Question-Aware Gaussian Experts for Audio-Visual Question Answering

Supplementary Material

A. Experimental Setup

A.1. Datasets

MUSIC-AVQA.[1] We use the MUSIC-AVQA dataset [2] to train and test our model. This dataset is a comprehensive benchmark specifically designed for audio-visual reasoning tasks. It contains 9,288 music performance videos sourced from YouTube, totaling over 150 hours of content. The dataset features 22 different instruments and includes 45,869 question-answer pairs. The questions are categorized into audio-only, visual-only, and audio-visual types based on 33 templates. These templates cover a range of reasoning categories, such as existence, location, quantity, comparison, and temporal aspects. MUSIC-AVQA excels in challenging models with complex spatio-temporal reasoning, making it a crucial benchmark for evaluating advanced audio-visual understanding.

MUSIC-AVQA-R.[2] We also evaluate our model's performance on MUSIC-AVQA-R [6], an extended version of the original MUSIC-AVQA dataset designed to test the model's robustness. This extension restructures and significantly expands the test set, with a particular emphasis on rare cases and out-of-distribution samples. It includes 211,572 restructured questions, offering a more comprehensive evaluation across a wider range of question types beyond basic template-based questions. This makes MUSIC-AVQA-R a rigorous benchmark for assessing models' capabilities in spatial-temporal reasoning and their ability to manage complex multimodal interactions.

MUSIC-AVQA-v2.0.[3] We assess our model on MUSIC-AVQA-v2.0 [5], an improved version of the original MUSIC-AVQA dataset that addresses data bias issues. This updated dataset offers a more balanced benchmark for audio-visual question answering, containing 53,573 question-answer pairs with a broader range of musical ensembles and more complex audio-visual interactions. To reduce bias, the authors manually curated 1,230 additional musical performance videos from YouTube and created 8,100 new QA pairs to supplement the original dataset. These updates ensure more balanced answers across various question templates.

A.2. Implementation Details

Disjoint-Centered Gaussian Experts. To reduce overlaps in the regions of influence among multiple Gaussian distributions, we initialize the center positions $\mathbf{u}_{\text{fixed}}$ of the E Gaussian experts with a predefined margin between them (Algorithm 1). While minor overlap may occur due to the Gaussian widths, the centers remain non-overlapping, reducing redundant temporal influences. To refine temporal segments, learnable offsets adjust the fixed positions, keeping centers within constrained margins. This ensures each expert focuses on distinct temporal ranges, capturing question-relevant segments more effectively while minimizing redundancy and maintaining expert specialization.

Patch Merging. To ensure a fair comparison with TSPM [4], we utilize the patch merging strategy from ToMe [1], which merges similar visual tokens within each transformer block. This involves dividing tokens into subsets, calculating similarities, and applying mean fusion to generate merged token features. By adopting this method, we align the feature extraction pipeline with TSPM for consistency in the experimental setup. For more detail on patch merging, please refer to ToMe[4] and TSPM[5].

Method	CLIP Encoder	A-QA	V-QA	AV-QA	Avg
PSTP-Net [3]	B/32	70.91	77.26	72.57	73.52
TSPM [4]	B/32	**76.91**	81.92	72.57	75.81
QA-TIGER	B/32	76.66	**83.69**	**72.61**	**76.26**
PSTP-Net [3]	L/14	73.87	79.19	71.76	74.10
TSPM [4]	L/14	76.91	83.61	73.51	76.79
QA-TIGER	L/14	**78.58**	**85.14**	**73.74**	**77.62**

Table A. Results for different Encoders (CLIP-B/32 and CLIP-L/14) used for both visual and textual feature extraction.

B. More Experimental Results

Since some prior studies evaluate the model with CLIP-B/32 encoder, we evaluate our method under the same setting to enable direct comparison. Previous studies, such as PSTP-Net [3] and TSPM [4], have shown that performance improves when transitioning from CLIP-B/32 to a more advanced feature extractor like CLIP-L/14. However, using CLIP-B/32 remains a relevant benchmark for assessing the baseline performance of methods. As shown in Table A, our method consistently outperforms PSTP-Net and TSPM, even with the smaller CLIP-B/32 encoder. This demonstrates the robustness of our approach, highlighting that its effectiveness is not solely reliant on high-capacity encoders.

[1] https://github.com/GeWu-Lab/MUSIC-AVQA
[2] https://github.com/reml-group/MUSIC-AVQA-R
[3] https://github.com/DragonLiu1995/MUSIC-AVQA-v2.0
[4] https://github.com/facebookresearch/ToMe
[5] https://github.com/GeWu-Lab/TSPM

Algorithm 1 Gaussian Experts Module

Input: Sentence-level question features: $q_s \in \mathbb{R}^D$. Video/Audio features: $\mathbf{v}_q, \mathbf{a}_q \in \mathbb{R}^{T\times D}$. Visual/Audio-related patch features : $\mathbf{p}_v, \mathbf{p}_a \in \mathbb{R}^{T\times D}$.
Output: Temporal integrated features: Aggregated visual-related patch features $\tilde{v}_{p_v} \in \mathbb{R}^D$, aggregated audio-related patch features $\tilde{v}_{p_a} \in \mathbb{R}^D$. Aggregated audio features $\tilde{a} \in \mathbb{R}^D$.

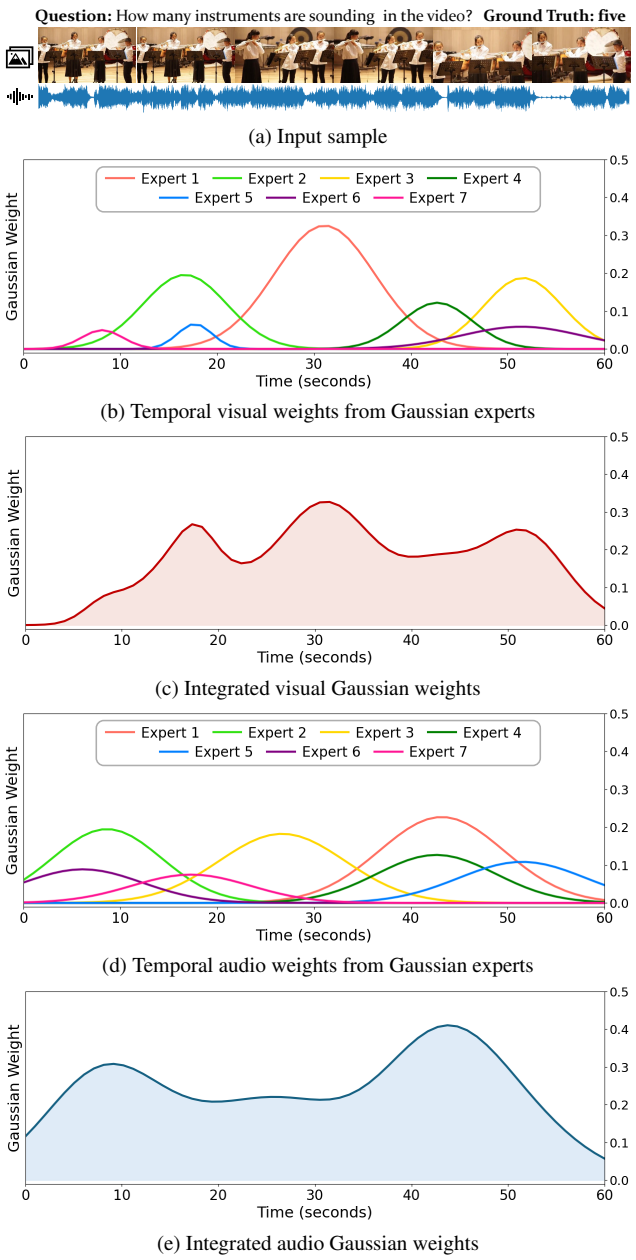
Initialization: Initialize the center of E experts to the central positions of E segments.
margin $\leftarrow \frac{1}{2\times E}$ ▷ Margin between Gaussian centers
$\mathbf{u}_{\text{fixed}} \leftarrow \left[\text{margin} + i \cdot \frac{1-2\cdot\text{margin}}{E-1} \text{ for } i = 0 \text{ to } E-1\right]$

1. Question-Guided Attention:
$\mathbf{v}'_q,\ \mathbf{a}'_q \leftarrow \text{CA}(q_s, \mathbf{v}_q, \mathbf{v}_q),\ \text{CA}(q_s, \mathbf{a}_q, \mathbf{a}_q)$

2. Calculate Experts Probability:
$\mathbf{r}_v, \mathbf{r}_a \leftarrow \text{Softmax}(\text{Router}(\mathbf{v}'_q)), \text{Softmax}(\text{Router}(\mathbf{a}'_q))$

3. Gaussian Weight Generation:
$u_{\text{offset}|v}, \sigma_v \leftarrow$ Gaussian Generator$(\mathbf{v}'_q)$
$u_{\text{offset}|a}, \sigma_a \leftarrow$ Gaussian Generator$(\mathbf{a}'_q)$
Adjust centers and normalize widths:
$u^i_v \leftarrow u^i_{\text{fixed}} + \text{Tanh}(u^i_{\text{offset}|v}) \cdot \text{margin}$
$u^i_a \leftarrow u^i_{\text{fixed}} + \text{Tanh}(u^i_{\text{offset}|a}) \cdot \text{margin}$
$\sigma^i_v,\ \sigma^i_a \leftarrow \text{Sigmoid}(\sigma^i_v),\ \text{Sigmoid}(\sigma^i_a)$
Generate temporal Gaussian weights:
for $i \leftarrow 1$ to E **do**
$\mathbf{g}_v[i],\ \mathbf{g}_a[i] \leftarrow \mathcal{N}(\mathbf{u}^i_v, (\sigma^i_v)^2),\ \mathcal{N}(\mathbf{u}^i_a, (\sigma^i_a)^2)$
$\mathbf{g}_v[i],\ \mathbf{g}_a[i] \leftarrow \frac{\mathbf{g}_v[i]}{\max(\mathbf{g}_v[i])},\ \frac{\mathbf{g}_a[i]}{\max(\mathbf{g}_a[i])}$
end for

4. Integration of Experts Output:
$\mathbf{p}_v = \mathbf{v}_q + \mathbf{p}_v$
$\mathbf{p}_a = \mathbf{v}_q + \mathbf{p}_a$
$\tilde{a} \leftarrow \sum_{i=1}^{E} \mathbf{g}^i_a \mathbf{r}^i_a \mathcal{E}^i(\mathbf{a}_q)$
$\tilde{v}_{p_v}, \tilde{v}_{p_a} \leftarrow \sum_{i=1}^{E} \mathbf{g}^i_v \mathbf{r}^i_v \mathcal{E}^i(\mathbf{p}_v), \sum_{i=1}^{E} \mathbf{g}^i_v \mathbf{r}^i_v \mathcal{E}^i(\mathbf{p}_a)$
return $\tilde{v}_{p_v}, \tilde{v}_{p_a}, \tilde{a}$



(a) Input sample
(b) Temporal visual weights from Gaussian experts
(c) Integrated visual Gaussian weights
(d) Temporal audio weights from Gaussian experts
(e) Integrated audio Gaussian weights

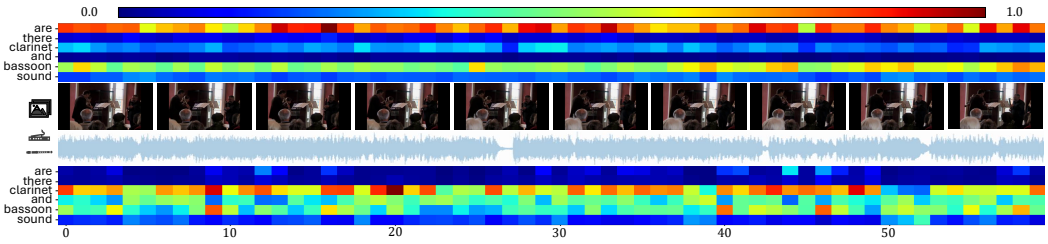
Figure A. Visualization of temporal weights from Gaussian experts for visual and audio modalities, integrated to focus on question-relevant frames for accurate predictions.

C. In-Depth Visualization of QA-TIGER

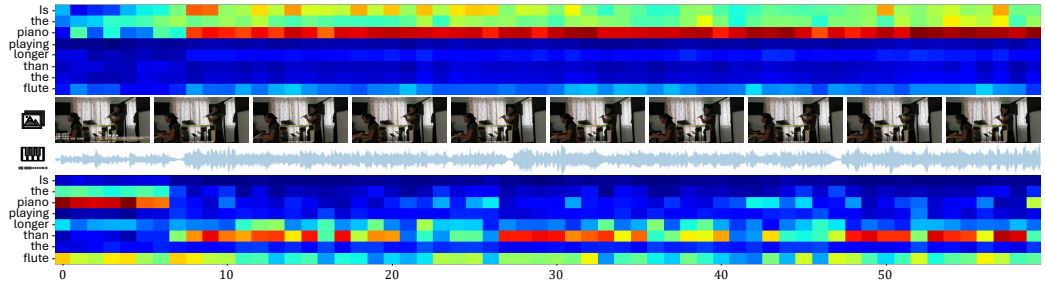
We demonstrate how QA-TIGER dynamically integrates temporal and multimodal information through Gaussian experts and question-aware fusion across diverse scenarios.

C.1. Temporal Integration of Gaussian Experts

Gaussian experts adaptively focus on specific temporal regions based on the input and question. Figure A illustrates the temporal weights generated by the seven Gaussian experts employed in the model (see Figures Ab and Ad). These graphs show how the weights are combined to emphasize the model's focus on distinct temporal segments across visual and audio modalities (Figures Ad and Ae). Each Gaussian curve demonstrates that the experts specialize in specific, minimally overlapping temporal regions, dynamically adjusting their focus based on the question and modality. By integrating these Gaussian weights, the model selectively attends to different frames according to the assigned expert and modality, even when processing the same question. This frame-level and modality-specific information is then utilized in the question-guided reasoning stage.

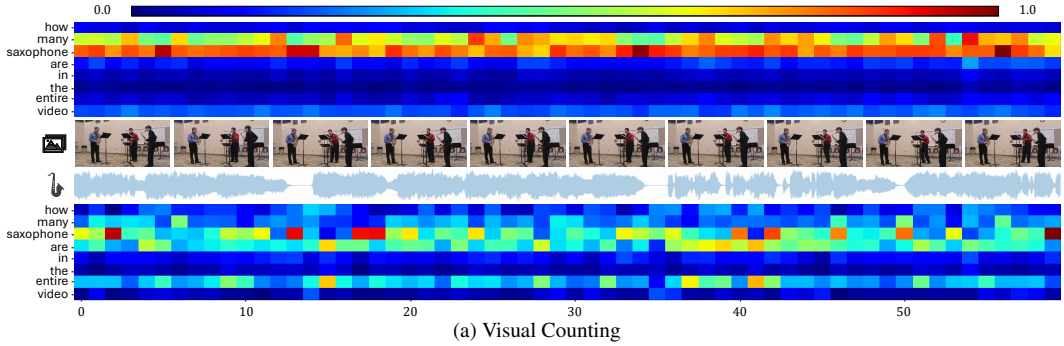


(a) Audio Counting

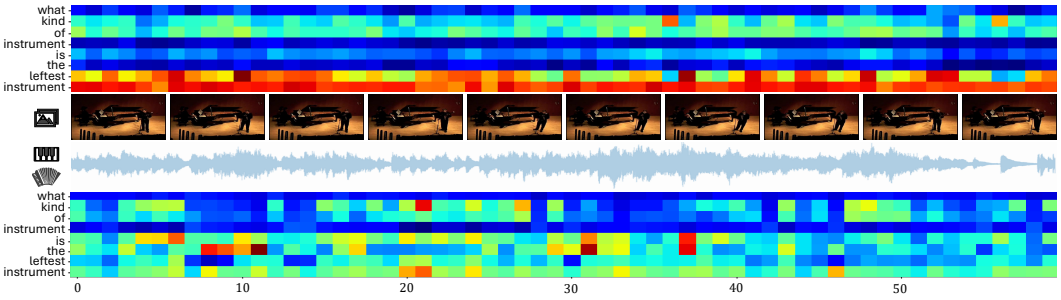


(b) Audio Comparative

Figure B. Valid attention visualization for audio questions.



(a) Visual Counting



(b) Visual Location

Figure C. Valid attention visualization for visual questions.

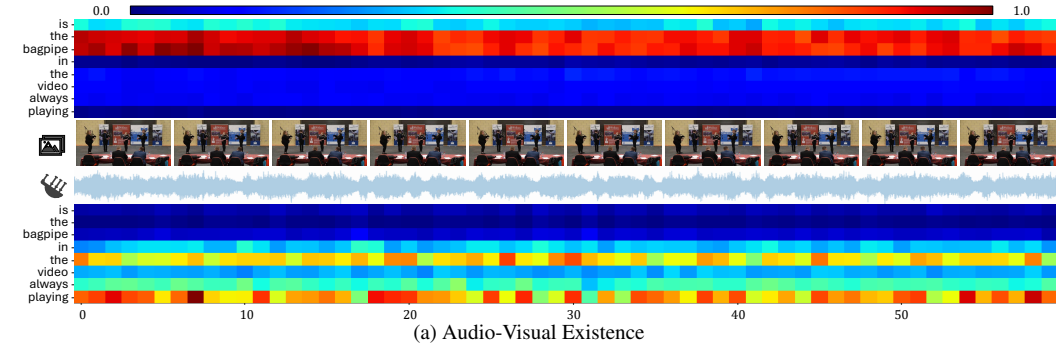
C.2. Visualization of Question-Aware Fusion

We visualize the attention of the question-aware fusion module across nine question types to highlight the effectiveness of our method. These nine question types are categorized into three main groups: Audio-QA (A-QA), Visual-QA (V-QA), and Audio-Visual-QA (AV-QA). A-QA focuses solely on auditory cues, V-QA relies exclusively on visual information, while AV-QA requires the integration of both modalities to address multi-modal questions effectively. Note that attention is presented in two parts: visual modality (top) and audio modality (bottom).

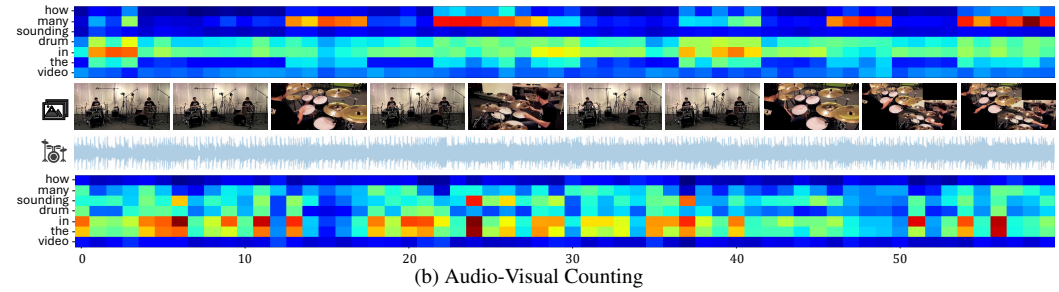
C.2.1. Valid Cases

***Audio* Counting.** In the visual modality, attention focuses on the "bassoon" with two players clearly visible, while the "clarinet" is partially obscured (Figure Ba). The audio modality highlights the "clarinet", ensuring recognition of both instruments. Interestingly, the word "are" also gains attention, likely due to its role in framing the question involving counting and the presence of instruments.

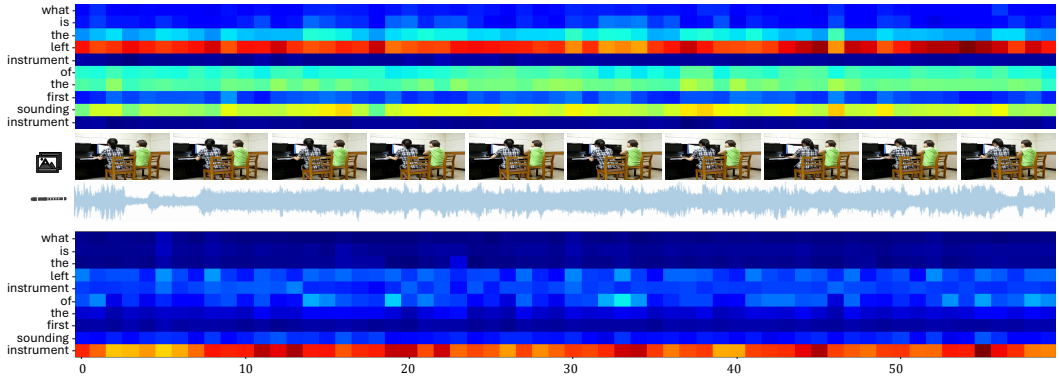
***Audio* Comparative.** The attention mechanism dynamically shifts between modalities to adapt to changing contexts. Initially, the audio attention focuses on the "piano"



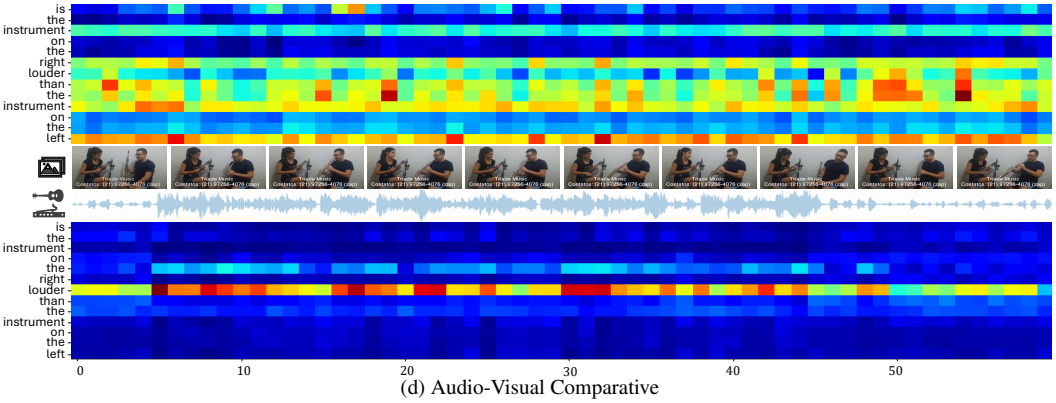
(a) Audio-Visual Existence



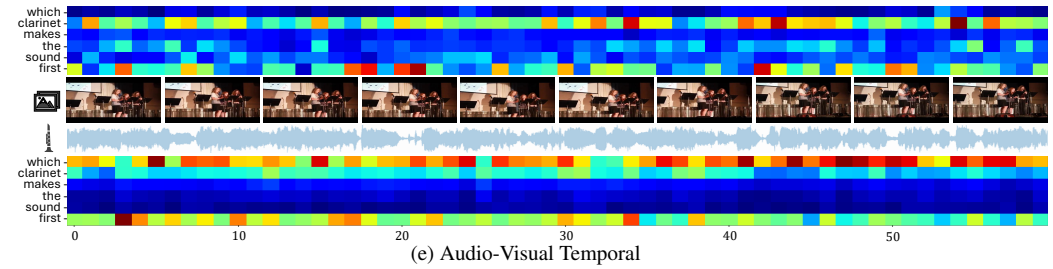
(b) Audio-Visual Counting



(c) Audio-Visual Location



(d) Audio-Visual Comparative



(e) Audio-Visual Temporal

Figure D. Valid attention visualization for audio-visual type.

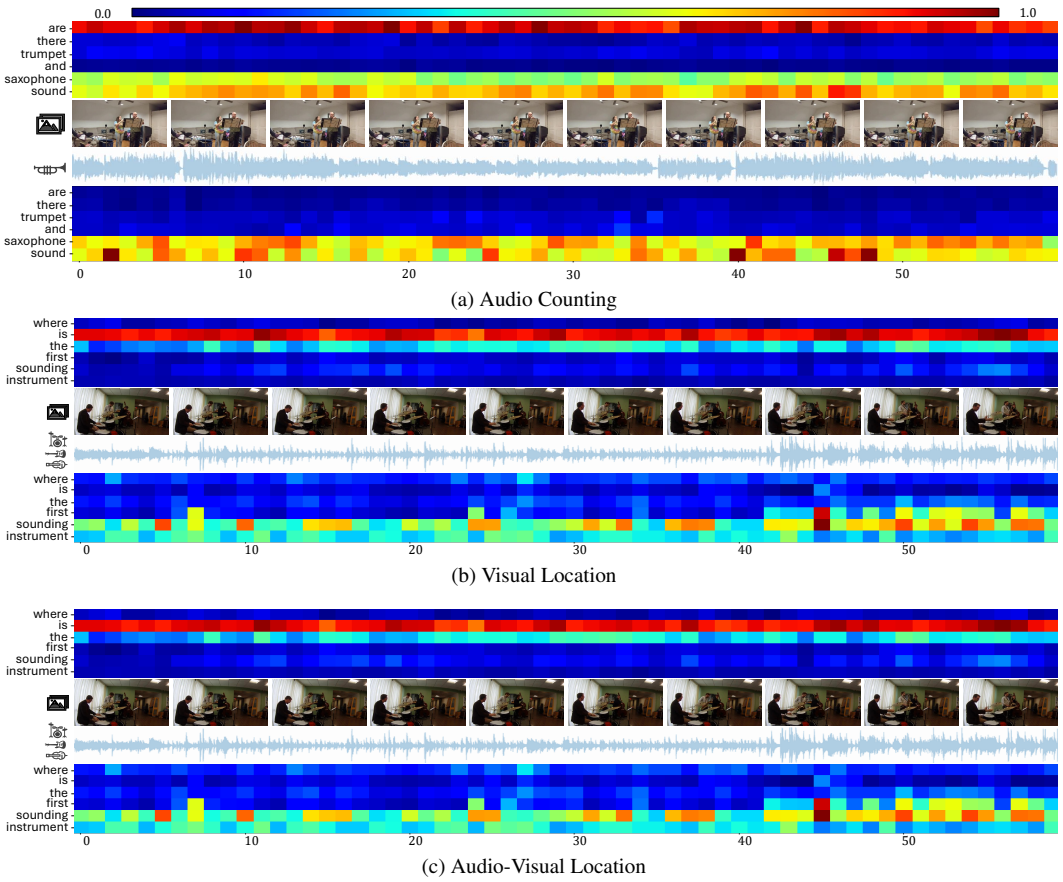


Figure E. Attention visualization in failure cases.

due to its prominent sound. As the "flute" becomes more dominant, the visual attention compensates by identifying that the "piano" continues to be played, even though its audio presence has diminished. This interaction is illustrated in Figure Bb.

***Visual* Counting.** In the visual modality, attention focuses on "many" and "saxophone," leveraging visual cues to estimate the number of instruments. Meanwhile, the audio modality complements this by highlighting the distinct auditory features of saxophone sounds, helping to identify and differentiate instances throughout the video in Figure Ca.

***Visual* Location.** The fusion module focuses on "kind of" and "leftest instrument" in the visual modality, using spatial cues to locate the leftmost instrument. In contrast, the audio modality emphasizes "kind of" and "instrument" to classify its type. Together, these modalities effectively balance spatial and categorical aspects as illustrated in Figure Cb.

***Audio-Visual* Existential.** When observing the entire video from a distance, the visual modality primarily focuses on the "bagpipe", determining whether it is consistently visible throughout the scene. Meanwhile, the audio modality emphasizes the words "always" and "playing", assessing whether the bagpipe consistently produces sound. This interplay between the modalities is illustrated in Figure Da.

***Audio-Visual* Counting.** The fusion module adapts its attention to focus on counting-related cues across modalities. In the visual modality, attention emphasizes "drum" and the word "many" through close-up images, enabling accurate counting of the drums. At the same time, it highlights "sounding" and "drum" in the audio modality to distinguish individual drum sounds as depicted in Figure Db.

***Audio-Visual* Location.** In this synthetic video, with only the flute sound present, the model uses spatial and auditory cues. Visually, it focuses on "left" and "sounding" to locate the instrument, while auditorily, it emphasizes "instrument" to classify its type, as shown in Figure Dc.

***Audio-Visual* Comparative.** The module focuses on "instrument," "right," "left," and "louder" to identify spatial locations in visual modality. Meanwhile, it emphasizes "louder" to analyze sound intensities in audio modality. This complementary approach enables the model to tackle the question effectively, as illustrated in Figure Dd.

***Audio-Visual* Temporal.** The attention module balances visual and auditory cues to identify the specific clarinet that produces the first sound. In the visual modality, attention focuses on "clarinet" and "first," using motion cues to detect active clarinets. To compensate for visually occluded clarinets, the audio modality emphasizes "which," "clarinet," and "first," helping to highlight the source of the initial sound in Figure De.

C.2.2. Failure Cases

***Audio* Counting.** The attention incorrectly highlights "saxophone" and related auditory features instead of "trumpet," as shown in Figure Ea. Given that only trumpet sounds are present, this misclassification likely stems from the model confusing the trumpet sound with the similar auditory characteristics of a saxophone. Visually, the model also fails to correctly identify "trumpet," possibly due to one being partially obscured and the other blending into the background because of similar coloring with the performers' clothing. This suggests that the model overly relies on auditory cues when visual distinctions are less prominent, leading to confusion between visually and aurally similar objects.

***Visual* Location.** The failure comes from limitations in both modalities. Visually, the absence of the "bagpipe" forces reliance on auditory cues. However, with no "bagpipe" sound present, overlapping flute and bassoon sounds may have been misclassified, as shown in Figure Eb. This highlights the challenge of distinguishing similar-sounding instruments in multi-modal reasoning. The issue likely stems from the model's difficulty in separating distinct auditory features when instrument sounds overlap, compounded by the lack of visual confirmation.

***Audio-Visual* Location.** Auditorily, the module captures sound-related cues effectively. Visually, attention is drawn to less critical words like "is," which provide some contextual relevance. However, this focus reduces the emphasis on essential keywords such as "where" and "first," which are crucial for temporal understanding, as shown in Figure Ec. Such attention patterns suggest that the model may have overemphasized certain contextual cues while underutilizing spatial and temporal keywords, leading to an incorrect prediction. This indicates a need for better balancing of context and question-specific focus.

C.3. Qualitative Comparison of Temporal Gaussian

This section focuses on two main points: (i) For all nine question types, we demonstrate that the proposed temporal Gaussian approach outperforms conventional sampling methods, such as uniform sampling and Top-K frame selection, by efficiently utilizing the entire temporal sequence and focusing on critical segments in Figure F, G and H. (ii) In Figure I, we examine cases where the proposed method underperforms compared to other sampling techniques, offering insights into areas for future improvement. For comparison, we use ST-AVQA [2] for the uniform sampling and TSPM [4] for the Top-K frame selection.

C.3.1. Valid Cases

***Audio* Counting.** For the question "How many musical instruments were heard throughout the video?" in Figure Fa, the Top-K method focuses on selecting frames where musical instruments are most visually prominent. However, key is to identify segments where audio signals from all playing instruments are strongest. QA-TIGER's audio Gaussian effectively captures these moments, enabling it to count the number of instruments played accurately.

***Audio* Comparative.** For a question like "Is the piano playing longer than the violin?" in Figure Fb, the Top-K approach focuses only on frames that include the piano and violin. However, since it considers only a limited number of N frames, it struggles to make accurate comparisons when the question requires analyzing the entire temporal span, such as for "longer." Uniform sampling also fails to consider the entire sequence, making it challenging to derive accurate answers. In contrast, QA-TIGER applies adaptive weights across the entire sequence, allowing for more efficient and accurate comparisons.

***Visual* Counting.** QA-TIGER predicts the number of cellos in a video by analyzing both close-up and full-shot scenes of cello performances, as illustrated in Figure Fc. In comparison, the uniform sampling may occasionally include frames where all cellos are visible, resulting in a correct prediction. However, if the frame order changes, the prediction could easily be wrong. Meanwhile, the Top-K method focuses only on the close-up frames where cellos are most visible. As a result, it predicts only "three" cellos, even though "four" are actually being played.

***Visual* Location.** The Top-K approach primarily focuses on the "violin" itself, selecting frames that are strongly related to the violin but overlooking its right-hand side (Figure Ga). While it does select one frame with clues about the instrument to the right of the violin, a close-up of the piano shifts the focus away from this information. In contrast, QA-TIGER effectively concentrates on temporal segments that emphasize both the violin and its right-hand side. It assigns the lowest weight to the piano, ensuring that the most important details are prioritized.

***Audio-Visual* Existential.** In Figure Gb "Is the flute in the video always playing?" highlights a limitation of the Top-K method, which selects only frames where the flute is both present and actively being played. QA-TIGER takes a more balanced approach by allowing the audio Gaussian to focus on moments when the flute is playing, while the visual Gaussian also considers frames where the flute is not being played. This demonstrates that our method enables each

modality to independently emphasize different aspects of the question in a complementary manner.

***Audio-Visual* Counting.** QA-TIGER focuses on the segments where instruments are being played while broadly considering the overall context, as shown in Figure Gc. On the other hand, as observed in other question types, Top-K approach often misses other critical information since it selects only a limited number of frames centered around the specific instrument mentioned in the question. Uniform sampling exclusively uses frames containing only the two individuals playing instruments, leading to the limitation of predicting "two" instead of the correct answer, "four."

***Audio-Visual* Location.** In Figure Ha, QA-TIGER's audio Gaussian assigns relatively higher weights to the early parts of the sequence, effectively identifying the first instrument played. Meanwhile, the visual Gaussian focuses on wide shots where the positions of all instruments are visible, contributing to accurately answering the instruments and their locations. However, the uniform sampling approach fails to capture which instrument was played first, and the Top-K method only considers frames with the most prominent close-up of an instrument, limiting its ability to determine the precise location of the instruments.

***Audio-Visual* Comparative.** When comparing audio-visual content as in Figure Hb, the original video is too short, so the audio for the remaining frames was generated by repeating the last 1-second segment of the video. QA-TIGER accurately focuses only on the relevant segments of the original video across all modalities, enabling precise answer predictions. In contrast, the Top-K method incorrectly focuses on unrelated frames, resulting in wrong answers when paired with uniform sampling.

***Audio-Visual* Temporal.** For the question about which congas are played first, as in Figure Hc, the Top-K approach focuses on frames highlighting the congas in the early part of the sequence. Unfortunately, due to the limited number of frames, it fails to utilize information about other congas in the later part, leading to an incorrect answer. In contrast, our method effectively reflects the intent of the question by assigning higher weights to the early part through the audio Gaussian, while the visual Gaussian captures critical information such as the total number and locations of congas in the later part, resulting in the correct answer.

C.3.2. Failure Cases

***Visual* Counting.** While QA-TIGER focuses on frames with multiple instruments, it could miss finer details, as shown in Figure Ia. In the case of the given sample, where "ukulele" and "violin" coexist, QA-TIGER could struggle to accurately identify each instrument's appearance or distinguish between them if the audio of one instrument is overshadowed or sounds similar to the other.

***Audio-Visual* Existential.** In questions like Figure Ib, "Is the trumpet in the video always playing?", QA-TIGER's audio Gaussian effectively focuses on the portions of the audio signal where the trumpet sound is most prominent. However, in the final 5 seconds, the loud cheering from the audience overwhelms the trumpet sound, causing the model to misidentify it and give an incorrect prediction.

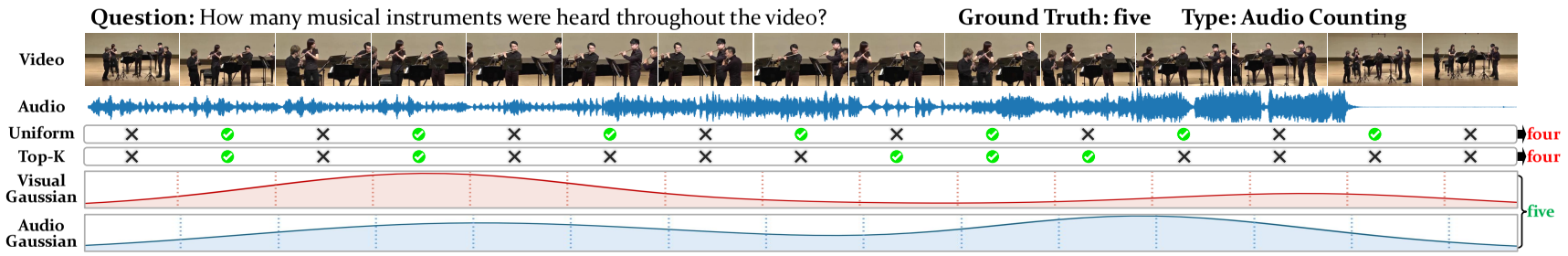
***Audio-Visual* Counting.** In cases like Figure Ic, the visual Gaussian effectively focuses on the early temporal segment where the ukulele is played, while the audio Gaussian prioritizes the acoustic guitar, which has a stronger audio signal but a similar sound to the ukulele. Although the weight is lower, QA-TIGER's consideration of the entire sequence allows it to include the ukulele sound from the early segment. This enables the model to correctly answer "two" for the question, "How many sounding ukuleles are in the video?"

Overall, these cases suggest opportunities to enhance QA-TIGER, such as implementing adaptive mechanisms to dynamically adjust the number of experts based on content, improving its ability to capture varying temporal complexities. Additionally, addressing external noise can further refine its performance across diverse scenarios.

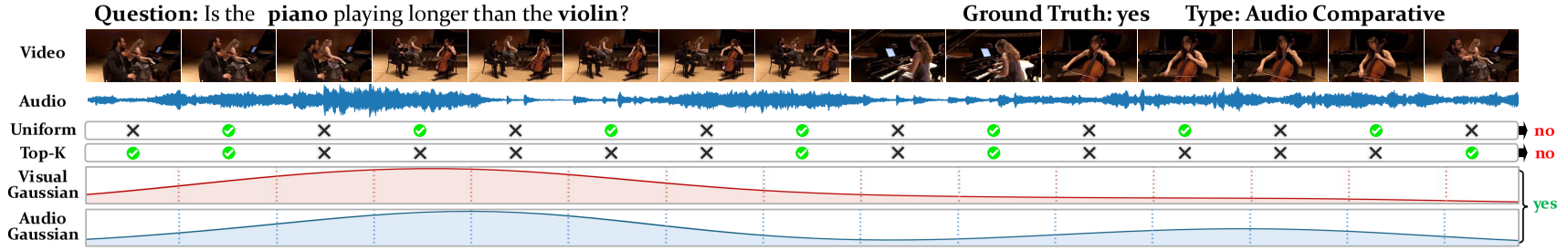
D. Discussion & Future Work

We provide supplementary material to complement the main paper by detailing experimental setups, additional results, and visualizations of QA-TIGER's mechanisms. QA-TIGER achieves state-of-the-art performance, leveraging Gaussian experts for precise temporal integration and effective alignment of question-specific audio-visual features with minimal redundancy. Compared to prior methods, it demonstrates superior accuracy in complex reasoning tasks, including temporal and comparative queries, while maintaining computational efficiency. In addition to quantitative improvements, visualizations show how QA-TIGER dynamically adjusts its attention across audio and visual modalities, effectively handling diverse question types.

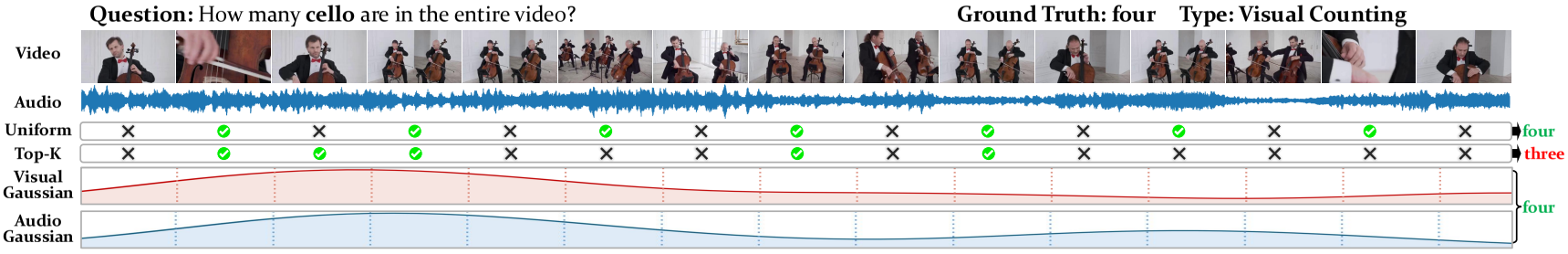
While QA-TIGER shows promising results both quantitatively and qualitatively, we aim to further enhance its adaptability and flexibility. The current Mixture of Experts (MoE) framework relies on a fixed number of experts, which may not fully capture the varying temporal complexities present in different audio-visual content. In this regard, it will be promising to develop adaptive mechanisms that dynamically adjust the number of experts based on the multimodal content, enabling the model to better represent and model varying temporal dynamics. Additionally, since AVQA models are often constrained to predefined answers, the integration of large language models into QA-TIGER can be investigated in the future. This integration allows for more flexible and natural language responses, broadening its applicability to more complex and diverse scenarios.



(a) Audio Counting

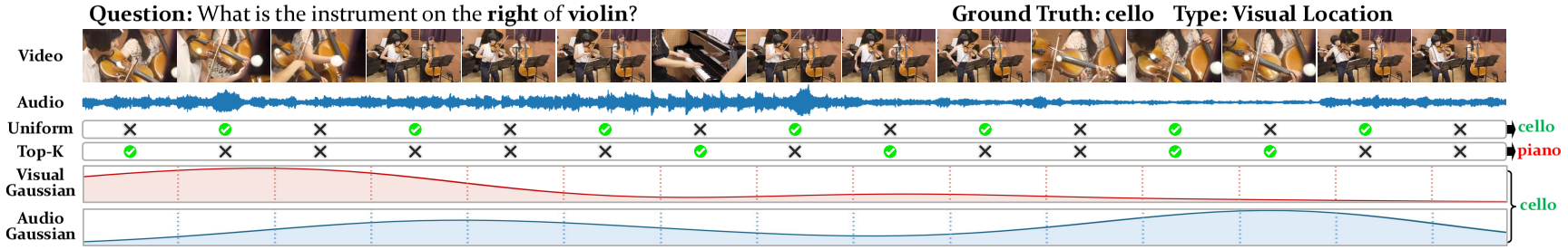


(b) Audio Comparative

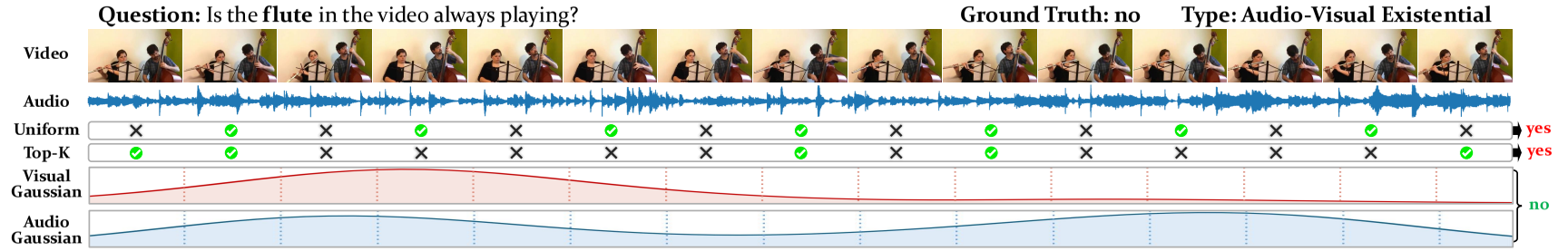


(c) Visual Counting

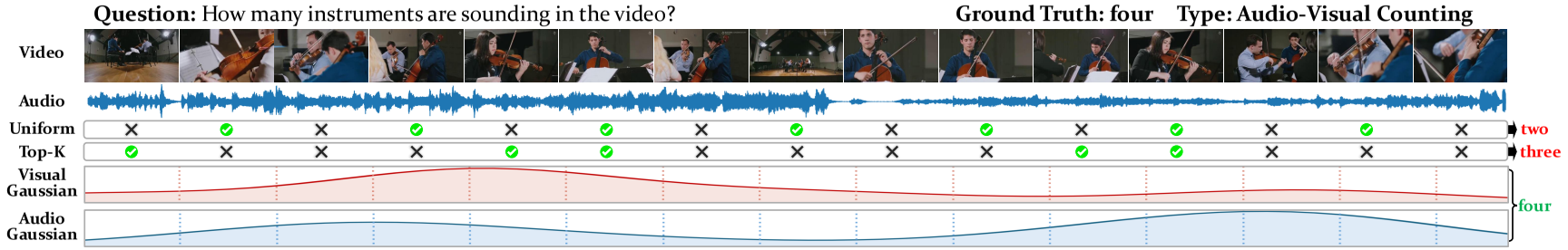
Figure F. Valid qualitative comparison with Uniform sampling and Top-K frame selection.



(a) Visual Location

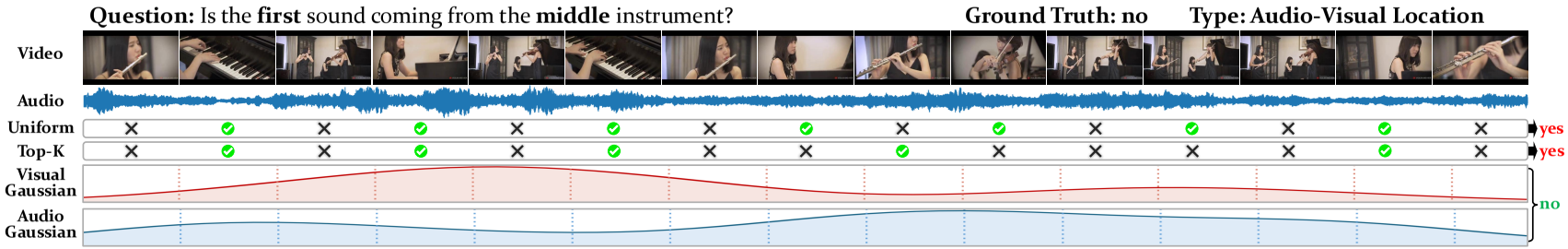


(b) Audio-Visual Existential

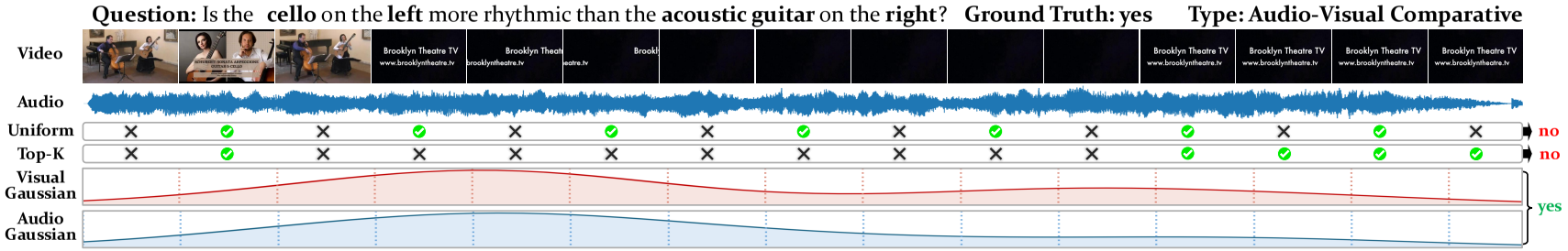


(c) Audio-Visual Counting

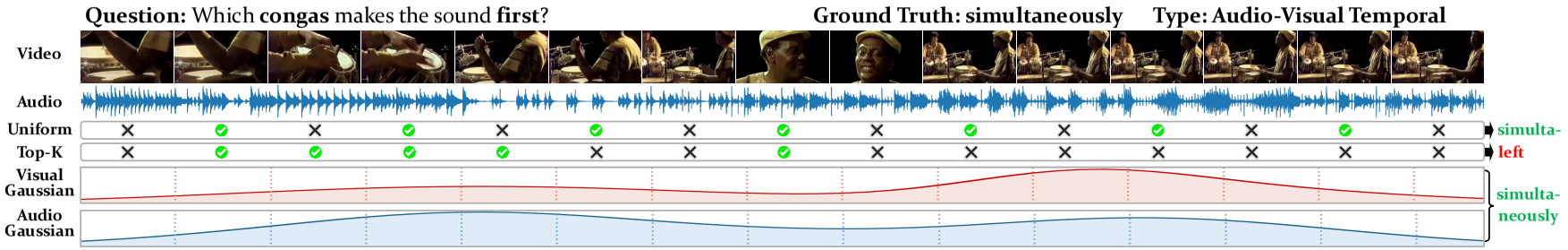
Figure G. Valid qualitative comparison with Uniform sampling and Top-K frame selection.



(a) Audio-Visual Location

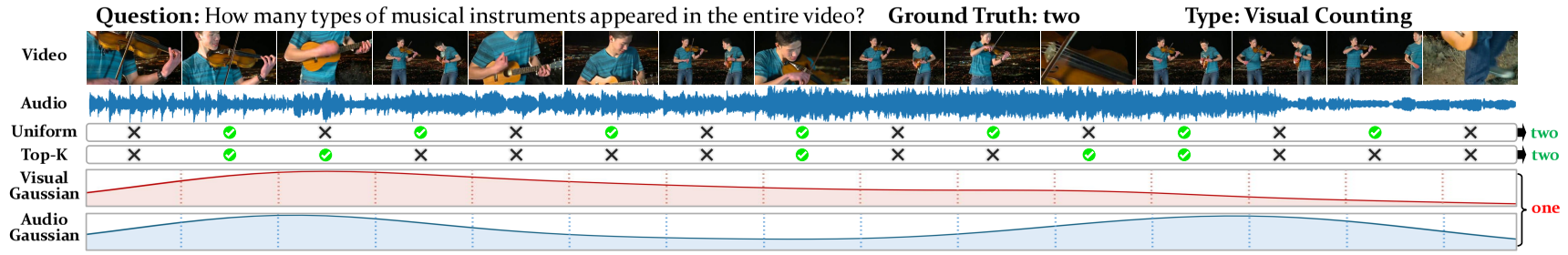


(b) Audio-Visual Comparative

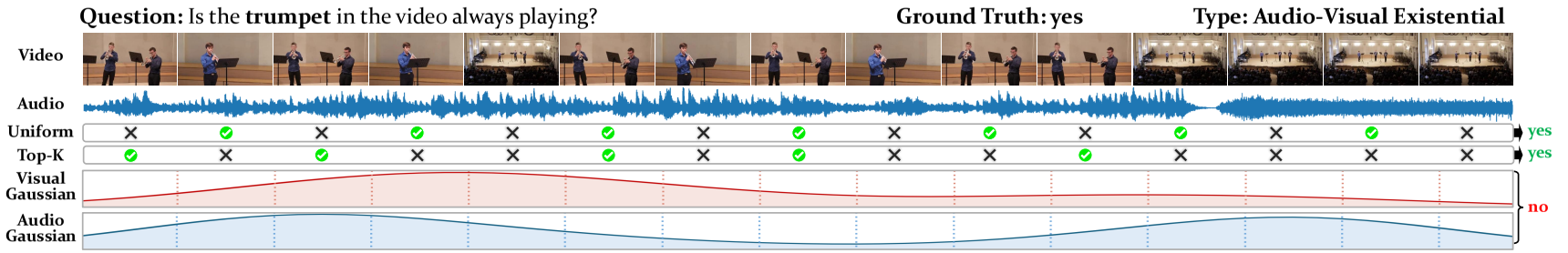


(c) Audio-Visual Temporal

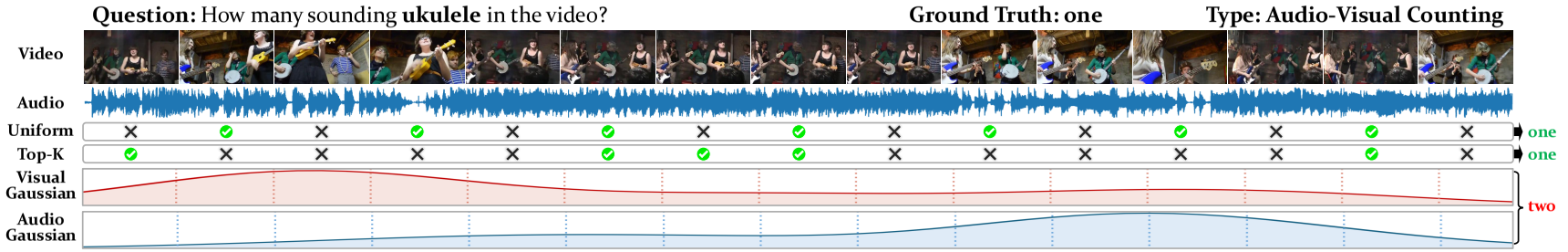
Figure H. Valid qualitative comparison with Uniform sampling and Top-K frame selection.



(a) Visual Counting



(b) Audio-Visual Existential



(c) Audio-Visual Counting

Figure 1. Failure qualitative comparison with Uniform sampling and Top-K frame selection.

References

[1] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *ICLR*, 2023.

[2] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *CVPR*, pages 19108–19118, 2022.

[3] Guangyao Li, Wenxuan Hou, and Di Hu. Progressive spatio-temporal perception for audio-visual question answering. In *ACM MM*, pages 7808–7816, 2023.

[4] Guangyao Li, Henghui Du, and Di Hu. Boosting audio visual question answering via key semantic-aware cues. In *ACM MM*, 2024.

[5] Xiulong Liu, Zhikang Dong, and Peng Zhang. Tackling data bias in music-avqa: Crafting a balanced dataset for unbiased question-answering. In *WACV*, pages 4478–4487, 2024.

[6] Jie Ma, Min Hu, Pinghui Wang, Wangchun Sun, Lingyun Song, Hongbin Pei, Jun Liu, and Youtian Du. Look, listen, and answer: Overcoming biases for audio-visual question answering. In *NeurIPS*, 2024.