Learning to Answer Questions in Dynamic Audio-Visual Scenarios (Supplementary Material)

Guangyao Li[1,†], **Yake Wei**[1,†], **Yapeng Tian**[3,†], **Chenliang Xu**[3], **Ji-Rong Wen**[1], **Di Hu**[1,2,*]

[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing
[2]Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing
[3]Department of Computer Science, University of Rochester, Rochester
[1]{guangyaoli, yakewei, jrwen, dihu}@ruc.edu.cn, [3]{yapengtian, chenliang.xu}@rochester.edu

Contents

1. Supplementary Video

In our demo video, we will provide video examples with sounds in our MUSIC-AVQA dataset and audio-visual question answering results. For more details, please check the demo.

2. Videos Collection

In this section, We introduce the details of MUSIC-AVQA dataset construction. According to *Wikipedia*, 22 kinds of instruments shown in Tab. 1 are divided into 4 categories: *String*, *Wind*, *Percussion* and *Keyboard*.

Table 1. Musical Instrument Classification

String	Wind	Percussion	Keyboard
violin	tuba	drum	accordion
cello	trumpet	xylophone	piano
guitar	suona	congas	
ukulele	bassoon		
erhu	clarinet		
guzheng	bagpipe		
pipa	flute		
bass	saxophone		
banjo			

2.1. Real Videos

In the MUSIC-AVQA dataset, three kinds of musical performance are covered to ensure the diversity, complexity and dynamic of audio-visual scenes: solo, ensemble of the same instrument (ESIT) and ensemble of different instruments (EDIT). The rule of EDIT is that each instrument is required to combine with one or more instruments in different categories. Specifically, we use permutation and combination methods for 22 instruments to ensure that all instrument combinations can be covered in the video as much as possible. For the duet case in EDIT, we consider all the combinations of 2 different categories of 22 instruments, which accordingly becomes a total of C_{22}^2 combinations. We search for related videos on YouTube according to these combinations styles. Meanwhile, for other ensemble forms in EDIT, such as trio, quartet, etc., we consider more than 2 different instrument combinations and retrieve related videos on YouTube.

In Fig. 1, we show the number of the combination of every two different instruments in the real video, counted from not only the duet video, but also the trio, quartet etc. The categories of musical instruments appearing in some videos are not in the 22 musical instruments and which are represented by *other*. As shown in Fig. 1, some instruments tend

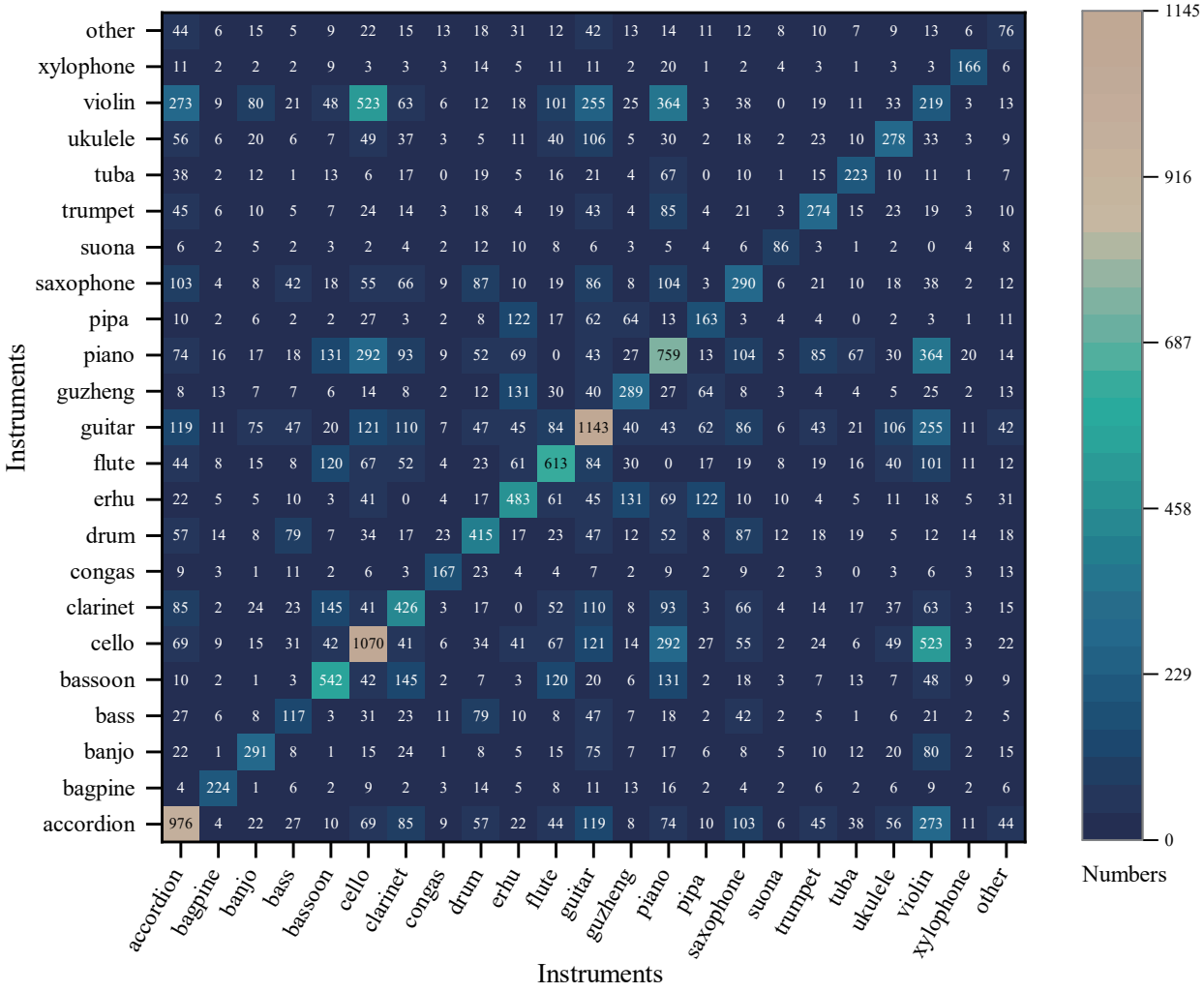


Figure 1. Number of combinations of different types of instruments, where the lighter the color, the more the number. And instruments outside the 22 instrument categories are denoted by *other*. The confusion matrix shows that the combination of different instruments is diversified.

to combine with some other instruments due to their coordination in music, such as *cello* and *violin* etc. Even though, we still do our best to find almost all kinds of combination of different instruments. These statistical results illustrate the diversity of the collected videos.

2.2. Synthetic Videos.

To further facilitate study on understanding and reasoning over complex multimodal scenes, we synthesize more challenging videos in which multiple visual objects and sounds are appeared with different associations.

For videos synthesized using solo scenes, we retrieve about another 1,500 videos from YouTube, w.r.t. above 22 instrument categories, and they are not included in the collected solo videos in Sec. 2.1 above.

Additionally, the number of solo videos for each instrument is between 50-80, and all the 1500 videos are randomly cut into 1 minute long. For simplicity, the cutted video is denoted as D. Then, we randomly select 750 videos from D and separate the sound track from them. The separated video (silent) and audio are represented by D_V and D_A, respectively. After that, we divide D into two types: M and N, where M contains D_V and D_A, and the rest videos in D except M is represented by N. Finally, we synthesize videos in the following three ways.

1) Audio overlay. We randomly select 500 audios and videos from D_A and N, respectively. Then we randomly select one audio and overlay it to one video, which generate one video contain single instrument in vision but with two instrument sounds.

2) Video stitching. We randomly select two different real videos then spatially stitch them into one video. Specially, we select 500 videos from D_A and N, respectively. Then these two different types of videos are randomly

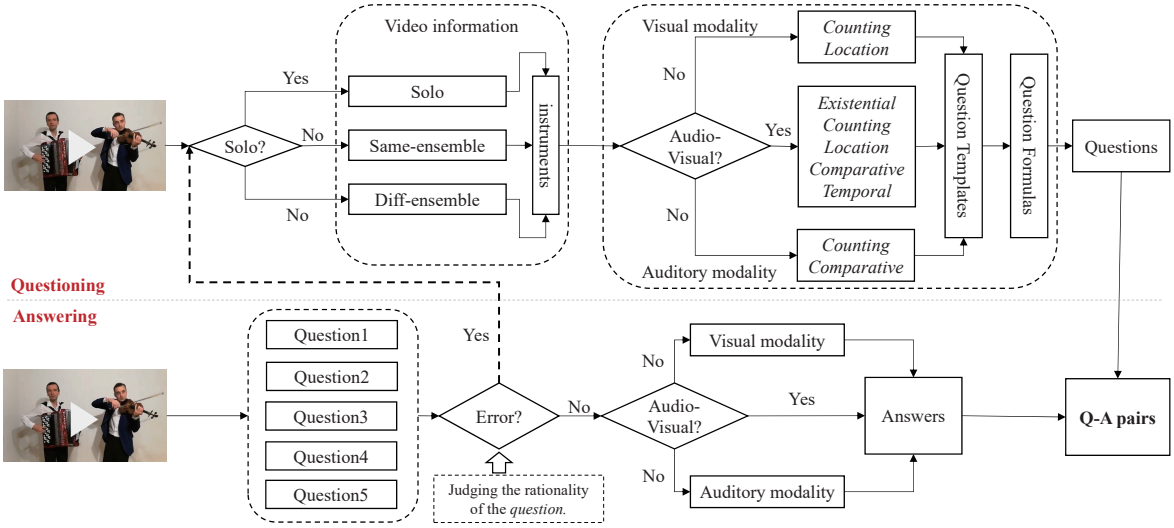


Figure 2. Labeling system contains *questioning* and *answering*. In the *questioning* section, the annotator is required to select the performance type of the video and the included instruments, and then *scene types*, *question types*, and *question templates*, and finally one *question* is automatically generated based on the previous selection. In the *answering* part, the annotator to judge whether the *question* is reasonable, and if it is unreasonable, the *question* will be labeled again. Then, the annotator answering the *question* according to video content, and finally one QA pair is produced.

stitched horizontally into one video, so that one video will contain the left and right instrument performance, but only one of them has sound.

3) Audio and video random matching. We replace the original sound of real videos with the sound track from another randomly selected video. In details, 500 samples are randomly selected from D_A and D_V, respectively. Then the audio in D_A is randomly superimposed on a video in D_V, hence the instrument and sound in the video do not match.

In addition, we also employ the above synthesizing operation on the ensemble videos, where about 1,000 videos are collected in the same way as ESIT and EDIT in Sec. 2.1, but the collected videos are not in the videos in Sec. 2.1. Finally, a total of about 1,867 synthetic videos are obtained, which constitutes the whole musical performance video set with the real-world ones.

3. QA pair Collection

3.1. Questions Design

In different modality scenarios, 33 question templates covering 9 question types are proposed. Tab. 2 shows 9 question types in different scenarios, and the specific 33 question templates are given in Sec. 7.

3.2. QA pairs Collection

We design an audio-visual question answering labeling system to collect questions, and all QA pairs are collected with this system. The flow chart of the labeling system is shown in Fig. 2. First, questions are required to raise w.r.t three different modality scenarios, namely *Audio-Visual*, *Visual* and *Audio*, to explore the different modal contents. Then, for each modality scenario, different question types are designed to meet the requirements of scene understanding and reasoning, such as *existential, counting, location*, etc. At last, for each question type, we design multiple question templates that consist of fixed sentence pattern and formulas.

Table 2. Three scenarios and their corresponding question types.

Audio-Visual	Visual	Audio
Existential Counting Location Comparative Temporal	Counting Location	Counting Comparative

3.3. QA pairs samples

The large-scale spatial-temporal audio-visual dataset that focuses on question-answering task, as shown in Fig. 3

4. Auxiliary experiments

4.1. Temporal modeling with shufled segments.

To better evaluate the Temporal Grounding (TG) module and answer the question, we exclude the Spatial Grounding module and shuffle each input video in the time dimension of AV+Q+TG model. Without shuffling, the performance on the temporal questions is 65.17 while the performance drops to 63.71 after shuffling. Since the TG module

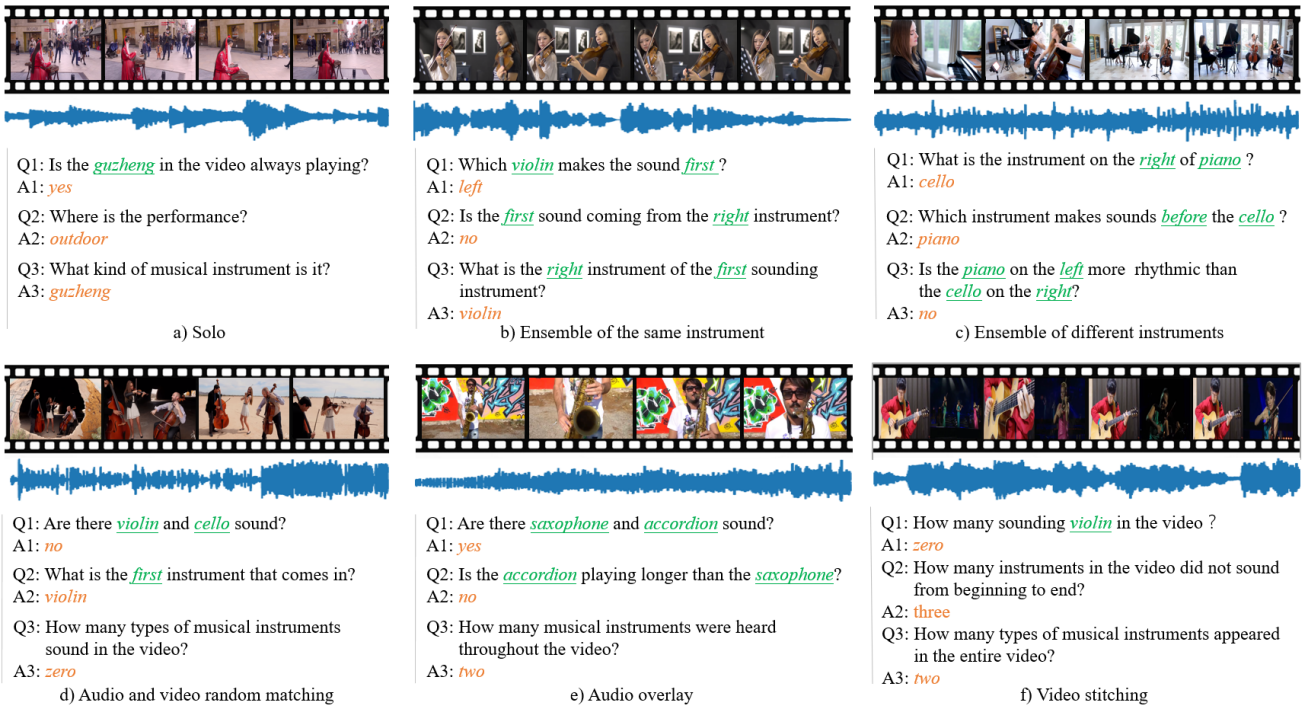


Figure 3. Different audio-visual scene types and their annotated QA pairs in the AVQA dataset. In the first row, a), b), and c) represent real musical performance videos, namely *solo*, *ensemble of the same instrument*, and *ensemble of different instruments*. In the second row, d), e), and f) represent the synthetic video, which are *audio and video random matching*, *audio overlay*, and *video stitching*, respectively.

does not explicitly encode the temporal order information of videos, shuffling the video segments does not affect the performance a lot. But the model with correct temporal information still achieves better on Temporal questions. One possible reason is that temporal-related words in questions, such as first and last, can implicitly help the model group to the corresponding temporal location. To further improve temporal question answering and strengthen temporal reasoning capability of our framework, it would be interesting to explore explicitly utilizing the temporal order information from the two modalities in the future.

4.2. Modeling with motion information

To further utilize the temporal information of the video, we use R(2+1)D network to extract motion features, which are fused to visual features. Our method with motion information achieves 71.75 on the released MUSIC-AVQA dataset, which is better than our method (71.53). According to the results, the model performance is boosted when combining motion information.

4.3. Experiments on existing video QA dataset

To explore whether the existing video QA dataset is suitable for AVQA task, we conduct experiments on the TVQA dataset [2], a large-scale video QA dataset based on 6 popular TV shows. Since the original TVQA framework does not take audio information as input, we add an audio encoder, a pre-trained VGGish [1] model, to extract audio features. Also, to be fair, we only take the ImageNet features as the video input, and the temporal dimension of question/answer features are squeezed by average operation. Different inputs are taken to comparison.

As the results shown in Tab. 3, both Q+V and Q+V+A methods are not superior to Q-only method based on common sense, which is consistent with the results reported in TVQA [2]. In addition, our method outperforms TVQA method in both visual-only and audio-visual inputs. But the introduced audio modality harms the performance of both methods. We consider the reason is that the sound in TVQA dataset is mainly human speech [2], and it is hard to modelling the interaction across both modalities. This phenomenon indicates that TVQA dataset is not quite suitable for the AVQA task which needs to explore the interactions between audio and visual components. In such a situation, our method still shows better robustness with less performance drop.

Table 3. **Experiments on TVQA dataset.** Q: Question. V: Video. A: Audio. *: TVQA method. †: Our method.

Method	Accuracy
Q-only*	43.50
Q+V*	41.70
Q+V+A*	41.45
Q+V†	42.01
Q+V+A†	41.95

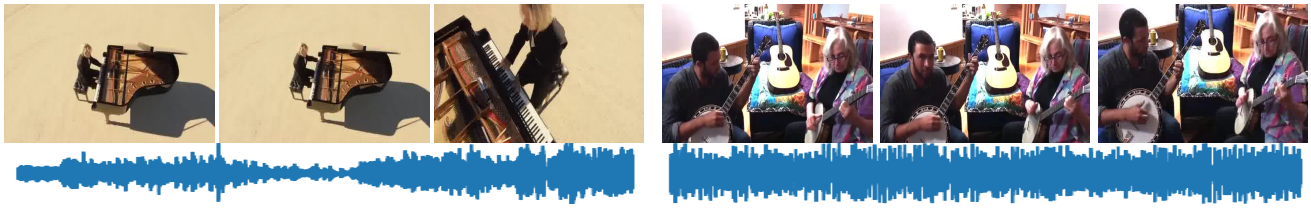


Figure 4. Ablation on input modalities. Left: leveraging the audio modality, the model A+Q can answer the correct instrument *piano* in the video. Right: with the help of the visual modality, V+Q recognizes two banjos in the video. However, the A+Q gives an wrong answer since it is more difficult to distinguish the number of sound sources in the same category for the audio.

Figure 5. Our audio-visual model predicts the correct answer but the individual audio and visual models fail. To answer this question, the model needs to perform multimodal scene understanding and temporal reasoning over the video.

5. Examples

To further study different input modalities and validate the effectiveness of the proposed model and compare to recent QA methods, we visualize some QA examples and have following findings:

First, audio improves question answering. The left example in Fig. 4 shows that the additional audio modality helps our model to answer the question. With the assistance of audio, the model can distinguish which instrument is playing. Second, visual modality is crucial. The visual modality is a strong signal for QA. One example is illustrated in the right of Fig. 4. In this case, recognizing sounds from complicated sound mixtures are very challenging, especially when two sounds are in the same category, while different sources are naturally isolated in the visual modality. The interesting results can support that auditory scene understanding can also benefit from visual perception. Third, multisensory perception boosts QA. An example is shown in Fig. 5. With recognizing sounding scenes and performing temporal reasoning, our audio-visual model can identify the sounding instrument trumpet after the accordion. From the results, we can learn that the two different modalities contain complementary information and multisensory perception is helpful for the fine-grained scene understanding task. Last but not least, to validate effectiveness of the proposed method, we compare it to a recent AVQA method: Pano-AVQA [3]. Several samples are provided in Fig. 6. We can find that our method, which explicitly constructs the association between audio and visual modalities and temporally aggregates audio and visual features, can predict correct answers to the questions and obtains superior performance.

6. Personal data/Human subjects

Videos in MUSIC-AVQA are public on YouTube, and annotated via crowdsourcing. We have explained how the data would be used to crowdworkers. Our dataset does not contain personally identifiable information or offensive content.

7. Question Templates

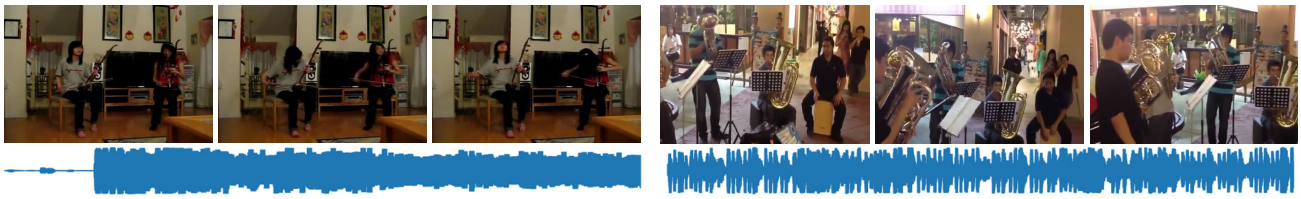
The 33 question templates in the AVQA dataset are shown in Table 4.

References

[1] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 776–780. IEEE, 2017.

[2] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

[3] Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2031–2041, 2021.



Question: Which erhu makes the sound first?
GT: simultaneously **Pano-AVQA:** left **Ours:** simultaneously

Question: How many sounding tuba in the video?
GT: four **Pano-AVQA:** three **Ours:** four

Question: Where is the first sounding instrument?
GT: right **Pano-AVQA:** left **Ours:** right

Question: What is the third instrument that comes in?
GT: accordion **Pano-AVQA:** saxophone **Ours:** accordion

Question: Where is the loudest instrument?
GT: right **Pano-AVQA:** left **Ours:** right

Question: Is this sound from the instrument in the video?
GT: no **Pano-AVQA:** yes **Ours:** no

Figure 6. Audio-visual question answering results. Our model can predict correct answers to the questions and is better than the recent AVQA method: Pano-AVQA [3].

Table 4. The 33 question templates.

Modalities	Question Types	Question Templates
Audio-Visual	Existential	Is this sound from the instrument in the video? Is the <*Object*> in the video always playing? Is there a voiceover?
	Counting	How many instruments are sounding in the video? How many types of musical instruments sound in the video? How many instruments in the video did not sound from beginning to end? How many sounding <*Object*> in the video?
	Location	Where is the <*LL*> instrument? Is the <*FL*> sound coming from the <*LR*> instrument? Which is the musical instrument that sounds at the same time as the <*Object*>? What is the <*LR*> instrument of the <*FL*> sounding instrument?
	Comparative	Is the instrument on the <*LR*> more rhythmic than the instrument on the <*RL*>? Is the instrument on the <*LR*> louder than the instrument on the <*RL*>? Is the <*Object*> on the <*LR*> more rhythmic than the <*Object*> on the <*RL*>? Is the <*Object*> on the <*LR*> louder than the <*Object*> on the <*RL*>?
	Temporal	Where is the <*FL*> sounding instrument? Which <*Object*> makes the sound <*FL*>? Which instrument makes sounds <BA> the <*Object*>?
Visual	Counting	Is there a <*Object*> in the entire video? Are there <*Object*> and <*Object*> instruments in the video? How many types of musical instruments appeared in the entire video? How many <*Object*> are in the entire video?
	Location	Where is the performance? What is the instrument on the <*LR*> of <*Object*>? What kind of musical instrument is it? What kind of instrument is the <*LRer*> instrument?
Audio	Counting	Is there a <*Object*> sound? How many musical instruments were heard throughout the video? How many types of musical instruments were heard throughout the video?
	Comparative	Is the <*Object1*> more rhythmic than the <*Object2*>? Is the <*Object1*> louder than the <*Object2*>? Is the <*Object1*> playing longer than the <*Object2*> ?