HarmonySet: A Comprehensive Dataset for Understanding Video-Music Semantic Alignment and Temporal Synchronization

Supplementary Material

A. More details of HarmonySet

A.1. Dataset Construction pipeline

Videos in HarmonySet are sourced from the YouTube Shorts platform and were crawled using 293 keywords we designed. The complete list of keywords is shown in Figure 6.

Figure 7 illustrates the multi-phase annotation process. The raw data we crawled includes videos along with their audio and metadata. The metadata includes the title, author, duration, and video width and height, as shown in Figure 8. We used PANNs [61] for music tagging and video filtering. The music tags generated by PANNs were added to the metadata, and videos without music were filtered out based on the tagging results. The filtering criterion was: if the labels with top 2 probabilities do not include ‘music’, the video is deleted. The filtered videos were then assigned to human annotators for detailed screening to **ensure video-music pair quality** and to **exclude non-ethical and sensitive content**. The instructions for the annotators were:

Music Check: *If there is no background music (e.g., pure human voice, pure environmental sound, pure noise, no sound, etc. Music mixed with human voice counts as having music), please flag the video. Listen to the entire video before making this determination, as music may only be present in a portion of the video.*

Content Suitability Check: *Carefully review the entire video for any content that is: Non-Ethical: This includes, but is not limited to, content that promotes or depicts illegal activities, harmful behavior, or discrimination. Sensitive: This includes, but is not limited to, content that is sexually suggestive, graphically violent, or exploits, abuses, or endangers children.*

Video Quality Check: *Please also assess the overall video quality. Flag any videos with technical issues, such as severe distortion, extremely poor resolution, or corrupted files.*

Human Annotation We conducted a rigorous annotator selection process. We recruited 120 annotators to pre-annotate 500 videos. The 120 annotators are all experts who have previous formal experience in video annotation work. After the pre-annotation, we retained 25 individuals who demonstrated both accuracy, diversity, and speed in their annotations. For videos with music, human annotators were to mark key time points and label tags. In the key time point annotation, annotators first identified moments representing visual narrative turning points or key points, then determined whether the music synchronously changed with the video at those moments. The instructions were:

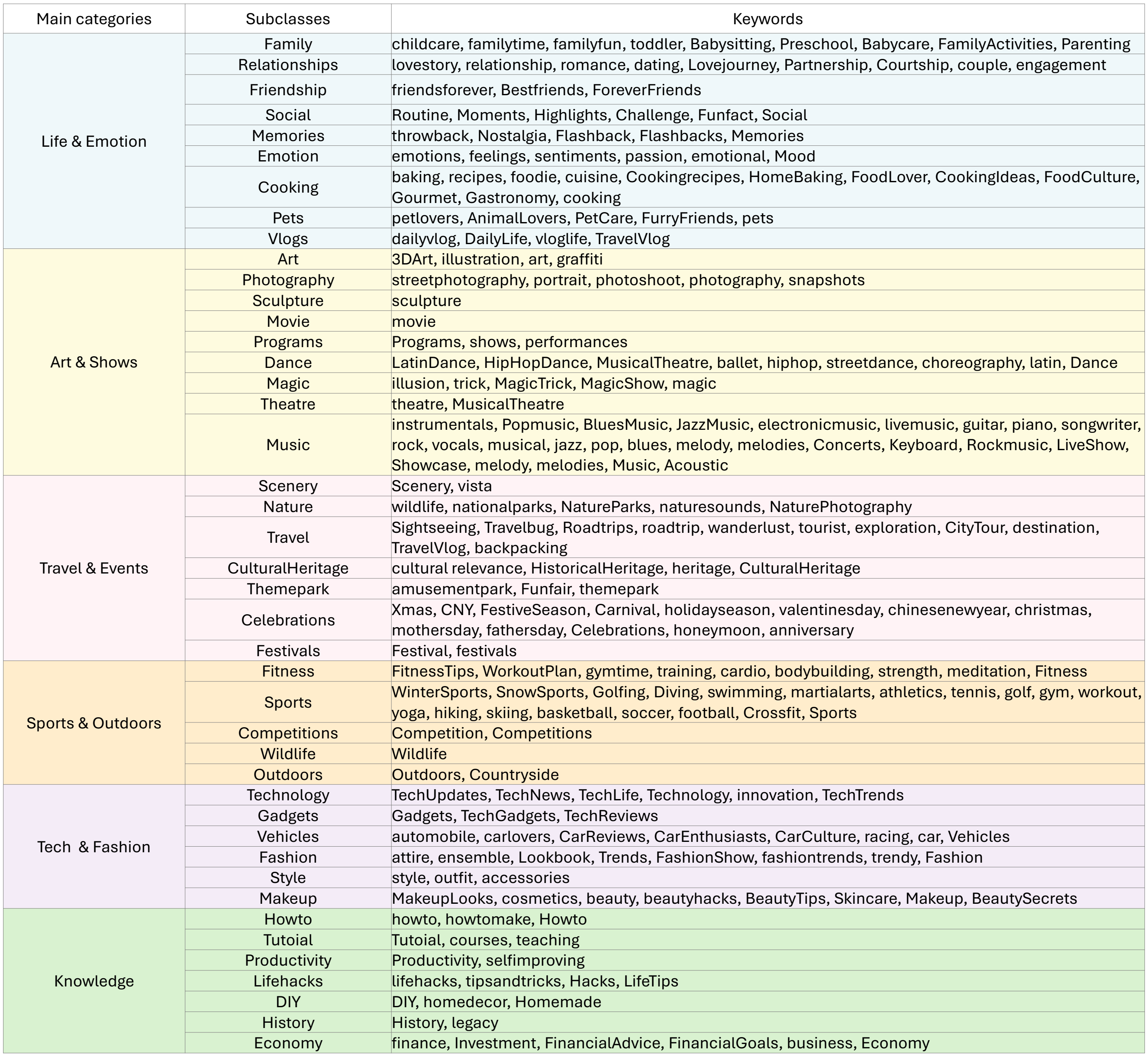
Please mark up to three important time points in the video. If there are no changes throughout the video, fill in 0. Then determine: A) The music changes precisely in sync with the video at the turning point; B) The music changes near the turning point but is not strongly synchronized; C) The music does not change when the video turns. The answer format should be: video timestamp + comma + uppercase letter option, separated by semicolons between time points. Example: 00:10,A; 00:20,B; 00:30,C (Non-synchronization means the visual changes but the music remains the same. Examples of synchronized music changes include [outfit change on beat], [music changes to a victorious tune after a basketball shot], [music reaches a climax as the video reaches its most exciting moment], etc.)

For label tagging, the structured label system is shown in Figure 9.

Automatic Annotation After human annotation, in the automatic annotation phase, the MLLM will receive the video and audio content, human annotation results, and required metadata as input. An example of the metadata is shown in Figure 8, where the video title and audio tags will be used in the automatic annotation process. The MLLM will generate detailed video-music alignment annotations, including semantic alignment and temporal synchronization understanding. We use Gemini 1.5 Pro as the MLLM for the automatic annotation phase, with specific instructions shown in Figure 10. In addition, to ensure the diversity of instructions and to avoid overestimation of performance, we employed multiple prompt templates for the instruction tuning data, as illustrated in Figure 11. These instructions convey the same underlying meaning while avoiding rigid patterns in sentence structure and word choice. The instructions in the dataset will be randomly assigned to one of these ten templates, promoting variability and enhancing the robustness of the training process. This not only helps in capturing a wider range of expressions but also mitigates the risk of the model becoming overly reliant on specific phrasing, thereby improving its generalization capabilities.

A.2. More statistics

The raw data crawled from the platform consists of 59,771 video-music pairs. After the first round of filtering, the number of video-music pairs was reduced to 49,610. Following a meticulous manual screening, the total number of videos was further reduced to 48,328. In addition to statistics including video categories, video duration, and the number of words



Main categories	Subclasses	Keywords
Life & Emotion	Family	childcare, familytime, familyfun, toddler, Babysitting, Preschool, Babycare, FamilyActivities, Parenting
	Relationships	lovestory, relationship, romance, dating, Lovejourney, Partnership, Courtship, couple, engagement
	Friendship	friendsforever, Bestfriends, ForeverFriends
	Social	Routine, Moments, Highlights, Challenge, Funfact, Social
	Memories	throwback, Nostalgia, Flashback, Flashbacks, Memories
	Emotion	emotions, feelings, sentiments, passion, emotional, Mood
	Cooking	baking, recipes, foodie, cuisine, Cookingrecipes, HomeBaking, FoodLover, CookingIdeas, FoodCulture, Gourmet, Gastronomy, cooking
	Pets	petlovers, AnimalLovers, PetCare, FurryFriends, pets
	Vlogs	dailyvlog, DailyLife, vloglife, TravelVlog
Art & Shows	Art	3DArt, illustration, art, graffiti
	Photography	streetphotography, portrait, photoshoot, photography, snapshots
	Sculpture	sculpture
	Movie	movie
	Programs	Programs, shows, performances
	Dance	LatinDance, HipHopDance, MusicalTheatre, ballet, hiphop, streetdance, choreography, latin, Dance
	Magic	illusion, trick, MagicTrick, MagicShow, magic
	Theatre	theatre, MusicalTheatre
	Music	instrumentals, Popmusic, BluesMusic, JazzMusic, electronicmusic, livemusic, guitar, piano, songwriter, rock, vocals, musical, jazz, pop, blues, melody, melodies, Concerts, Keyboard, Rockmusic, LiveShow, Showcase, melody, melodies, Music, Acoustic
Travel & Events	Scenery	Scenery, vista
	Nature	wildlife, nationalparks, NatureParks, naturesounds, NaturePhotography
	Travel	Sightseeing, Travelbug, Roadtrips, roadtrip, wanderlust, tourist, exploration, CityTour, destination, TravelVlog, backpacking
	CulturalHeritage	cultural relevance, HistoricalHeritage, heritage, CulturalHeritage
	Themepark	amusementpark, Funfair, themepark
	Celebrations	Xmas, CNY, FestiveSeason, Carnival, holidayseason, valentinesday, chinesenewyear, christmas, mothersday, fathersday, Celebrations, honeymoon, anniversary
	Festivals	Festival, festivals
Sports & Outdoors	Fitness	FitnessTips, WorkoutPlan, gymtime, training, cardio, bodybuilding, strength, meditation, Fitness
	Sports	WinterSports, SnowSports, Golfing, Diving, swimming, martialarts, athletics, tennis, golf, gym, workout, yoga, hiking, skiing, basketball, soccer, football, Crossfit, Sports
	Competitions	Competition, Competitions
	Wildlife	Wildlife
	Outdoors	Outdoors, Countryside
Tech & Fashion	Technology	TechUpdates, TechNews, TechLife, Technology, innovation, TechTrends
	Gadgets	Gadgets, TechGadgets, TechReviews
	Vehicles	automobile, carlovers, CarReviews, CarEnthusiasts, CarCulture, racing, car, Vehicles
	Fashion	attire, ensemble, Lookbook, Trends, FashionShow, fashiontrends, trendy, Fashion
	Style	style, outfit, accessories
	Makeup	MakeupLooks, cosmetics, beauty, beautyhacks, BeautyTips, Skincare, Makeup, BeautySecrets
Knowledge	Howto	howto, howtomake, Howto
	Tutoial	Tutoial, courses, teaching
	Productivity	Productivity, selfimproving
	Lifehacks	lifehacks, tipsandtricks, Hacks, LifeTips
	DIY	DIY, homedecor, Homemade
	History	History, legacy
	Economy	finance, Investment, FinancialAdvice, FinancialGoals, business, Economy

Figure 6. We employed a hierarchical keyword taxonomy for video collection, comprising six primary categories and 43 subcategories, yielding a total of 293 unique keywords. This taxonomy was meticulously designed to target videos with high-quality music. Keywords unrelated to music, such as *news broadcast* and *read*, were excluded to enhance the precision of the search and prioritize music-centric content. The resulting keyword set was then utilized to crawl and curate a collection of videos.

in the annotation text shown in Figure 2, we also compiled statistics on the frequency of music tags and keywords in video titles (with meaningless stop words like *I* and *the* removed). Figure 12 presents word clouds for music tags and video titles, illustrating the diversity in music genres, styles, and instruments, as well as the variety of content and themes showcased in the videos.

A.3. Benchmark

Metrics of HarmonySet-OE

Traditional language metrics are mainly sensitive to lexical variations and cannot identify changes in sentence semantics.

BLEU-4 BLEU [62], or Bilingual Evaluation Understudy, is a metric for evaluating machine translation by comparing N-grams of the translation to human references. BLEU-4 specifically evaluates the match of four-word sequences and

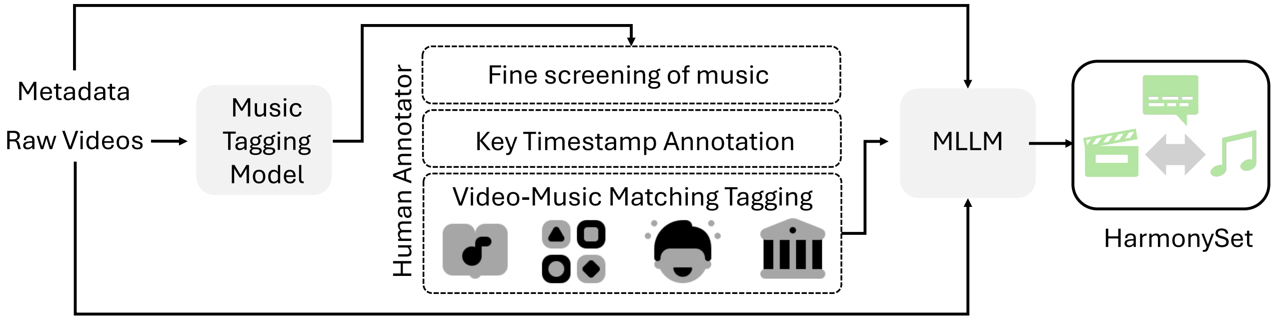
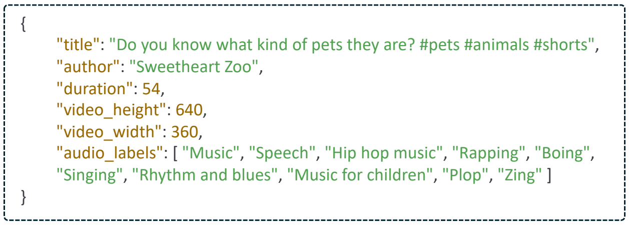


Figure 7. Multi-phase annotation process



```
{
    "title": "Do you know what kind of pets they are? #pets #animals #shorts",
    "author": "Sweetheart Zoo",
    "duration": 54,
    "video_height": 640,
    "video_width": 360,
    "audio_labels": [ "Music", "Speech", "Hip hop music", "Rapping", "Boing",
    "Singing", "Rhythm and blues", "Music for children", "Plop", "Zing" ]
}
```

Figure 8. An example of metadata of video-audio pairs. The top 10 most probable music labels generated by the PANNs are retained. Metadata used in the automated annotation generation pipeline includes video titles and music tags.

includes a Brevity Penalty to account for shorter translations.

ROUGE-L ROUGE [63], or Recall-Oriented Understudy for Gisting Evaluation, measures the overlap between generated and reference summaries. ROUGE-L focuses on the Longest Common Subsequence (LCS), which does not require consecutive word matches.

BLEURT BLEURT [72], or Bilingual Evaluation Understudy with Representations from Transformers, evaluates machine translation and natural language generation using pre-trained language models like BERT. It captures deep semantic information, addressing issues like synonym substitution and sentence rearrangement better than traditional metrics.

CIDEr CIDEr [73], or Consensus-based Image Description Evaluation, is designed for image captioning tasks. It uses TF-IDF weights to emphasize n-grams common in human annotations but rare in general descriptions, capturing more detailed information.

WUPS The Wu-Palmer similarity (WUPS) [74] measures the similarity of word senses based on their positions in the WordNet taxonomy. However, it struggles with words that are similar in form but different in meaning and cannot handle phrases or sentences effectively.

These traditional Natural Language Generation metrics lack the ability to understand and evaluate text with complex logic. In contrast, large language models are proven to be capable of comprehending text deeply. We employed GPT-4o to evaluate similarity between model responses and ground truth annotations in terms of their semantic meaning or the true intent they express. Specific prompt can be seen in Figure 13.

HarmonySet-MC Curation We followed the QA curation pipeline of widely used multiple-choice QA benchmarks like EgoSchema [10] and VideoMME [13] to design the extended benchmark HarmonySet-MC. We used a large language model to generate three incorrect answers for each annotation that is used as the correct answer. After iterating through multiple prompt versions, we ensured that the final multiple-choice questions were challenging yet reasonable, avoiding any form-based biases beyond semantics. Detailed prompt can be seen in Figure 14. For the LLM selection, we tested GPT4o, Claude, and GPT4, ultimately choosing GPT4o as the model for generating the multiple-choice questions.

B. More details of Experiment

B.1. Baseline Introduction

Our open-source model baselines include VideoLLaMA2 [23] and Video-SALMONN [24], both state-of-the-art video-audio multimodal large language models. Earlier MLLMs capable of processing both video and audio, such as Macaw-LLM [26], are no longer available due to a lack of maintenance and therefore were not included in our experiments.

Video-LLaMA2 VideoLLaMA2 is a Video Large Language Model designed for spatial-temporal modeling and audio understanding. VideoLLaMA2 comprises a vision-language branch, an audio-language branch, a Spatial-Temporal Convolution Connector (STC Connector), and a Large Language Model (LLM). The vision-language branch uses a CLIP image encoder (ViT-L/14) to process individual frames, then aggregates these features using a novel Spatial-Temporal Convolution Connector (STC Connector) designed to preserve spatial-temporal information efficiently. The audio-language branch transforms audio into spectrograms, encodes them using BEATs, and then uses an MLP to align the audio features with the LLM. The chosen LLMs for this

Rhythm & Synchronization	Music pace	fast	slow	medium	diverse							
	Rhythm alignment	precise	appropriate	irrelevant	mismatch							
Theme & Content	Content alignment	strongly related	indirectly related	unrelated	conflicting							
	Narrative supplement	suggestion	enhancement	reversing	independent	no supplement						
Emotion	Video Emotions	happy	sad	angry	fear	anxious	creative	hope	surprise	humor	relax	···
	Emotion alignment	consistent	partially consistent	contrast	imbalanced	unrelated						
	Emotional richness	rich	single	abrupt								
Culture	Culture elements	clothing	customs	cuisines	landmarks	no culture element						
	Cultural relevance	related	misuse	unrelated								

Figure 9. A hierarchical labeling system is employed for manual annotation, encompassing four primary aspects: Rhythm and Synchronization, Theme and Content, Emotion, and Culture. These aspects are further refined into nine sub-aspects, with labels representing factual information or degrees of match.

architecture are Mistral-Instruct and Qwen2-Instruct. The STC Connector prioritizes maintaining spatial-temporal order, minimizing token count, and mitigating information loss during downsampling through the use of 3D downsampling and RegStage convolution blocks.

Video-SALMONN Video-SALMONN is designed for obtaining fine-grained temporal information required by speech understanding. Video-SALMONN uses pre-trained encoders for visual (InstructBLIP), speech (Whisper), and non-speech audio (BEATs) inputs. These features are temporally synchronized, aligning audio and visual features at the video frame rate (2Hz). A Multi-Resolution Causal (MRC) Q-Former then processes these synchronized features at different time scales (e.g., 1, 5, and 10-second windows) to capture fine-grained audio-visual joint representations.

B.2. Implementation Details

For VideoLLaMA2 training, we utilized 4 NVIDIA H800 GPUs (a total of 32 GPUs). The training configuration followed the default settings of VideoLLaMA2-AV, except for the learning rate (lr), which was adjusted to 1e-5. The model was trained for 2 epochs on our dataset. All testing was conducted on a single NVIDIA H800 GPU (a total of 8 GPUs). Experiments in Table 2 used 16 frames for testing. For experiments in Table 10, which explore the impact of varying frame numbers, the number of frames used for testing matched the number of frames used during training (16/32/64 frames).

B.3. More analysis of main results

In Table 2, we present scores for each model across six main video categories and four evaluation aspects. A horizontal comparison across the six video categories reveals that models generally score lowest on knowledge-based videos and highest on arts and performance videos. This may be because knowledge-based videos often require deeper semantic understanding and factual recall. The criteria for evaluating knowledge-based videos might be more stringent, reflecting the need for accurate information retrieval. On the other hand, the evaluation of arts and performance videos could be more subjective and open to interpretation, potentially leading to higher scores. A vertical comparison across the four evaluation aspects shows that all models consistently score lowest on the cultural aspect, suggesting that understanding and evaluating cultural nuances remains a significant challenge for current models. This could be due to the inherent complexity and subjectivity of cultural interpretations, and the current models lack sufficient training data that adequately represents the diversity and depth of cultural contexts. This deficiency hinders their ability to accurately assess culturally relevant aspects of the videos.

While Gemini-1.5 Pro generally performs well, the HarmonySet-enhanced VideoLLaMA2 demonstrates that open-source models can achieve comparable or even superior performance. This highlights the potential of open-source development in the MLLM domain. However, the base VideoLLaMA2 and video-SALMONN lag signifi-

Gemini Prompt for Automated Annotation Generation

```
Analyze the provided video and music, developing a preliminary interpretation by generating the most fitting
description for each of the following four aspects: Rhythm and Synchronization, Theme and Content, Emotion,
and Cultural Relevance.
•  Rhythm and Synchronization: Analyze how well the music's tempo (fast or slow) matches the video's pace,
   the level of synchronization between the music's rhythm and the video's actions or cuts, and the overall
   rhythmic coherence between the two.
•  Theme and Content: Analyze how well the music's theme complements the video's content, whether the music
   enhances the video's narrative or message, and if there is any thematic dissonance between the two.
•  Emotion: Analyze how well the music's style and the video's emotional tone match, identify the specific
   emotions conveyed by both the music (e.g., joyful, melancholic, suspenseful) and the video, assess the
   level of emotional coherence between them, and consider whether the music elevates the emotional impact of
   the video.
•  Cultural Relevance: Analyze whether the music and video share any common cultural elements, assess the
   relevance and accuracy of the cultural elements used, and identify any specific geographical references in
   both the music and video and how well they align.
Focus your analysis on the detailed reason of match or mismatch between the video and the music for each
aspect. Describe elements of both the video and the music and explain why you believe they match or mismatch.

Alongside this video and music, You will be provided with manually labeled tags representing the ground
truth. Use these tags to refine and enhance your initial interpretations. Specifically, integrate the
timestamps indicated in the sync_answer.value tag into your analysis of Rhythm and Synchronization,
explaining how these synchronized (or unsynchronized) moments shape your overall understanding of the video
and music‘s rhythmic relationship. The labels offer detailed insights, including the sync_answer.value tag
which specifies key video timestamps and their corresponding musical synchronization. Within this tag, ‘A’
denotes precise synchronization between video transitions and musical changes; ‘B’ indicates a musical
change near the transition point, but without strong synchronization; and ‘C’ signifies no musical change
during the video transition. 0 means no key timestamps.

You will also be provided with the video’s metadata including title and music tags. Use this metadata to
supplement and support your interpretation. Music tags are only used for reference, and latter tags in the
ten tags are generally not associated with music. It cannot be assumed that tag necessarily represents
musically related content.

Finally, articulate your comprehensive and nuanced interpretation in your own words, without explicitly
mentioning the labels themselves. Ensure your interpretation accurately and thoroughly reflects the
information conveyed by the labels.
For each aspect, provide a rich and detailed explanation, exploring the nuances and subtleties you observe,
always focusing on the interplay and degree of alignment between the visual and auditory elements. You can
include your own opinion to fulfill your answer.

The output format should be:

Rhythm and Synchronization: <your detailed explanation>
Theme and Content: <your detailed explanation>
Emotion: <your detailed explanation>
Cultural Relevance: <your detailed explanation>
```

Figure 10. Prompt for generating automated annotation using Gemini 1.5 Pro. Inputs include video and audio content, manual annotated labels, and metadata. The model is tasked with providing a detailed understanding of the match across four aspects.

cantly, indicating that further research and development are needed to close the gap with closed-source models without relying on additional datasets like HarmonySet.

B.4. Experiments on General Audio-Visual Tasks

Table 7. Performance on AVSD for General Audio-Visual QA.

Metrics	BLEU	BLEU-4	ROUGE	BERT
VideoLLaMA2	0.28	0.19	0.33	0.87
VideoLLaMA2 (HarmonySet)	**0.30**	**0.21**	**0.35**	**0.89**

Table 7 presents the performance of VideoLLaMA2 on AVSD dataset before and after training with HarmonySet. AVSD is a widely used dataset for audio-visual question answering. It includes general audio-visual QA tasks, such as ”Do you hear any audio at all?” and ”Is there a violin sound in the background of the video?” From the table, it is evident that the VideoLLaMA2 trained with HarmonySet has achieved improved performance across all metrics. This indicates that HarmonySet not only enhances the model’s understanding of the relationships between

```
# Diversified Instructions

"<video>\nExamine the given video and soundtrack, assessing their alignment in four critical areas: rhythm and
timing, thematic elements, emotional resonance, and cultural significance. Offer a detailed analysis for each
dimension, using specific timestamps to highlight how moments of synchronization (or lack thereof) enhance the
overall effectiveness of the audio-visual experience.",

"<video>\nEvaluate the supplied video and music, focusing on their compatibility in four essential aspects:
rhythm and coordination, thematic relevance, emotional effect, and cultural importance. Provide an in-depth
explanation for each category, referencing specific timestamps to demonstrate how instances of synchronization
(or their absence) influence the overall impact of the combined audio-visual presentation.",

"<video>\nAssess the provided video alongside the music, analyzing their compatibility across four main
dimensions: rhythm and alignment, thematic substance, emotional influence, and cultural context. Deliver a
thorough explanation for each aspect, incorporating specific timestamps to illustrate how moments of
synchronization (or their lack) contribute to the overall effectiveness of the audio-visual experience.",

"<video>\nAnalyze the video and accompanying music, evaluating their compatibility in four key areas: rhythm
and harmony, thematic expression, emotional depth, and cultural relevance. Provide a comprehensive breakdown
for each dimension, using specific timestamps to show how moments of synchronization (or their absence) affect
the overall effectiveness of the audio-visual experience.",

"<video>\nInvestigate the provided video and its music, assessing their compatibility across four fundamental
dimensions: rhythm and synchronization, thematic content, emotional impact, and cultural significance. Offer a
detailed analysis for each aspect, referencing specific timestamps to illustrate how moments of synchronization
(or their lack) enhance the overall effectiveness of the combined audio-visual experience.",

"<video>\nReview the given video and soundtrack, focusing on their compatibility in four critical areas: rhythm
and synchronization, thematic elements, emotional resonance, and cultural relevance. Provide a thorough
explanation for each dimension, incorporating specific timestamps to highlight how moments of synchronization
(or their absence) contribute to the overall effectiveness of the audio-visual experience.",

"<video>\nCritique the provided video and music, evaluating their compatibility across four key dimensions:
rhythm and timing, thematic content, emotional impact, and cultural significance. Deliver a comprehensive
analysis for each aspect, using specific timestamps to illustrate how moments of synchronization (or their
absence) influence the overall effectiveness of the audio-visual experience.",

"<video>\nExplore the relationship between the provided video and music, assessing their compatibility in four
main areas: rhythm and synchronization, thematic relevance, emotional effect, and cultural importance. Provide
an in-depth explanation for each dimension, referencing specific timestamps to demonstrate how moments of
synchronization (or their lack) affect the overall impact of the audio-visual presentation.",

"<video>\nAnalyze the video and its accompanying music, evaluating their compatibility across four essential
dimensions: rhythm and coordination, thematic substance, emotional influence, and cultural context. Offer a
detailed breakdown for each aspect, incorporating specific timestamps to illustrate how moments of
synchronization (or their absence) contribute to the overall effectiveness of the audio-visual experience.",

"<video>\nExamine the provided video and music, focusing on their compatibility in four critical areas: rhythm
and alignment, thematic expression, emotional depth, and cultural relevance. Provide a comprehensive
explanation for each dimension, using specific timestamps to show how moments of synchronization (or their lack)
enhance the overall effectiveness of the combined audio-visual experience.",

"<video>\nAnalyze the provided video and music, evaluating their compatibility across four key dimensions:
rhythm and synchronization, thematic content, emotional impact, and cultural relevance. Provide a comprehensive
explanation for each aspect, incorporating specific timestamps to illustrate how moments of synchronization (or
their absence) contribute to the overall effectiveness of the combined audio-visual experience."
```

Figure 11. Diversified instructions for HarmonySet data

video and music but also proves beneficial for conventional tasks, such as perception. We will further investigate the specific insights and capabilities that HarmonySet can provide, aiming to deepen our understanding of its impact on model performance and its potential applications in various audio-visual tasks.

B.5. More details of ablations

Full ablation on different training data Table 9 shows the results of VideoLLaMA2 when not trained, trained with 10k MLLM auto-generated data, trained with 10k HarmonySet data, and trained with the entire HarmonySet data. It provides a more intuitive comparison on different types of training data, demonstrating the effectiveness and importance of incorporating human knowledge through HarmonySet.

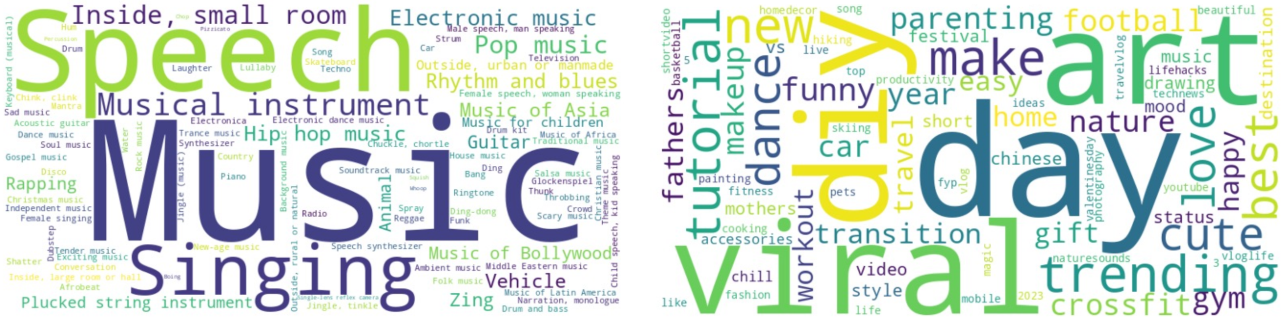


Figure 12. (Left) Word cloud visualizations of high-frequency music tags extracted by PANNs (excluding stop words). The larger the word, the higher its frequency. This showcases the diversity of musical genres, instruments, and cultures. (Right) Word cloud visualizations of video titles (excluding stop words). The word cloud demonstrates the wide range of video types and diverse scenes included in the dataset.

# GPT Prompt for HarmonySet-OE Evaluation

```
Please evaluate the similarity between the following response and the correct answer in terms of their
semantic meaning or the true intent they express. Provide scores based on the following four aspects:
1. Rhythm and Synchronization
2. Thematic Content
3. Emotional Impact
4. Cultural Relevance

Provide a score for each aspect on a scale of 1 to 10, where a higher score indicates better semantic
similarity. Please output a single line containing only four values indicating the scores for four aspects,
respectively. Do not output  any words other than scores. The 4 scores are separated by a comma. Remember to
rate these 4 aspects respectively, and score for one aspect should not be affect be others. If an aspect is
not addressed or is irrelevant, give a lower score.

User Response: {response}\n\nCorrect Answer: {correct_answer}\n\n
```

Figure 13. Detailed prompt for using GPT4o as the evaluation metric for HarmonySet-OE. GPT4o will receive the correct answer and the model's output response and output the similarity of the response to the true answer in four aspects, assessing semantic and factual similarity rather than mere word matching.

Table 8. Performance of humans and models on 100 questions from HarmonySet-OE. Results show a noticeable gap between even the best model's performance and human performance, highlighting the limitations of current models in generating open-ended responses.

	R & S	T	E	C
VideoLLaMA2 (HarmonySet)	5.49	5.10	5.25	4.77
Human	**7.38**	**7.02**	**7.57**	**6.32**

Ablation on Number of Frames In Table 10 we provide the results of ablation on number of frames across all six categories. Increasing frames from 16 to 32 demonstrably improves performance, highlighting the importance of sufficient temporal context. However, the performance degradation with 64 frames reveals that more frames do not necessarily translate to better results, especially for short-form videos. This suggests potential overfitting, information redundancy, or an unfavorable cost-benefit ratio regarding computational resources. Crucially, this indicates that effectively tackling our dataset's challenges doesn't require excessive computation. A moderate frame count (32 in this instance) appears to strike an optimal balance, maximizing performance while minimizing computational burden. This underscores the possibility of creating efficient and effective solutions for short-form video analysis without resorting to computationally intensive strategies, and emphasizes the importance of optimizing frame selection based on video characteristics.

B.6. Human performance on HarmonySet-OE

We also evaluated both human and model performance on HarmonySet-OE, shown in Table 8. Due to the complexity of generating open-ended answers manually, we randomly selected 100 questions from HarmonySet-OE and

GPT Prompt for HarmonySet-MC Curation

```
Please design multiple-choice questions based on the following subtitles in four aspects: Rhythm and
Synchronization, Theme and Content, Emotion, and Cultural Relevance. For each aspect, use the given
subtitles as the correct option, and design three other options that are similar but have some errors. The
correct options should contain all information that the subtitles provide.

Note that the correct option can slightly modify the language of the subtitles but should not change the
meaning. Each option should have similar length. Especially in the rhythm section, where the correct option
may contain a timestamp, the confusingly wrong option may also contain some incorrect timestamp or incorrect
description of the correct timestamp to avoid the correct answer being too obvious.

Ensure that the incorrect options are distinguishable from the correct option but not too simple or
completely unrelated. Also, provide the correct answer's letter (A, B, C, or D) randomly positioned among
the options. The correct answer should only include the letter without any other words.

Provided subtitles:\n\n{input_subtitle}\n\n

The output format should be:
Rhythm and Synchronization:
A.<one option>
B.<one option>
C.<one option>
D.<one option>
Correct Answer: <The correct choice such as 'B'>

Theme and Content:
A.<one option>
B.<one option>
C.<one option>
D.<one option>
Correct Answer: <The correct choice such as 'A'>

Emotion:
A.<one option>
B.<one option>
C.<one option>
D.<one option>
Correct Answer: <The correct choice such as 'C'>

Cultural Relevance:
A.<one option>
B.<one option>
C.<one option>
D.<one option>
Correct Answer: <The correct choice such as 'D'>
```

Figure 14. Detailed prompt for generating HarmonySet-MC using GPT4o based on HarmonySet annotations. Each question includes one correct option derived from the dataset and three distractor options designed to be similar in structure, length, and theme, but containing identifiable factual errors.

collected answers from three different annotators per question. Human-generated answers were evaluated using the same methodology applied to the models. We compared human performance against VideoLLaMA2 (HarmonySet), the best-performing model in our main experiment. Results show a noticeable gap between even the best model's performance and human performance, highlighting the limitations of current models in generating open-ended responses. However, the performance gap between humans and models on the OE task is smaller (e.g., in the cultural aspect, the human score is 6.32, while the model score is 4.77) compared to the multiple-choice task (e.g., human accuracy on the cultural aspect is 93.81%, while the best model accuracy is only 50.40%). This smaller gap in open-ended responses might be attributed to the higher cost for humans to produce long-form text, whereas models can achieve higher scores by increasing text richness and length. This suggests that the OE task presents certain challenges even for humans.

Table 9. Full ablation on impact of different training data. Results reveal that training with 10,000 automatically generated annotations provides minimal performance improvement and even hinders performance on Theme and Emotion aspects, suggesting potential inaccuracies or misleading information in the auto-generated data. In contrast, training with HarmonySet data consistently enhances performance, with greater improvements observed with larger training sets. This demonstrates the effectiveness and importance of incorporating human knowledge through HarmonySet.

Models	Metrics	Life & Emotion	Art & Performance	Travel & Events	Sports & Outdoors	Knowledge	Tech & Fashion	Overall
VideoLLaMA2 (Vanilla)	R & S	3.89	4.80	4.56	4.01	3.39	3.54	4.15
	T	4.09	4.83	4.93	3.89	3.44	3.71	4.29
	E	4.36	5.01	5.02	4.08	3.44	3.49	4.38
	C	2.95	3.46	3.69	2.56	2.32	2.52	3.05
VideoLLaMA2 (10k, F.A.)	R & S	4.56	5.05	4.97	4.28	4.20	4.17	4.59
	T	4.20	5.01	4.85	3.49	3.36	3.41	4.16
	E	4.29	4.76	4.67	4.03	3.76	3.79	4.28
	C	3.53	3.98	3.67	2.93	3.13	3.02	3.44
VideoLLaMA2 (10k, HamrmonySet)	R & S	4.69	5.58	5.30	4.49	4.36	4.25	4.86
	T	4.66	5.02	4.98	4.40	4.39	4.29	4.70
	E	4.64	5.43	5.26	4.06	3.85	3.78	4.66
	C	3.99	4.30	4.25	2.97	3.79	3.34	3.89
VideoLLaMA2 (Full HamrmonySet)	R & S	5.43	6.35	6.03	4.94	5.33	4.83	**5.55**
	T	5.12	5.21	5.03	4.84	5.21	4.85	**5.06**
	E	5.25	6.41	5.84	4.00	4.88	4.47	**5.26**
	C	4.87	4.98	4.72	3.31	5.23	4.09	**4.62**

Table 10. Table 6 records the average results across six types of videos. This table presents the detailed results of the frame number ablation experiments. Experiments using VideoLLaMA2 trained with varying numbers of video frames (16, 32, and 64) show optimal performance with 32 frames. Performance degrades with 64 frames, indicating that using too many frames may lead to information redundancy and performance degradation.

Models	Metrics	Life & Emotion	Art & Performance	Travel & Events	Sports & Outdoors	Knowledge	Tech & Fashion	Overall
16 Frames	R & S	5.43	6.35	6.03	4.94	5.33	4.83	5.55
	T	5.12	5.21	5.03	4.84	5.21	4.85	5.06
	E	5.25	6.41	5.84	4.00	4.88	4.47	5.26
	C	4.87	4.98	4.72	3.31	5.23	4.09	4.62
32 Frames	R & S	5.47	6.37	6.07	4.98	5.39	4.87	**5.59**
	T	5.14	5.23	5.08	4.85	5.23	4.86	**5.08**
	E	5.26	6.44	5.88	4.05	4.90	4.50	**5.29**
	C	4.89	5.01	4.76	3.32	5.26	4.14	**4.65**
64 Frames	R & S	5.38	6.31	5.97	4.86	5.27	4.76	5.49
	T	5.03	5.11	4.88	4.72	5.10	4.71	4.94
	E	5.22	6.36	5.80	3.95	4.82	4.40	5.21
	C	4.76	4.89	4.68	3.22	5.10	4.02	4.53