Gemma-4-31B Instruct LLM [22] as a thinking model on our dataset. As this model cannot consume the raw music video files, we instead provide the extracted music, audio, and visual features for each transition pair as structured text input. First we establish a baseline and then evaluate its performance when augmented with CKG facts following the retrieval and injection procedure described in the methodology.

## 6. RESULTS AND DISCUSSION

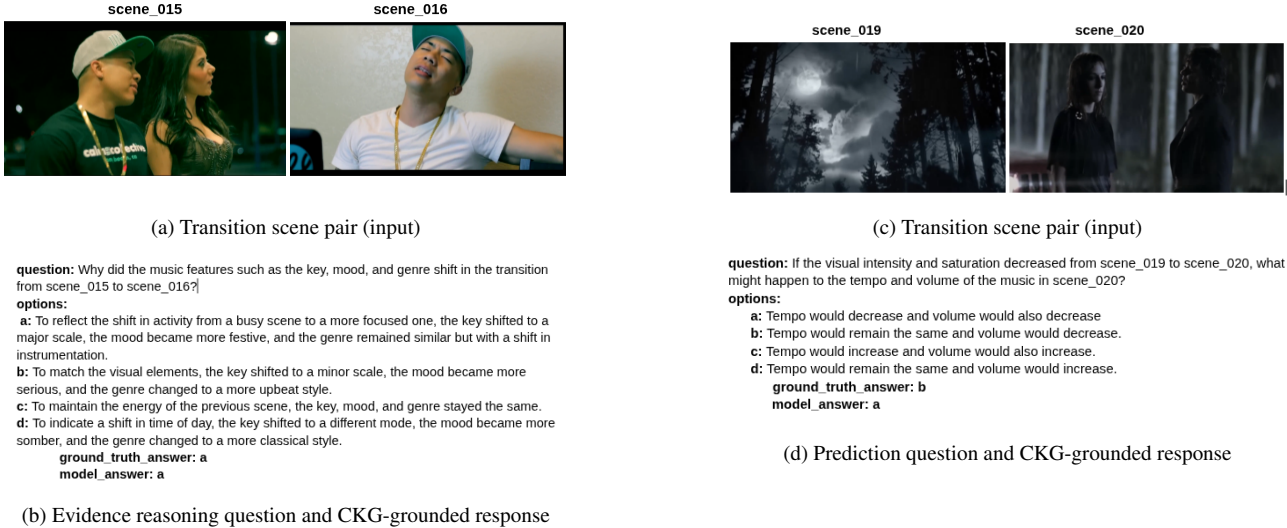
### 6.1 Improving MCQ answering with CKG

#### 6.1.1 Baseline results

Examining the baseline models provides a picture of how the VLMs behave on our dataset without any external causal grounding. As shown in Table 1, Qwen 2.5 Omni achieves an overall accuracy of 66.37%, the lowest among the three baselines. Counterfactual questions prove to be its most significant weakness, with an accuracy of only 56.82%; barely above random chance for a 4-option MCQ; while prediction-type questions are comparatively more tractable (76.57%). This pattern is consistent with the nature of the task: prediction questions require forward temporal inference from observed visual context, whereas

Model Architecture	Type	Evidence Reasoning	Prediction	Counterfactual	Overall
<i>Without Knowledge Graph Grounding</i>					
Qwen-2.5-Omni-7B	VLM	65.73	76.57	56.82	66.37
MiniCPM-o-4.5	VLM	74.06	71.11	69.39	71.51
Gemma-4-31B Instruct	LLM	78.20	73.37	74.11	75.22
<i>With Causal Knowledge Graph (CKG) Grounding</i>					
Qwen-2.5-Omni-7B + CKG	VLM	73.35	78.90	68.98	73.74
MiniCPM-o-4.5 + CKG	VLM	77.69	71.67	72.90	74.09
Gemma-4-31B Instruct + CKG	LLM	79.17	76.68	72.29	76.05

**Table 1:** Accuracy (%) of the Multiple Choice Question (MCQ) task highlighting the performance gains achieved by integrating Causal Knowledge Graph (CKG) facts into VLM and LLM architectures. Random chance baseline is 25%. We note that VLM models are provided the raw music video, whereas the LLM model is provided with the extracted features.



**Figure 3:** Qualitative examples of responses generated by Qwen-2.5-Omni-7B integrated with the CKG on evidence reasoning and prediction question types.

counterfactual questions demand reasoning over hypothetical scenarios with no direct perceptual grounding in the video, a capacity that appears limited in a 7B VLM without external causal context.

MiniCPM-o-4.5 performs substantially better overall (71.51%), with a more balanced accuracy profile across categories (Evidence: 74.06%, Prediction: 71.11%, Counterfactual: 69.39%). Notably, while Qwen shows a pronounced weakness in counterfactual reasoning, MiniCPM’s counterfactual performance (69.39%) is within 5 points of its evidence reasoning score, suggesting that its instruction-tuning and thinking mode contribute to more robust causal generalization.

The best-performing baseline is Gemma-4-31B Instruct (75.22% overall), with relatively strong counterfactual performance (74.11%). It is important to note, however, that Gemma operates on structured feature representations rather than raw video, which simplifies the inference problem and reduces the perceptual complexity of the task. This input modality difference makes direct comparison with the VLMs somewhat asymmetric, though it does not diminish the validity of the CKG integration analysis conducted separately for each model.

### 6.1.2 Causal Knowledge Graph Grounded Architecture

Integrating CKG facts into the model prompts yields consistent accuracy improvements, with the magnitude of improvement varying substantially across models and question types as shown in Table 1.

Qwen-2.5-Omni-7B, benefits most from CKG augmentation, gaining 7.37 percentage points overall (from 66.37% to 73.74%). The category-level breakdown reveals a particularly striking result: the largest absolute gain occurs on counterfactual questions, where accuracy improves by 12.16 points (from 56.82% to 68.98%). This is precisely the question type where Qwen’s baseline was weakest, suggesting that the model lacked the internalized causal knowledge necessary to reason about hypothetical audio-visual scenarios, and that the CKG directly compensates for this deficit by supplying the relevant causal links at inference time. Evidence reasoning also improves substantially (+7.62 points), while prediction questions show a more modest gain (+2.33 points), consistent with the baseline already being strong in that category.

MiniCPM-o-4.5, augmented with the CKG, improves by 2.58 points overall (71.51% to 74.09%). Unlike Qwen, the gains are distributed more uniformly across question types: evidence reasoning improves by 3.63 points, counterfactual by 3.51 points, and prediction by a marginal 0.56 points. This uniform improvement profile suggests that for

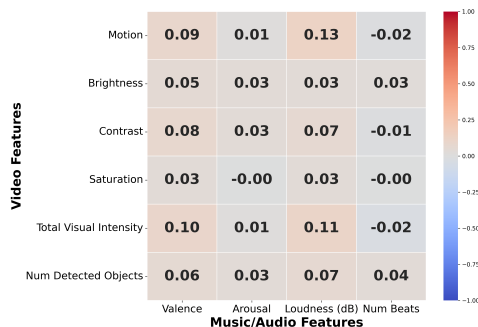
a model with stronger baseline causal reasoning capacity, the CKG provides consistent contextual grounding rather than targeted remediation of a specific reasoning weakness. The smaller absolute gain compared to Qwen is expected given MiniCPM’s stronger starting point.

For Gemma-4-31B Instruct, the CKG yields the smallest overall improvement (+0.83 points, from 75.22% to 76.05%), with gains concentrated in evidence reasoning (+0.97) and prediction (+3.31) categories. Notably, counterfactual accuracy *decreases* by 1.82 points (from 74.11% to 72.29%) when the CKG is integrated. We observe that when the thinking LLM is given too many facts, it tends to hallucinate or over-rely on the provided context, which disrupts the logical consistency required for counterfactual inference. This suggests that for large models with rich internalized causal knowledge, injecting excessive graph-derived constraints can restrict rather than support their capacity for open-ended hypothetical reasoning.

### 6.1.3 Model Capacity and CKG Benefit: An Inverse Scaling Observation

A consistent pattern emerges across all three models: the benefit from CKG augmentation is inversely related to model capacity. The 7B Qwen model gains 7.37 percentage points overall, the 9B MiniCPM gains 2.58 points, and the 31B Gemma gains only 0.83 points. This trend is consistent with the hypothesis that larger models internalize more causal world knowledge during pre-training, reducing the marginal contribution of explicit external knowledge at inference time. Conversely, smaller models operating on tasks that demand causal reasoning beyond their parametric capacity stand to gain the most from structured knowledge injection.

This observation has a practical implication for benchmark design and system deployment: CKG augmentation is a particularly effective compensatory mechanism for resource-constrained inference settings, where deploying a smaller model augmented with a lightweight knowledge graph may approach or match the performance of a larger model without graph grounding. In our experiments, Qwen-2.5-Omni-7B with CKG (73.74%) substantially narrows the gap to the unaugmented MiniCPM-o-4.5 (71.51%) and approaches MiniCPM with CKG (74.09%), despite having fewer than half the parameters.



**Figure 4:** Spearman rank correlation heatmap illustrating cross-modal dependencies between video features (y-axis) and audio/music features (x-axis)

## 6.2 Cross-modal Dependencies in Dataset

To evaluate audio-visual dependencies, we compute Spearman rank correlations across all 22,011 scenes (Figure 4). The positive correlations between Total Visual Intensity (TVI) and loudness, and between motion and arousal, provide quantitative validation that KARMA-MV is grounded in genuine cross-modal dependencies rather than incidental co-occurrence or dataset artifacts. Furthermore, the correlation between visual saturation and valence aligns with prior work on color-affect mapping, suggesting that even low-level visual features contribute to audio-visual coherence. Together, these patterns establish a quantitative baseline for benchmarking future causal frameworks in the music-video domain.

## 6.3 Discussion and Future Work

Causal reasoning about music-video relationships is inherently difficult even for trained human annotators: correctly attributing a change in musical key or rhythmic density to a specific visual shift requires simultaneous expertise in music theory and visual analysis, making human annotation unreliable. This is precisely why KARMA-MV adopts an automated pipeline grounded in high-accuracy specialized feature extractors and a well-grounded state-of-the-art LLM, trading the noise of lay annotation for consistent, reproducible feature-grounded generation.

Future work could address this through adaptive retrieval strategies that modulate the amount of injected causal context based on model size or confidence, reducing the anchoring effect for larger models. Additional directions include richer causal interventions beyond scene transitions, broader genre coverage, and more expressive graph representations that better capture the graded nature of audio-visual causal influence.

## 7. CONCLUSION

We introduced KARMA-MV<sup>1</sup>, a large-scale benchmark for causal question answering in music videos, targeting how changes in visual content drive changes in music and audio. The dataset is constructed entirely through automated, feature-grounded generation, reflecting the genuine difficulty of the task for human annotators. Experiments across three state-of-the-art models demonstrate that explicit causal knowledge graph grounding yields consistent gains—particularly for smaller models, where the CKG most directly compensates for limited internalized causal knowledge. The inverse scaling pattern observed, where larger models benefit less from external causal context, suggests that structured knowledge injection is most valuable in resource-constrained deployment settings. The dataset, causal knowledge graph as well as the VLM/LLM setup are available open-source<sup>2</sup>.

KARMA-MV establishes a scalable testbed and strong baseline for causal audio-visual reasoning, with clear di-

<sup>1</sup> <https://huggingface.co/datasets/amaai-lab/karma-mv>

<sup>2</sup> <https://github.com/AMAai-Lab/Karma-MV>

rections for future work in adaptive knowledge retrieval, richer causal interventions, and broader genre coverage.

## 8. ACKNOWLEDGMENTS

This work has received funding from grant no. SUTD SKI 2021\_04\_06 and from MOE grant no. MOE-T2EP20124-0014

## 9. AI USAGE STATEMENT

We acknowledge the use of Gemini and Claude for grammar improvements.

## 10. REFERENCES

- [1] Qwen Team, “Qwen2.5: A party of foundation models,” Qwen Blog, September 2024. [Online]. Available: <https://qwenlm.github.io/blog/qwen2.5/>
- [2] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu *et al.*, “Qwen2 technical report,” 2024. [Online]. Available: <https://arxiv.org/abs/2407.10671>
- [3] J. Li, L. Niu, and L. Zhang, “From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA, 18–24 Jun. 2022, pp. 21 273–21 282.
- [4] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, “Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 8992–8999, Apr. 2020. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/6431>
- [5] K. Khurana and U. Deshpande, “Video question-answering techniques, benchmark datasets and evaluation metrics leveraging video captioning: A comprehensive survey,” *IEEE Access*, vol. 9, pp. 43 799–43 823, 2021.
- [6] M. Tapaswi, Y. Zhu, R. Stiefelbogen, A. Torralba, R. Urtasun, and S. Fidler, “MovieQA: Understanding stories in movies through question-answering,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, 27–30 Jun. 2016, pp. 4631–4640.
- [7] J. Lei, L. Yu, M. Bansal, and T. Berg, “TVQA: Localized, compositional video question answering,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Oct–Nov 2018, pp. 1369–1379.
- [8] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao, “ActivityNet-QA: a dataset for understanding complex web videos via question answering,” ser. AAAI’19/IAAI’19/EAAI’19. AAAI Press, 2019. [Online]. Available: <https://doi.org/10.1609/aaai.v33i01.33019127>
- [9] J. Xiao, X. Shang, A. Yao, and T.-S. Chua, “NExT-QA: Next phase of question-answering to explaining temporal actions,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, June 2021, pp. 9772–9781. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00965>
- [10] G. Li, Y. Wei, Y. Tian, C. Xu, J.-R. Wen, and D. Hu, “Learning to answer questions in dynamic audio-visual scenarios,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 19 108–19 118.
- [11] B. Castellano, “PySceneDetect.” [Online]. Available: <https://github.com/Breakthrough/PySceneDetect>
- [12] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “librosa: Audio and music signal analysis in python,” in *Proc. 14th Python in Science Conf. (SCIPY 2015)*, Austin, USA, 06–12 Jul. 2015, pp. 18–25.
- [13] J. Kang and D. Herremans, “Towards unified music emotion recognition across dimensional and categorical models,” *arXiv preprint arXiv:2502.03979*, 2025.
- [14] A. Chopra, A. Roy, and D. Herremans, “MIRFLEX: Music information retrieval feature library for extraction,” in *Proc. of the Late-Breaking Demo Session of the 25th International Society for Music Information Retrieval Conf. (ISMIR)*, San Francisco, United States, 2024.
- [15] P. Alonso-Jiménez, D. Bogdanov, J. Pons, and X. Serra, “Tensorflow audio models in essentia,” in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 266–270.
- [16] G. Jocher, A. Chaurasia, and J. Qiu, “Ultralytics yolov8,” 2023, version 8.0.0. [Online]. Available: <https://github.com/ultralytics/ultralytics>
- [17] Z. Xu and Y. Dang, “Data-driven causal knowledge graph construction for root cause analysis in quality problem solving,” *International Journal of Production Research*, vol. 61, no. 10.
- [18] A. Hagberg, P. Swart, and D. Chult, “Exploring network structure, dynamics, and function using networkx,” 06 2008.
- [19] J. Pearl, “Causal inference,” *Causality: objectives and assessment*, pp. 39–58, 2010.
- [20] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, B. Zhang, X. Wang, Y. Chu, and J. Lin, “Qwen2.5-omni technical report,” 2025. [Online]. Available: <https://arxiv.org/abs/2503.20215>

- [21] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He *et al.*, “Minicpm-v: A gpt-4v level mllm on your phone,” *arXiv preprint arXiv:2408.01800*, 2024.
- [22] Google, “Gemma 4 31b it,” 2026. [Online]. Available: <https://huggingface.co/google/gemma-4-31B-it>