

# NarrAD: Supplementary Materials

## 1. Instruction Prompts

To solve the subtasks for AD generation, we leverage GPT-4o with prompt engineering. The instruction prompts utilized for each subtask are as follows: Fig. 1 is the instruction prompt for scene recognition, which involves finding the scene heading that best matches the given video segment’s background. Fig. 2 is the instruction prompt for generating AD by incorporating the narrative context of the movie script with the video segment. Fig. 3 is the instruction prompt for dividing complex sentences into the minimal semantic units. Fig. 4 is the instruction prompt for identifying semantic units that contain overlapping information between adjacent ADs. Fig. 5 is the instruction prompt for curating high-priority semantic units and reconstituting them into a single sentence.

**Instruction**

Find where the given video frames belong to among the backgrounds below.  
List all the likely candidates.

Just answer a number of the background. Do not say your reasoning.  
If there are multiple matched backgrounds, list all numbers separated by ,

List of scene headings

Video frames: Video frames

Figure 1. Instruction prompt for scene recognition. (§4.1)

You will be provided with a movie script delimited by triple quotes and several video frames. Your task is to generate audio description of the video using the given movie script in up to 10 words. Audio description is a short narration that explain the video to visually impaired individuals. It should contain name of characters and narrative context of the situation provided from the movie script. If the script does not contain the information needed to describe the video then simply describe it. You should describe the whole video in one sentence, rather than describing each video frame separately. Here are some examples of various audio descriptions of other movies.

AD Examples

Video frames

Figure 2. Instruction prompt for AD generation. (§4.2)

## 2. LLM Evaluator

We use an LLM as an evaluator to assess AD from more diverse perspectives. Fig. 7 shows the instruction prompt

**Instruction**

Your task is a recursive sentence splitting into semantic units, which are typically composed of a simple subject-predicate-object structure. You have to generate an intermediate representation that presents a simple and more regular structure. Each semantic unit should have only one information piece. Format of response should be unit1\*unit2\*unit3\*...

**Examples for In-context learning**

Example input 1    Example output 1

Example input 2    Example output 2

**Target sentence**

Target sentence

Figure 3. Instruction prompt for sentence splitting. (§4.3)

**Instruction**

You are given a sentence pairs consisting of two sentences separated by \*. Your task is to determine whether the two sentences in each pair convey similar information. The two sentences don't need to exactly be same. If the information of another sentence can be inferred from one sentence, it is accepted as the same. If two sentences contain similar information, answer O or answer X.

**Examples for In-context learning**

Example input 1    Example output 1

Example input 2    Example output 2

**Target sentence**

Unit1\*Unit2

Figure 4. Instruction prompt for finding duplicates. (§4.3)

**Instruction**

You are given several semantic units separated by \*. Each semantic unit represents a piece of information in a simple subject-predicate-object structure. Your task is to combine these units into a single sentence that contains all the information from the units and is grammatically correct and natural.

**Examples for In-context learning**

Example input 1    Example output 1

Example input 2    Example output 2

**Target sentence**

Unit1\*Unit2\*Unit3 ....

Figure 5. Instruction prompt for iterative reconstitution. (§4.3)

used to measure the SegEval score. For SegEval, we set the segment size  $L = 5$  and the context window size  $W = 3$ . Fig. 8 shows the instruction prompt used to measure the LLM-AD-eval score. For LLM-AD-eval, we compare the target and reference at text level ( $L = 1$ ) and sequence level ( $L = 5$ ). We use GPT-4o-mini as an evaluator.

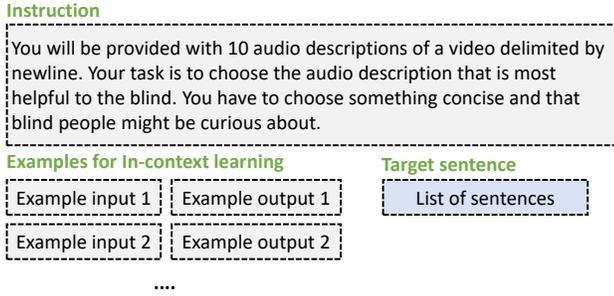


Figure 6. Instruction prompt for AD selection.

### 3. User Study Details

#### 3.1. Design, Procedure and Measures

Using a single factor between-subjects design, we randomly assigned participants to one of four AD types (NarrAD, autoAD-Zero, Expert-created, or MM-Narrator) to evaluate their effectiveness. All participants were randomly assigned to one of the four AD types and listened to an audio excerpt of one of the three movie: *The Ides of March* (political drama, dialogue-light), *How Do You Know* (romantic comedy, dialogue-heavy), and *Charlie St. Cloud* (fantasy, dialogue-light). The comprehension test consisting of seven true/false questions about the movie’s content. They answer each statement with “True”, “False”, or “I don’t know”. The number of correct answers is used to measure movie comprehension. Afterward, participants watched the same excerpt again and evaluated the AD based on usefulness (2 items: e.g., “The screen description aids in comprehending the movie content; Cronbach’s alpha = 0.89, 0.75, 0.70, respectively) and specificity (1 item: The screen description provides all the necessary information in sufficient detail to understand the movie). using 7-point scales (1: very strongly disagree, 7: very strongly agree). They also rated their likelihood to recommend the AD they experienced (i.e., “How much would you recommend the screen descriptions you experienced to a blind friend?) on a 7-point scale (1: not at all, 7: very much). The questions for comprehension test and surveys are shown in Fig. 9. Video samples for study can be found at <https://bit.ly/4aSwOTr>.

#### 3.2. Detailed Statistical Analysis of Results

A multivariate analysis of variance (MANOVA) revealed a significant effect of AD type on all measures ( $p$ -values  $< .05$ ). To further investigate these effects, we conducted single-step multiple comparison process utilizing the Dunnett method [1], comparing the NarrAD group with each of the other groups.

**Comprehension.** For *The Ides of March*, the comprehension score for NarrAD ( $M = 6.24$ ,  $SD = 1.07$ ) was significantly higher than those for autoAD-Zero ( $M = 1.6$ ,  $SD =$

$1.2$ ,  $p < .001$ ), Expert-created ( $M = 4.16$ ,  $SD = 2.22$ ,  $p < .001$ ), and MM-Narrator ( $M = 1.82$ ,  $SD = 1.68$ ,  $p < .001$ ).

For *How Do You Know*, the comprehension score for NarrAD ( $M = 4.06$ ,  $SD = 1.16$ ) was significantly higher than those for autoAD-Zero ( $M = 2.27$ ,  $SD = 1.27$ ,  $p < .001$ ) and MM-Narrator ( $M = 2.02$ ,  $SD = 1.48$ ,  $p < .001$ ) but did not significantly differ from Expert-created ( $M = 3.98$ ,  $SD = 1.66$ ,  $p = .98$ ).

For *Charlie St. Cloud*, the comprehension score for NarrAD ( $M = 4.25$ ,  $SD = 1.44$ ) was significantly higher than those for autoAD-Zero ( $M = 3.54$ ,  $SD = 1.18$ ,  $p = .04$ ) and MM-Narrator ( $M = 2.98$ ,  $SD = 1.59$ ,  $p < .001$ ) but did not significantly differ from Expert-created ( $M = 4.51$ ,  $SD = 1.63$ ,  $p = .71$ ).

**Usefulness.** For *The Ides of March*, NarrAD was rated significantly higher in usefulness ( $M = 4.33$ ,  $SD = 1.54$ ) compared to autoAD-Zero ( $M = 1.59$ ,  $SD = 1.04$ ,  $p < .001$ ), Expert-created ( $M = 3.17$ ,  $SD = 1.7$ ,  $p < .001$ ), and MM-Narrator ( $M = 1.83$ ,  $SD = 1.56$ ,  $p < .001$ ).

For *How Do You Know*, NarrAD was rated marginally significantly higher in usefulness ( $M = 3.54$ ,  $SD = 1.35$ ) compared MM-Narrator ( $M = 2.86$ ,  $SD = 1.52$ ,  $p = .093$ ) but did not differ from autoAD-Zero ( $M = 2.99$ ,  $SD = 1.92$ ,  $p < .22$ ) and Expert-created ( $M = 3.77$ ,  $SD = 1.62$ ,  $p = .83$ ).

For *Charlie St. Cloud*, NarrAD was rated marginally significantly higher in usefulness ( $M = 3.76$ ,  $SD = 1.59$ ) compared to autoAD-Zero ( $M = 2.98$ ,  $SD = 1.89$ ,  $p = .05$ ) but did not differ from Expert-created ( $M = 3.74$ ,  $SD = 1.61$ ,  $p = 1.00$ ) and MM-Narrator ( $M = 3.17$ ,  $SD = 1.61$ ,  $p = .18$ ).

**Specificity.** For *The Ides of March*, the specificity rating for NarrAD ( $M = 4.25$ ,  $SD = 1.62$ ) was significantly higher than those for autoAD-Zero ( $M = 1.56$ ,  $SD = 1.11$ ,  $p < .001$ ) and MM-Narrator ( $M = 2.39$ ,  $SD = 1.77$ ,  $p < .001$ ). Its specificity rating did not significantly differ from Expert-created ( $M = 4.02$ ,  $SD = 1.80$ ,  $p = .80$ ).

For *How Do You Know*, the specificity rating for NarrAD ( $M = 4.12$ ,  $SD = 1.72$ ) was significantly higher than those for autoAD-Zero ( $M = 3.02$ ,  $SD = 2.07$ ,  $p = .008$ ) and MM-Narrator ( $M = 3.18$ ,  $SD = 1.84$ ,  $p = .026$ ) but did not significantly differ from Expert-created ( $M = 4.60$ ,  $SD = 1.56$ ,  $p = .43$ ).

For *Charlie St. Cloud*, the specificity rating for NarrAD ( $M = 3.7$ ,  $SD = 1.87$ ) was did not differ from autoAD-Zero ( $M = 3.00$ ,  $SD = 2.02$ ,  $p = .15$ ), Expert-created ( $M = 4.04$ ,  $SD = 1.74$ ,  $p = .68$ ), and MM-Narrator ( $M = 3.45$ ,  $SD = 1.77$ ,  $p = .84$ ).

**Likelihood of recommendation.** For *The Ides of March*, participants indicated that they would be more likely to recommend NarrAD to a blind friend ( $M = 4.51$ ,  $SD = 1.76$ ) compared to autoAD-Zero ( $M = 1.48$ ,  $SD = 1.01$ ,  $p < .001$ ), Expert-created ( $M = 3.7$ ,  $SD = 1.98$ ,  $p = .033$ ), and MM-Narrator ( $M = 1.98$ ,  $SD = 1.51$ ,  $p < .001$ ).

For *How Do You Know*, participants indicated that they would be more likely to recommend NarrAD to a blind friend ( $M = 4.14$ ,  $SD = 2.0$ ) compared to autoAD-Zero ( $M = 2.96$ ,  $SD = 1.87$ ,  $p = .006$ ) and MM-Narrator ( $M = 2.73$ ,  $SD = 1.84$ ,  $p < .001$ ). It did not significantly differ from Expert-created ( $M = 4.57$ ,  $SD = 1.85$ ,  $p = .54$ ).

For *Charlie St. Cloud*, participants indicated that they would be more likely to recommend NarrAD to a blind friend ( $M = 3.73$ ,  $SD = 1.77$ ) compared to MM-Narrator ( $M = 2.82$ ,  $SD = 1.88$ ,  $p = .04$ ) but not significantly differ from autoAD-Zero ( $M = 3.24$ ,  $SD = 2.06$ ,  $p = .43$ ) and Expert-created ( $M = 3.88$ ,  $SD = 1.82$ ,  $p = .95$ ).

As the likelihood of recommendation often reflects the overall evaluation of a product or a service [2], this result suggests that NarrAD outperformed the other ADs in terms of overall user satisfaction.

## 4. Implementation Details

In this work, we use the GPT-4o, as a multimodal LLM. In dialogue synchronization (§4.1), we extract movie dialogues using Automatic Speech Recognition through the Google Cloud Video Intelligence API. Then, to calculate the Levenshtein distance between the movie script dialogue and the movie dialogue, we use the Python FuzzyWuzzy library. To ensure accurate matching results, we set the threshold for word similarity at 67. In AD generation (§4.2), we sample 8 frames from video frames extracted at 5fps. In the movie script, only stage directions are utilized, excluding dialogues. For the same input, we generate a total of 10 outputs and then select the sentence that is most useful for visually impaired individuals by using GPT-4o with the prompt in Fig. 6 In information curation (§4.3), we prioritize curation based on content that overlaps with adjacent ADs. We define an AD within 20 seconds before and after as an adjacent ADs. For *Hansel & Gretel: Witch Hunters* (2013), due to the unavailability of the movie script, a synopsis from IMDb is used instead to generate AD. Therefore, for an in-depth analysis of video-to-script retrieval (§5.5), we exclude the movie.

## References

- [1] Charles W Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121, 1955. 2
- [2] Markus Groth. Customers as good soldiers: Examining citizenship behaviors in internet service deliveries. *Journal of management*, 31(1):7–27, 2005. 3

Suppose you are a visually impaired person, and you will be "watching" a movie videoclip with audio description (AD). Here, you are requested to provide feedback (via reasoning and marking) on the performance of two AI assistants ("ASSISTANT1" and "ASSISTANT2") for automatic AD generation task:

**Evaluation Steps:**

1. you will be given <Context ADs>, <ASSISTANT1-output>, and <ASSISTANT2-output>, where <Context ADs> shows a few contextual human-annotated ADs but leaves "<PRESENT-SEGMENT>" empty to be filled with one or multiple AD(s) generated by two AI assistants (i.e., <ASSISTANT1-output> and <ASSISTANT2-output>).
2. you will read though the <Context ADs> and <ASSISTANT1-output> and <ASSISTANT2-output>, and then measure the AD generation quality of the two AI assistants in terms of coherence aspect.
3. you will complete the following five sections IN ORDER (namely, <Assistant1-Reasoning>, <Assistant2-Reasoning>, <Comparison-Reasoning>, <Assistant1-Score>, and <Assistant2-Score>).

**HINT:**

1. <Context ADs> will be used to provide the context of the movie scene. If it contains no valid ADs, it means that the current evaluation metric (coherence) will not take contextual information into account.
2. <Assistant1-Reasoning> and <Assistant2-Reasoning> will be used to record your reasoning and comments (with supporting evidence) on the coherence aspect of the ADs generated by two AI assistants, respectively;
3. <Comparison-Reasoning> will be used to record your feedback (with supporting evidence) for comparisons between the two AI assistants (with respect to the coherence aspect), which will be used to support the below two marking sections;
4. <Assistant1-Score> and <Assistant2-Score> will be used to record your AD generation coherence scores (from "1" to "10", where "1" indicates the worst and "10" indicates the excellent) of the two AI assistants, respectively.

**Evaluation Criteria:**

- Evaluation criteria:

Evaluation Criteria

-----Evaluation Starts-----

<Context ADs>

Previous ground-truth ADs

\* At present: ---PLACEHOLDER for "<PRESENT\_SEGMENT>" to be generated by AI assistants below---

Future ground-truth ADs

<ASSISTANT1-output>

Current generated ADs

<ASSISTANT2-output>

Current ground-truth ADs

Please make sure you read and understand these instructions carefully, and complete the following five sections IN ORDER:

- (1) firstly reason them individually within "<Assistant1-Reasoning>" and "<Assistant2-Reasoning>";
- (2) secondly compare two assistants within "<Comparison-Reasoning>"; and
- (3) finally mark them within "<Assistant1-Score>" and "<Assistant2-Score>"

**Coherence Criteria:** Determines whether <PRESENT-SEGMENT> logically connects to the given <Context ADs>. A coherent text flows smoothly and deepen the movie understanding for the visually impaired.

**Specificity Criteria:** Measures the level of detail in the generated <PRESENT-SEGMENT>, assessing if it is sufficiently detailed and/or focused for the <Context ADs>.

Figure 7. Instruction prompt for SegEval.

You are an intelligent chatbot designed for evaluating the quality of generative outputs for movie audio descriptions. You are given N consecutive audio descriptions describing a particular scene from a movie, separated by \*. Your task is to compare the predicted audio descriptions with the correct audio descriptions and determine its level of match, considering mainly the visual elements like actions, objects, and interactions. Here's how you can accomplish the task:

Please evaluate the following movie audio description pair:

Correct Audio Description:

Predicted Audio Description:

Provide your evaluation only as a matching score where the matching score is an integer value between 0 and 5, with 5 indicating the highest level of match. Please generate the response in the form of a Python dictionary string with keys 'score', where its value is the matching score in INTEGER, not STRING. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. For example, your response should look like this: {'score': }

Figure 8. Instruction prompt for LLM-AD-eval.

**Rating 1 to 7 ( 1 : Very Strongly Disagree, 7 : Very Strongly Agree )**

**Usefulness**

1. With the help of the screen description, I could understand the situation without watching the video.
2. The screen description helped with immersion in the movie and made the viewing experience more satisfying.

**Specificity**

3. The screen description provides all the necessary information in sufficient detail to understand the movie.

**Likelihood of recommendation.**

4. How much would you recommend the screen descriptions you experienced to a blind friend?

**Answer each item with True, False, or I don't know**

**[The ideo of march]**

- In the beginning, Stephen and Molly went to a clinic.
- In the beginning, Stephen and Molly went to a hotel.
- Stephen and Molly waited in the waiting room.
- Stephen and Molly waited in the hotel lobby.
- Stephen and Molly looked comfortable and joyful.
- Stephen and Molly looked anxious and depressed.
- Paul lay on the bed, talking on his phone.

**[How do you know]**

- Lisa took her phone charger from her bag.
- George opened a refrigerator.
- George got a phone call from his dad.
- George got a phone call from his friend.
- Lisa opened a refrigerator.
- While Lisa was on the phone, George was watching from the kitchen.
- While Lisa was on the phone, George was sitting across from her at the table.

**[Charlie St. Cloud]**

- Charlie took a baseball.
- Charlie took a cell phone.
- Charlie checked a sunset chart on a wall.
- Charlie checked a boat schedule on a wall.
- Charlie is wearing a T-shirt.
- Charlie walked through a sunlit forest, carrying a bag.
- Charlie showed a toy to Sam.

Figure 9. The questions for comprehension test and surveys.