Shot-by-Shot: Film-Grammar-Aware Training-Free Audio Description Generation

Supplementary Material

This appendix is organised as follows:

- **Implementation details:** In Appendix A, we provide additional details on film grammar prediction, action score evaluation settings, and the exact text instructions.
- **Evaluation metrics for AD generation:** In Appendix B, we elaborate on the key metrics used to measure AD quality.
- **Additional experimental results:** In Appendix C, we provide additional results for AD generation.
- **Human alignment with action scores:** In Appendix D, we provide a thorough description of the human agreement study for action scores, including curated test sets, a correlation analysis, and an inter-rater agreement study.
- **Quantitative results:** In Appendix E, we present the complete results for AD generation on MAD-Eval.
- **Qualitative visualisations:** Appendix F includes more detailed visualisations, along with an in-depth analysis of failure cases and assisted AD generation.

A. Implementation details

Thread structure prediction setting. To predict thread structure in a zero-shot manner, we resize video frames to 224p and employ DINOv2 [20] ViT-g/14 w. reg to extract spatial features. During matching score prediction, we consider a 5×5 mask neighbourhood for cost volume computation and set the softmax temperature to $\tau=0.1$. To determine the relationship between each pair of shots, we apply a threshold of $\epsilon=0.3$ to the matching score $s^{i,j}$.

Shot scale classification setting. For shot scale classification, we fine-tune the last 6 layers of the DINOv2 [20] ViT-B/14 on the MovieShots training set. During evaluation, we use the middle frame of each shot as input.

The averaged shot scale of the current shots (i.e. the effective shot scale S_{eff}) is used to guide the incorporation of additional factors in Stage I prompts. Specifically,

$$\text{Stage I factor} \mathrel{+}= \begin{cases} \text{Facial expression}, & \text{if } S_{\text{eff}} \leq 1.5 \\ \text{Key object}, & \text{if } 2 \leq S_{\text{eff}} \leq 3 \\ \text{Environment}, & \text{if } S_{\text{eff}} \geq 3.5 \\ \text{None}, & \text{otherwise} \end{cases}$$

Action score evaluation setting. To obtain character-free GT action sentences, we employ LLaMA3.1-70B [16] for pre-processing in two steps: (i) *Character information removal:* Character names are replaced with appropriate pronouns using the LLM with the prompt provided in Algorithm 3; (ii) *Action sentence extraction:* Each AD sentence is split into subsentences, each containing a single action. To achieve this, the LLM is prompted with instructions in Algorithm 4.

During hierarchical prediction parsing, we use spaCy[1] to extract action phrases and corresponding verb lemmas from predicted sentences.

During similarity score computation, to extract sentence embeddings, we apply gte-Qwen2-7B-instruct [12], which supports optional text prompt input as guidance. We set the prompt to: *"Retrieve relevant passages that involve similar actions, with particular focus on the verbs."*, further emphasising actions and verbs during similarity matching.

Additionally, to establish the LLM-based baseline, we define evaluation criteria as outlined in Algorithm 5 and use them to prompt LLaMA-3.1-70B and GPT-4o.

GPT-4o setup for AD generation. For both stages, we use `gpt-4o-2024-08-06` [19] as the base model. For visual token extraction, the "detail" parameter is set to "low".

Text instructions for AD generation. The prompts for AD generation are provided in Algorithms 1 and 2 for Stage I and Stage II, respectively. The Stage I prompt is designed for both Qwen2-VL and GPT-4o, while the Stage II prompt is tailored for LLaMA3 and GPT-4o.

B. Evaluation metrics for AD generation

CIDEr [24] measures text similarity by computing a weighted word-matching score, emphasising n-gram overlap while accounting for term frequency and importance through TF-IDF [22] weighting.

Recall@k/N [9] is a retrieval-based metric that evaluates whether predicted texts can be distinguished from their temporal neighbours. Specifically, for each predicted AD, it checks whether the AD can be retrieved at a top-k position within a neighbourhood of N ADs. Following prior work [8, 11], we report Recall@1/5 on CMD-AD and TV-AD, and Recall@5/16 on MAD-Eval.

LLM-AD-Eval [11] employs LLM agents (LLaMA3-8B [16] | LLaMA2-7B [6]) as evaluators to compare ground truth ADs with predictions, generating a matching score ranging from 1 (lowest) to 5 (highest).

For MAD-Eval, we additionally report the performance on conventional metrics including ROUGE-L [13], SPICE [2], METEOR [4], and BLEU-1 [21].

The fixed and distinct set of character names in each video can bias conventional captioning metrics. For instance, TF-IDF weighting in CIDEr assigns high importance to character names. To provide a more independent measure of character names and other AD content, we consider CRITIC and the new action score for CMD-AD and TV-AD evaluation.

CRITIC [11] measures the accuracy of character names in

[1] https://spacy.io/models/en

Stage I VideoLLM	Stage II LLM	CMD-AD				TV-AD			
		CIDEr	CRITIC	Action (Stage I)	Action (Stage II)	CIDEr	CRITIC	Action (Stage I)	Action (Stage II)
Qwen2.5-VL-7B [3]	LLaMA3-8B [16]	24.1	**49.7**	35.5	27.2	26.5	**43.6**	36.4	23.8
VideoLLaMA3-7B [30]	LLaMA3-8B [16]	22.1	45.1	34.2	24.8	26.8	41.3	**39.9**	23.8
InternVL2.5-8B [5]	LLaMA3-8B [16]	24.0	46.0	35.0	28.1	28.3	41.2	36.0	**24.2**
Qwen2-VL-7B [26]	LLaMA3-8B [16]	**26.3**	47.8	**38.2**	**28.4**	**31.1**	42.2	38.2	23.9
Qwen2-VL-7B [26]	Gemma3-12B [23]	**26.4**	45.9	–	**30.6**	30.0	43.2	–	**24.7**
Qwen2-VL-7B [26]	Qwen3-8B [28]	26.3	45.2	–	28.4	30.7	**43.4**	–	**24.7**
Qwen2-VL-7B [26]	LLaMA3-8B [16]	26.3	**47.8**	–	28.4	**31.1**	42.2	–	23.9
GPT-4o [19]	GPT-4o [19]	26.1	49.1	40.2	32.5	34.2	46.5	41.0	27.4

Table A1. **Different open-source VideoLLMs (for Stage I) and LLMs (for Stage II).** As a reference, the last row reports results using the proprietary GPT-4o model. Note, we additionally report action scores for predicted Stage I description, whereas other metrics, including CIDEr, CRITIC, and Action (Stage II), measure the Stage II AD quality.

Thread structure	Stage I VideoLLM	TV-AD subset		
		CIDEr	CRITIC	Action
✗	Qwen2.5-VL-7B [3]	23.0	42.2	23.2
✓	Qwen2.5-VL-7B [3]	23.9 ↑0.9	43.0 ↑0.8	23.2 0.0
✗	VideoLLaMA3-7B [30]	24.9	42.1	21.9
✓	VideoLLaMA3-7B [30]	25.5 ↑0.6	41.1 ↓1.0	22.7 ↑0.8
✗	InternVL2.5-8B [5]	25.9	40.3	21.8
✓	InternVL2.5-8B [5]	27.1 ↑1.2	41.9 ↑1.6	23.1 ↑1.3
✗	Qwen2-VL-7B [26]	28.8	42.0	22.6
✓	Qwen2-VL-7B [26]	30.7 ↑1.9	42.7 ↑0.7	22.9 ↑0.3

Table A2. **Thread structure injection for different open-source VideoLLMs.** The base model for Stage II is LLaMA3-8B. Thread structure information is injected only into subsets predicted to exhibit thread structures ($\sim$60% in TV-AD).

Exp.	CMD-AD			TV-AD		
	CIDEr	CRITIC	Action	CIDEr	CRITIC	Action
1	26.3	47.8	28.4	31.1	42.2	23.9
2	26.9	47.7	28.4	30.7	41.9	23.8
3	26.4	48.3	28.3	30.5	43.0	23.7
4	26.3	47.3	28.5	30.8	42.0	23.5
5	26.5	48.3	28.4	31.5	41.9	23.7
Mean	26.5	47.9	28.4	30.9	42.2	23.7
STD	0.2	0.4	0.1	0.4	0.5	0.1

Table A3. **Repeated (multi-run) experiments.** Results shown are from five independent runs (Stage I + Stage II) using different random seeds.

predicted ADs. It first resolves character ambiguity in GT ADs by applying a coreference model to replace pronouns with corresponding character names. During evaluation, the intersection-over-union (IoU) of predicted and ground truth character names is computed.

Action Score, as detailed in Sec. 4 of the main text, evaluates the quality of predicted actions (i.e., verbs, object nouns, etc.) while minimising the influence of character name variations. Throughout this paper, unless otherwise specified, we use the action score to assess Stage II AD outputs, i.e., "Action" refers to "Action (Stage II)".

Method	Shot Partition	Film Grammar	Stage I VideoLLM	Stage II LLM	Total
AutoAD-Zero	–	–	2.18s	0.64s	2.82s
Ours	0.09s	0.12s	2.82s	0.72s	3.75s

Table A4. **Inference time analysis.**

C. Additional experimental results

Different VideoLLMs for Stage I. Tab. A1 compares our AD generation performance using different open-source VideoLLMs in Stage I, validating our choice of Qwen2-VL-7B as the default model.

Beyond the Stage II action scores presented in the main text, we also report Stage I action scores as a direct indicator of dense description performance. In general, Stage I action scores are noticeably higher than their Stage II counterparts, suggesting that some ground truth actions are captured in dense descriptions but are not selected for the final AD outputs. This further supports the validity of our assisted AD generation protocol, where multiple candidate ADs with different actions are extracted from dense descriptions and await further selection.

Different LLMs for Stage II. Tab. A1 also compares different options for the Stage II LLM. Overall, the choice of LLM has a relatively minor impact compared to the Stage I VideoLLM. The default LLaMA3-8B achieves overall strong performance.

Thread structure injection for different VideoLLMs. We investigate how injecting thread structure into different open-source VideoLLMs affects AD generation. Specifically, we evaluate performance on TV-AD, which contains a large proportion of thread-structured video clips. As shown in Tab. A2, incorporating thread information leads to general performance boosts across various VideoLLMs.

Repeated AD generation. Tab. A3 reports results from five independent runs of our two-stage AD generation pipeline, each adopting a different random seed. The results are largely consistent across runs, indicating the stability of the AD generation process. In particular, the CRITIC results exhibit the highest variance, followed by CIDEr, while the action score

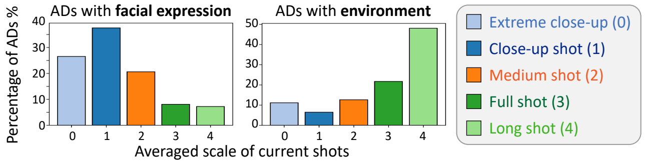


Figure A1. **Correlations between descriptive factors in ADs and averaged shot scales.**

remains relatively stable across repeated experiments.

Inference time analysis. We report the inference time *per AD* in Tab. A4. Our added components (e.g. shot partitioning, film grammar prediction) incur minimal overhead. The main cost increase arises from sampling more contextual frames for Stage I inputs. Overall, our method maintains reasonable efficiency—given averaged input clip duration (including contextual shots) is 19.82s—while delivering clear performance gains.

Investigation on factor–scale correlations. To verify our assumptions about the strong correlations between descriptive factors in ADs and shot scales (as mentioned in **??**), we used GPT-4o to extract key elements (e.g. environments) mentioned in GT ADs from 100 movies, and analysed their correlation with shot scales. As shown in Fig. A1, facial expressions predominantly occur in (extreme) close-up shots, while environmental cues align with full/long shots. These observations align with the assumption and generalise across real-world movies.

D. Human alignment with action scores

To assess how the proposed action score aligns with human knowledge, we curate a test set that measures the correlation between action scores and human scoring.

Scoring criteria. The human-annotated scores measure whether the ground truth (GT) action is described in the descriptions, ranging from 0 to 3 based on the following scoring criteria:

- Score 0 - *GT action is unrelated to any action in PD*
- Score 1 - *GT action is loosely related to an action in PD*
- Score 2 - *GT action is similar in meaning to an action in PD*
- Score 3 - *GT action exactly matches with an action in PD, using the same verb*

where PD stands for Predicted Description. Additional guidelines and scoring examples are provided in Tab. A5.

Test Set Formulation. We construct two test sets, namely the "paragraph set" and the "sentence set," corresponding to the scoring of (Stage I) dense descriptions and (Stage II) AD sentences, respectively.

The paragraph set consists of 300 ground truth (GT) ADs, each paired with a predicted paragraph. In total, around 460 character-free GT actions are extracted, with each action-paragraph pair manually annotated by five workers using the 0–3 scoring scale, as described in the previous section. For ADs containing multiple GT actions, the final human score is obtained by averaging the manually assigned scores across different actions, meaning the resultant score may not always be an integer.

The sentence set contains 500 GT ADs with approximately 890 actions. Given the video clip described by the GT AD, instead of generating AD predictions from a VLM, we use a human-narrated AD from alternative sources for the same clip as the prediction. Similarly, the final human score is computed by averaging the scores across different actions within each GT AD.

Correlation between human scoring and metrics. We plot the human-annotated scores against the scores reported by each metric, as shown in Fig. A2. Most conventional metrics (blue and pink) fail to align with human evaluations of action predictions. In contrast, both LLM-based metrics and our action scores effectively assess the quality of action predictions in AD sentences (Fig. A2, bottom).

However, when evaluating longer paragraphs against the GT action, LLM-based metrics struggle, whereas action scores maintain a high correlation with human judgments (Fig. A2, top). Quantitative results on human-metric correlations are provided in Tab. A6.

Inter-rater analysis for action score. To obtain an upper bound on action scores for predicted AD sentences, we measure the agreement between two versions of human-annotated ADs for the same movie clip at different temporal IoUs (i.e. inter-rater agreement [11]). As shown in Tab. A7, the action score increases monotonically as the temporal IoU increases.

E. Quantitative results

Tab. A8 presents the complete quantitative comparison on MAD-Eval, where our method outperforms all training-free approaches and remains competitive with state-of-the-art models fine-tuned on GT ADs.

F. Qualitative visualisations

Fig. A3 presents additional visualisations for CMD-AD and TV-AD, comparing our method against other approaches. For more visualisations, please refer to the Supp. Videos.

Example failure cases are illustrated in Fig. A4. The top example highlights a hallucination issue in the prediction by Qwen2-VL + LLaMA3 (Ours), where the VideoLLM model misidentifies a "gift" as "candy" and infers an incorrect action of "eating," which does not occur. When a stronger base model is used (Ours* with GPT-4o), this issue is mitigated.

Additionally, the current method struggles to incorporate broader (story-level) contextual understanding, as demonstrated in the bottom example of Fig. A4. Specifically, the model fails to describe the return of an insect and instead focuses on the sleeping woman. This limitation could potentially be addressed by incorporating more abstract information, which could be extracted from existing movie plots or summarised from a longer temporal context.

Visualisation of intermediate outputs. Fig. A5 provides more detailed visualisations, including intermediate results on thread structure and shot scale predictions, as well as Stage I descriptions. It also illustrates how the predicted shot scales influence the formulation of Stage I factors.

Score	Note	GT action	Prediction
0	**Unrelated**: Completely irrelevant in actions, objects, environments, etc.	*He indicates that he was just joking.*	*The video begins with a man sitting under an umbrella in the rain, wearing a dark jacket with red accents and a blue shirt underneath. The scene transitions to show him lying on the ground, wet from the rain, with his arms outstretched and his eyes closed. Another person, dressed in a white shirt and black vest, reaches towards him from above.*
		His new companion cuddles back up to him.	*The flames start to die down.*
1	**Loosely related**: Semantically or causally relevant actions; Similar objects, environments, etc.	*He points the remote at the word's main menu.*	*1. Main characters: Morty and Michael Newman. 2. Actions: Both characters are standing in a room with futuristic digital screens and symbols around them. They appear to be engaged in an activity involving the screens, possibly* ***interacting with them or using devices****. 3. Character-character interactions: The two characters seem to be focused on the screens, possibly discussing or collaborating on something related to the content displayed.*
		They step away from the window.	*The four of them* ***leave the cabin*** *and go to their horses tethered to a rail at the side.*
2	**Similar meaning**: Similar verb, objects, environments, etc.	*He dives into an Olympic-sized swimming pool.*	*The movie clip shows a young boy running out of a house wearing only a diaper. He then jumps over a fence and runs across a lawn. The boy continues running and* ***jumps into a swimming pool****. Throughout the video, the boy's movements are energetic and playful, and he appears to be having fun.*
		She carries a laundry basket outside the house.	*Caroline* ***brings some washing*** *into the garden.*
3	**Exactly same match:** Exactly the same verb; Same/similar objects, environments, etc.	*He lights a cigar.*	*In the movie clip, a man is seen sitting at a table, while another man enters the room and approaches him. The second man takes out a* ***cigarette*** *and* ***lights it****, while the first man watches him. The scene is set in a dimly lit room with a table and chairs in the foreground, and a window in the background.*
		She stares glumly at the night sky.	*At the palace, Jasmine wanders out into her balcony and* ***stares up at the stars****.*

Table A5. **Example of human-annotated scores** assessing whether the ground truth (GT) action is accurately captured in the predictions. For each score, examples of a paragraph prediction and a sentence prediction are provided.

Metric	Paragraph		Sentence	
	Pearson	Spearman	Pearson	Spearman
CIDEr [24]	0.205	0.264	0.412	0.528
ROUGE-L [13]	0.305	0.280	0.526	0.512
METEOR [4]	0.462	0.406	0.602	0.641
BLEU-1 [21]	0.265	0.264	0.477	0.481
SPICE [2]	0.022	0.048	0.031	0.012
BERTScore [33]	0.377	0.393	0.508	0.507
LLM-based (LLaMA3.1-70B [16])	0.569	0.491	0.779	0.798
LLM-based (GPT-4o [19])	0.742	0.678	0.797	0.807
Action Score (w/o verb matching)	0.735	0.728	0.765	0.790
Action Score (w verb matching)	**0.749**	**0.729**	**0.806**	**0.820**

Table A6. **Comparison of action score with other metrics.** The listed metrics measure the similarity between predicted paragraphs/sentences and ground truth actions. The reported values indicate the correlation (i.e. alignment) between these metrics and human-annotated scores.

tIoU	#movies	#AD pairs	CIDEr	R@1/5	Action	LLM-AD-Eval
0.8	315	4447	61.5	71.2	45.6	3.04 | 3.24
0.9	267	999	69.8	80.4	47.6	3.53 | 3.34
0.95	148	229	73.9	–	47.8	3.57 | 3.45

Table A7. **Inter-rater analysis on CMD-AD,** where two versions of human-annotated ADs for the same movie clip are compared under different temporal intersection-over-union (tIoU) thresholds.

Visualisation for assisted AD generation. Fig. A6 presents additional examples with Stage I descriptions, from which multiple AD candidates are extracted. Among the five AD predictions, the one that best aligns with the ground truth (based on the averaged CIDEr and action score) is highlighed.

In the top example, multiple actions are present in the Stage I dense description (e.g. *"kiss"*, *"hand on neck"*, *"eyes closed"*, etc.), resulting in AD candidates that differ in both subjects and actions. In contrast, the middle example contains fewer actions (e.g. *"shoot blue energy"* and *"look"*/*"observe"*). In this case, the AD candidates primarily vary in style, such as changes in subjects or sentence structures, providing varied options for selection.

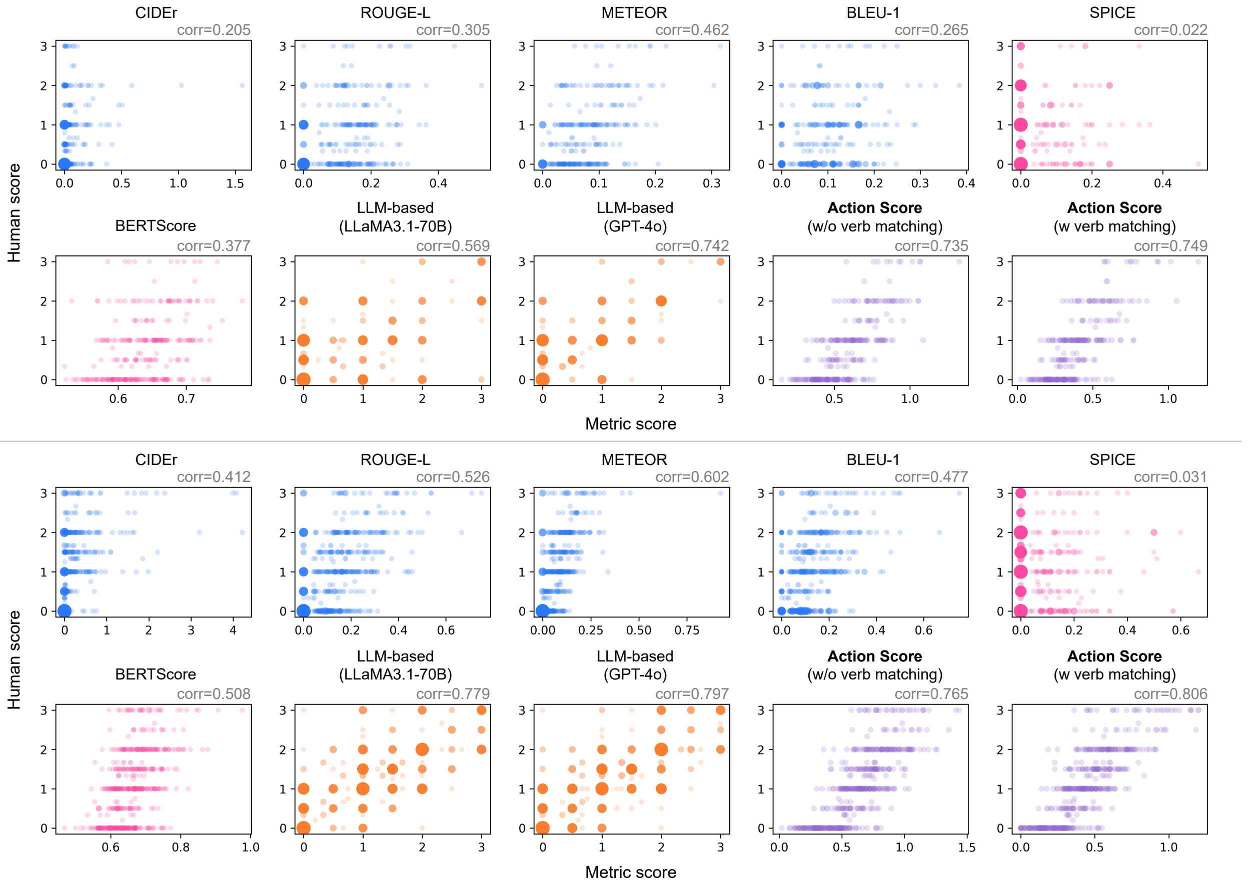
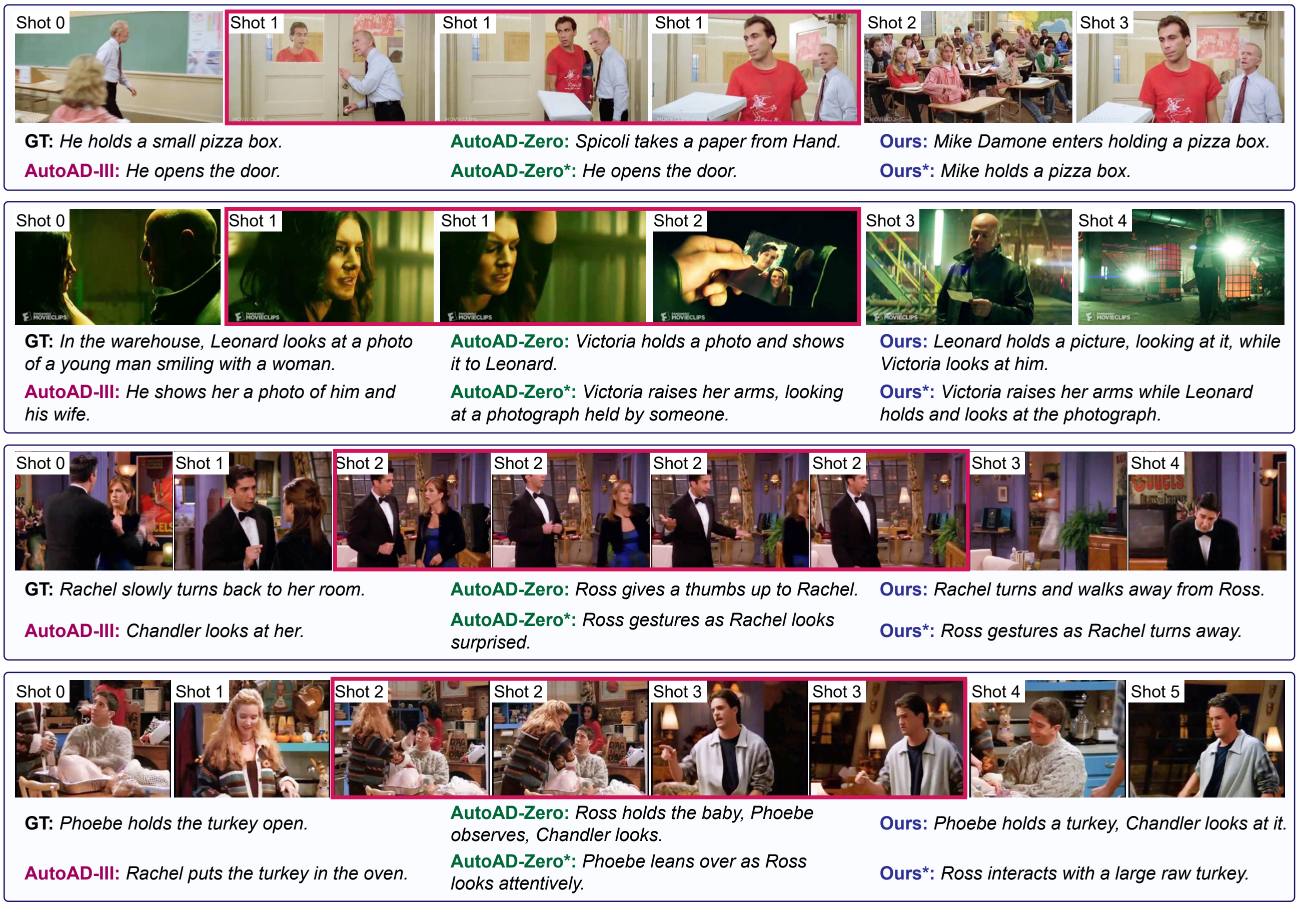


Figure A2. **Human agreement of different metrics for action evaluation,** where human-annotated scores are compared with metric scores. These scores measure the quality of the Stage I description paragraph (top) or the Stage II output AD (bottom). We consider various metrics, including word-matching-based metrics (blue), semantic-based metrics (pink), LLM-based metrics (orange), and our proposed action scores (purple). The Pearson correlation between human and metric scoring is reported. Within the scatter plots, we use colour depth and marker size to indicate density – larger and darker markers represent more data points at a single position. Zoom in for a clearer view.

Method	VLM	LLM	Training -free	Propriet. model	MAD-Eval					
					CIDEr	R@5/16	Rouge-L	SPICE	METEOR	BLEU-1
ClipCap [17]	CLIP-B32	GPT-2	✗	✗	4.4	36.5	8.5	1.1	–	–
CapDec [18]	–	–	✗	✗	6.7	–	8.2	1.4	–	–
AutoAD-I [9]	CLIP-B32	GPT-2	✗	✗	13.4	42.1	11.9	4.4	–	–
AutoAD-II [10]	CLIP-B32	GPT-2	✗	✗	19.5	51.3	13.4	–	–	–
AutoAD-III [11]	EVA-CLIP	LLaMA2-7B	✗	✗	24.0	52.8	13.9	6.1	5.5	13.1
MovieSeq [15]	CLIP-B16	LLaMA2-7B	✗	✗	24.4	51.6	15.5	7.0	–	–
DistinctAD [8]	CLIP$_{AD}$-B16	LLaMA3-8B	✗	✗	27.3	56.0	**17.6**	8.3	–	–
UniAD [25]	CLIP-L14	LLaMA-8B	✗	✗	**28.2**	54.9	17.2	–	–	–
Video-LLaMA [32]	Video-LLaMA-7B	–	✓	✗	4.8	33.8	–	–	–	–
Video-BLIP [29]	Video-BLIP	–	✓	✗	5.0	35.2	–	–	–	–
AutoAD-Zero [27]	VideoLLaMA2-7B	LLaMA3-8B	✓	✗	22.4	47.0	14.4	7.3	6.6	15.1
AutoAD-Zero [27]	Qwen2-VL-7B	LLaMA3-8B	✓	✗	23.6	51.3	14.6	7.8	6.6	13.6
Ours	Qwen2-VL-7B	LLaMA3-8B	✓	✗	25.0	50.6	14.7	7.8	7.2	**16.2**
VLog [1]	BLIP-2 + GRIT	GPT-4	✓	✓	1.3	42.3	7.5	2.1	–	–
MM-Vid [14]	GPT-4V	–	✓	✓	6.1	46.1	9.8	3.8	–	–
MM-Narrator [31]	Azure API + CLIP-L14	GPT-4V	✓	✓	9.8	–	12.8	–	7.1	10.9
MM-Narrator [31]	Azure API + CLIP-L14	GPT-4	✓	✓	13.9	49.0	13.4	5.2	6.7	12.8
LLM-AD [7]	GPT-4V	–	✓	✓	20.5	–	13.5	–	–	–
AutoAD-Zero [27]	GPT-4o	GPT-4o	✓	✓	25.4	54.3	14.3	8.1	6.7	13.7
Ours	GPT-4o	GPT-4o	✓	✓	26.9	**56.4**	15.0	**8.5**	**7.4**	15.9

Table A8. **Quantitative comparison on MAD-Eval.** For training-free methods, "VLM" and "LLM" refer to the models used in separate stages, while for fine-tuned models, they denote the pre-trained components within an end-to-end model.



GT: *He holds a small pizza box.*

AutoAD-III: *He opens the door.*

AutoAD-Zero: *Spicoli takes a paper from Hand.*

AutoAD-Zero*: *He opens the door.*

Ours: *Mike Damone enters holding a pizza box.*

Ours*: *Mike holds a pizza box.*

GT: *In the warehouse, Leonard looks at a photo of a young man smiling with a woman.*

AutoAD-III: *He shows her a photo of him and his wife.*

AutoAD-Zero: *Victoria holds a photo and shows it to Leonard.*

AutoAD-Zero*: *Victoria raises her arms, looking at a photograph held by someone.*

Ours: *Leonard holds a picture, looking at it, while Victoria looks at him.*

Ours*: *Victoria raises her arms while Leonard holds and looks at the photograph.*

GT: *Rachel slowly turns back to her room.*

AutoAD-III: *Chandler looks at her.*

AutoAD-Zero: *Ross gives a thumbs up to Rachel.*

AutoAD-Zero*: *Ross gestures as Rachel looks surprised.*

Ours: *Rachel turns and walks away from Ross.*

Ours*: *Ross gestures as Rachel turns away.*

GT: *Phoebe holds the turkey open.*

AutoAD-III: *Rachel puts the turkey in the oven.*

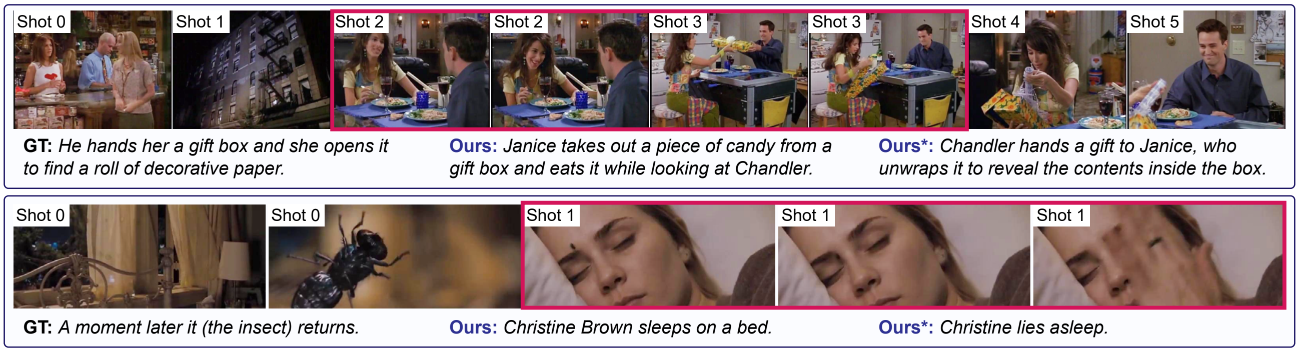
AutoAD-Zero: *Ross holds the baby, Phoebe observes, Chandler looks.*

AutoAD-Zero*: *Phoebe leans over as Ross looks attentively.*

Ours: *Phoebe holds a turkey, Chandler looks at it.*

Ours*: *Ross interacts with a large raw turkey.*

Figure A3. **More qualitative visualisations.** Current shots (corresponding to AD intervals) are outlined by red boxes for illustration purposes only. Training-free methods with "*" adopt GPT-4o (otherwise Qwen2-VL + LLaMA3). Examples from top to bottom are taken from *Fast Times at Ridgemont High* (1982), *Extraction* (2015), *Friends* (S3E2), and *Friends* (S1E9), respectively.



GT: *He hands her a gift box and she opens it to find a roll of decorative paper.*

Ours: *Janice takes out a piece of candy from a gift box and eats it while looking at Chandler.*

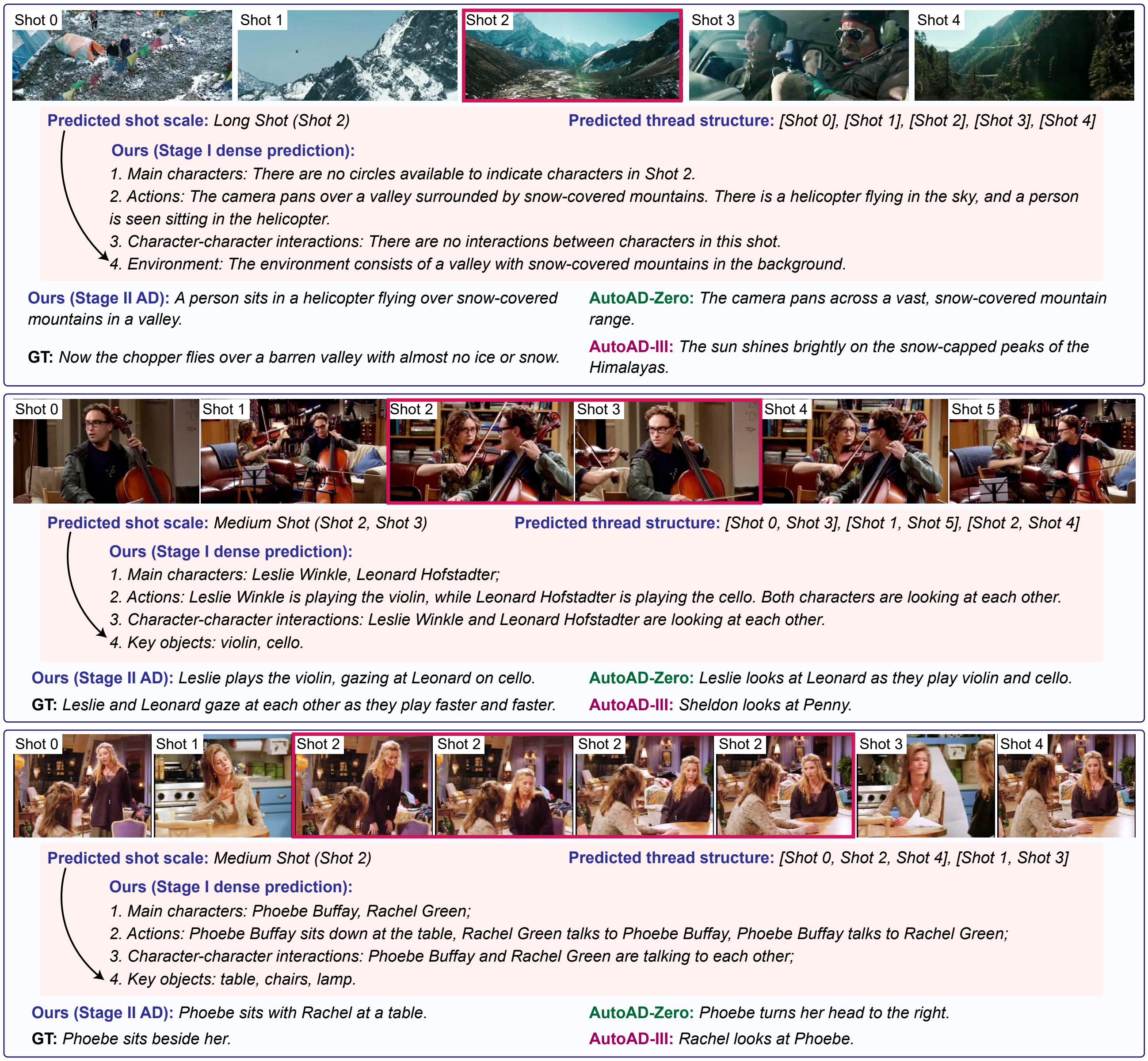
Ours*: *Chandler hands a gift to Janice, who unwraps it to reveal the contents inside the box.*

GT: *A moment later it (the insect) returns.*

Ours: *Christine Brown sleeps on a bed.*

Ours*: *Christine lies asleep.*

Figure A4. **Failure case visualisations.** Current shots (corresponding to AD intervals) are outlined by red boxes for illustration purposes only. Training-free methods with "*" adopt GPT-4o (otherwise Qwen2-VL + LLaMA3). Examples from top to bottom are taken from *Friends* (S3E4), and *Drag Me to Hell* (2009), respectively.



Predicted shot scale: *Long Shot (Shot 2)*

Predicted thread structure: *[Shot 0], [Shot 1], [Shot 2], [Shot 3], [Shot 4]*

Ours (Stage I dense prediction):

1. Main characters: There are no circles available to indicate characters in Shot 2.

2. Actions: The camera pans over a valley surrounded by snow-covered mountains. There is a helicopter flying in the sky, and a person is seen sitting in the helicopter.

3. Character-character interactions: There are no interactions between characters in this shot.

4. Environment: The environment consists of a valley with snow-covered mountains in the background.

Ours (Stage II AD): *A person sits in a helicopter flying over snow-covered mountains in a valley.*

GT: *Now the chopper flies over a barren valley with almost no ice or snow.*

AutoAD-Zero: *The camera pans across a vast, snow-covered mountain range.*

AutoAD-III: *The sun shines brightly on the snow-capped peaks of the Himalayas.*

Predicted shot scale: *Medium Shot (Shot 2, Shot 3)*

Predicted thread structure: *[Shot 0, Shot 3], [Shot 1, Shot 5], [Shot 2, Shot 4]*

Ours (Stage I dense prediction):

1. Main characters: Leslie Winkle, Leonard Hofstadter;

2. Actions: Leslie Winkle is playing the violin, while Leonard Hofstadter is playing the cello. Both characters are looking at each other.

3. Character-character interactions: Leslie Winkle and Leonard Hofstadter are looking at each other.

4. Key objects: violin, cello.

Ours (Stage II AD): *Leslie plays the violin, gazing at Leonard on cello.*

GT: *Leslie and Leonard gaze at each other as they play faster and faster.*

AutoAD-Zero: *Leslie looks at Leonard as they play violin and cello.*

AutoAD-III: *Sheldon looks at Penny.*

Predicted shot scale: *Medium Shot (Shot 2)*

Predicted thread structure: *[Shot 0, Shot 2, Shot 4], [Shot 1, Shot 3]*

Ours (Stage I dense prediction):

1. Main characters: Phoebe Buffay, Rachel Green;

2. Actions: Phoebe Buffay sits down at the table, Rachel Green talks to Phoebe Buffay, Phoebe Buffay talks to Rachel Green;

3. Character-character interactions: Phoebe Buffay and Rachel Green are talking to each other;

4. Key objects: table, chairs, lamp.

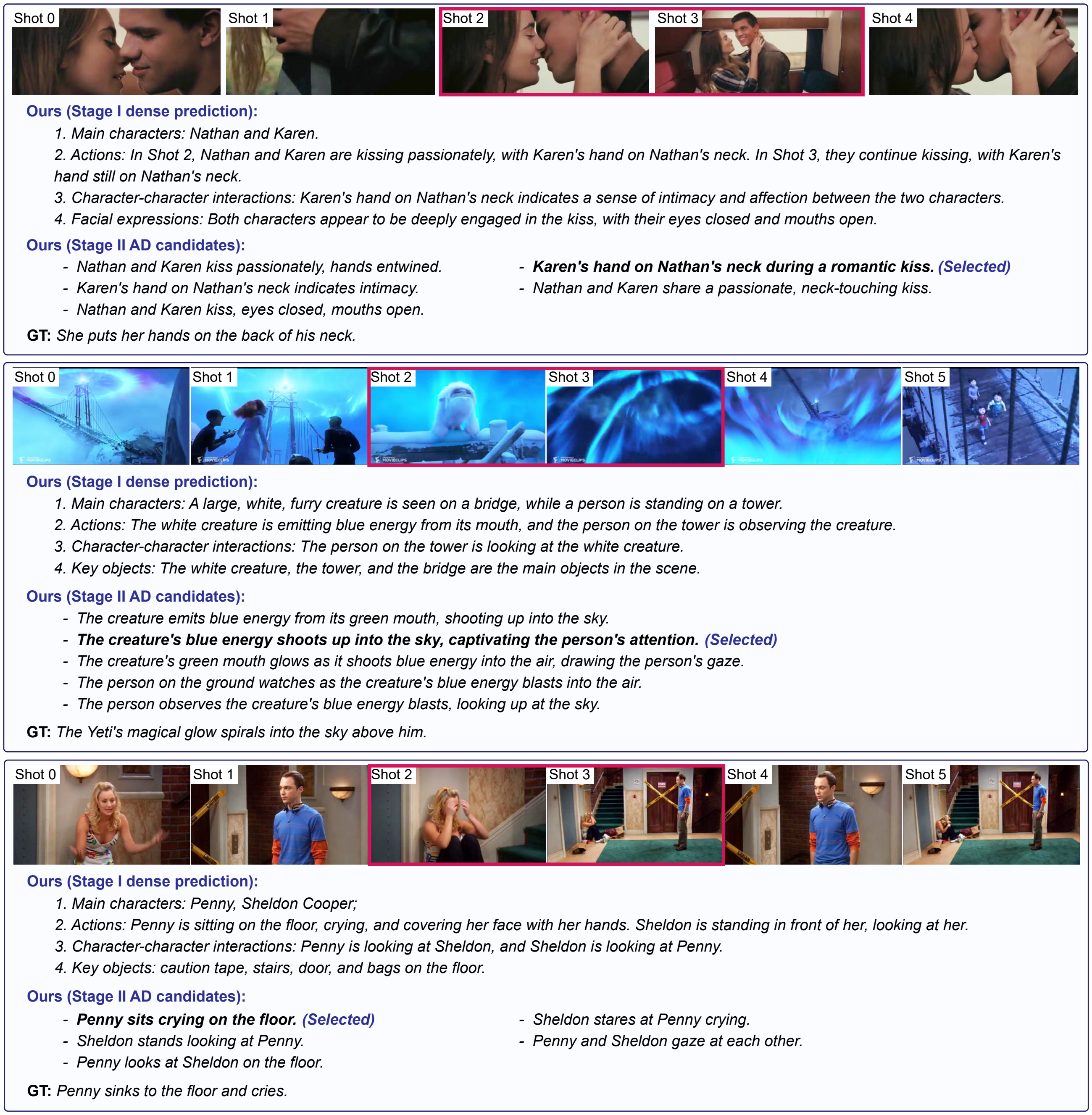
Ours (Stage II AD): *Phoebe sits with Rachel at a table.*

GT: *Phoebe sits beside her.*

AutoAD-Zero: *Phoebe turns her head to the right.*

AutoAD-III: *Rachel looks at Phoebe.*

Figure A5. **Detailed visualisations including intermediate results,** such as predicted thread structures and shot scales, as well as Stage I dense descriptions. Current shots (corresponding to AD intervals) are outlined by red boxes for illustration purposes only. Training-free methods adopt Qwen2-VL + LLaMA3 as base models. Examples from top to bottom are taken from *Everest* (2015), *The Big Bang Theory* (S1E5), and *Friends* (S1E12), respectively.



Ours (Stage I dense prediction):

1. *Main characters: Nathan and Karen.*
2. *Actions: In Shot 2, Nathan and Karen are kissing passionately, with Karen's hand on Nathan's neck. In Shot 3, they continue kissing, with Karen's hand still on Nathan's neck.*
3. *Character-character interactions: Karen's hand on Nathan's neck indicates a sense of intimacy and affection between the two characters.*
4. *Facial expressions: Both characters appear to be deeply engaged in the kiss, with their eyes closed and mouths open.*

Ours (Stage II AD candidates):

- *Nathan and Karen kiss passionately, hands entwined.*
- *Karen's hand on Nathan's neck indicates intimacy.*
- *Nathan and Karen kiss, eyes closed, mouths open.*
- ***Karen's hand on Nathan's neck during a romantic kiss.** (Selected)*
- *Nathan and Karen share a passionate, neck-touching kiss.*

GT: *She puts her hands on the back of his neck.*

Ours (Stage I dense prediction):

1. *Main characters: A large, white, furry creature is seen on a bridge, while a person is standing on a tower.*
2. *Actions: The white creature is emitting blue energy from its mouth, and the person on the tower is observing the creature.*
3. *Character-character interactions: The person on the tower is looking at the white creature.*
4. *Key objects: The white creature, the tower, and the bridge are the main objects in the scene.*

Ours (Stage II AD candidates):

- *The creature emits blue energy from its green mouth, shooting up into the sky.*
- ***The creature's blue energy shoots up into the sky, captivating the person's attention.** (Selected)*
- *The creature's green mouth glows as it shoots blue energy into the air, drawing the person's gaze.*
- *The person on the ground watches as the creature's blue energy blasts into the air.*
- *The person observes the creature's blue energy blasts, looking up at the sky.*

GT: *The Yeti's magical glow spirals into the sky above him.*

Ours (Stage I dense prediction):

1. *Main characters: Penny, Sheldon Cooper;*
2. *Actions: Penny is sitting on the floor, crying, and covering her face with her hands. Sheldon is standing in front of her, looking at her.*
3. *Character-character interactions: Penny is looking at Sheldon, and Sheldon is looking at Penny.*
4. *Key objects: caution tape, stairs, door, and bags on the floor.*

Ours (Stage II AD candidates):

- ***Penny sits crying on the floor.** (Selected)*
- *Sheldon stands looking at Penny.*
- *Penny looks at Sheldon on the floor.*
- *Sheldon stares at Penny crying.*
- *Penny and Sheldon gaze at each other.*

GT: *Penny sinks to the floor and cries.*

Figure A6. **Visualisations for assisted AD generation,** where multiple AD candidates are extracted from the dense description, and the best among the five candidates is highlighted in **bold**. Current shots (corresponding to AD intervals) are outlined with red boxes for illustration purposes only. Our method uses Qwen2-VL and LLaMA3 as base models. The examples, from top to bottom, are taken from *Abduction* (2011), *Abominable* (2019), and *The Big Bang Theory* (S2E3), respectively.

Algorithm 1 Stage I text prompt

```
# Thread information injection
if exist(thread_structure):
    # {thread_structure}: e.g., "[Shot 1, Shot 3] share the same camera setup."
    thread_structure_text = "Finally, in one sentence, briefly explain why {thread_structure}\n"
else:
    thread_structure_text = ""

# Additional factors suggested by shot scales
factor_numbers = "four"
if effect_shot_scale <= 1.5:
    additional_factor_text = "4. Describe the facial expressions of characters.\n"
elif effect_shot_scale >= 2 and effect_shot_scale <= 3:
    additional_factor_text = "4. Describe the key objects that characters interact with.\n"
elif effect_shot_scale >= 3.5:
    additional_factor_text
      = "4. Describe the environment, focusing on the location, furniture, entrances and exits, etc.\n"
else:
    additional_factor_text = ""
    factor_numbers = "three"

# Stage I prompt
# {video_type}: "movie" or "TV series"
# {key_shots}: the middle shots (e.g., "[Shot 2, Shot 3]")
# {label_type}: "circles"
# {char_text}: character information (e.g., "Possible characters: Sheldon Cooper (red), ...")
prompt = (
    "Please watch the following
       {video_type} clip, where different shot numbers are labelled on the top-left of each frame.\n"
    f"Please briefly describe what happened in {key_shots} in the {factor_numbers} steps below:\n"
    f"1. Identify main characters (if {label_type} are available){char_text};\n"
    "2. Describe the actions of characters, i.e., who is doing what, focusing on the movements;\n"
    "3. Describe the interactions between characters, such as looking;\n"
    f"{additional_factor_text}"
    f"Note, colored {label_type
      } are provided for character indications only, DO NOT mention them in the description. "
    f"{thread_structure_text}"
    "Make sure you do not hallucinate information.\n"
    "### Answer Template ###\n" # Base format
      , need to be adjusted based on additionally factors, and whether the thread structure is injected
    "Description:\n"
    "1. Main characters: '';\n"
    "2. Actions: '';\n"
    "3. Character-character interactions: ''."
)
```

Algorithm 2 Stage II text prompt

```
# Stage II system prompt
# {video_type}: "movie" or "TV series"
sys_prompt = (
    f"[INST] <<
      SYS>>\nYou are an intelligent chatbot designed for summarizing {video_type} audio descriptions. ”
    "Here’s how you can accomplish
      the task:------##INSTRUCTIONS: you should convert the predicted descriptions into one sentence. ”
    "You should directly start the answer with the converted results
       WITHOUT providing ANY more sentences at the beginning or at the end. \n<</SYS>>\n\n{} [/INST] ”
)

# Dataset dependent information
if dataset == "CMD-AD":
    verb_list = [’look’, ’turn’, ’take’, ’hold’, ’pull’, ’walk’, ’run’, ’watch’, ’stare’, ’grab
      ’, ’fall’, ’get’, ’go’, ’open’, ’smile’] # top-15 lemma verb in the corresponding training set
    speed_factor = 0.275 # averaged (duration / number of words in AD) in the training set
elif dataset == "TV-AD":
    verb_list = [’look’, ’walk’, ’turn’,
      ’stare’, ’take’, ’hold’, ’smile’, ’leave’, ’pull’, ’watch’, ’open’, ’go’, ’step’, ’get’, ’enter’]
    speed_factor = 0.2695
elif dataset == "MAD-Eval":
    verb_list = [’look’, ’turn’, ’sit
      ’, ’walk’, ’take’, ’stand’, ’watch’, ’hold’, ’pull’, ’see’, ’go’, ’open’, ’smile’, ’run’, ’get’]
    speed_factor = 0.5102

# Single AD generation / multiple AD candidate outputs (as an assistant)
if not assistant_mode: # Single AD
    pred_text = "Provide the AD from a narrator perspective.\n”
else: # Multiple ADs
    pred_text = "Provide 5 possible ADs from a narrator perspective, each offering a valid and distinct
       summary by emphasizing different key characters, actions, and movements present in the scene.\n”

# Stage II user prompt
# {text_pred}: Stage I dense description outputs
# {duration}: duration of the AD interval
# {example_sentence}: 10 randomly sampled AD sentences from training sets
user_prompt = (
    "Please summarize the
       following description for one movie clip into ONE succinct audio description (AD) sentence.\n”
    f"Description: {text_pred}\n\n”

    "Focus on the most attractive characters, their actions, and related key objects.\n”
    "For characters, use their first names, remove titles such as ’Mr.’ and ’Dr.’. If names
      are not available, use pronouns such as ’He’ and ’her’, do not use expression such as ’a man’.\n”
    "For actions, avoid mentioning the camera, and do not focus on ’talking’.\n”
    "For objects,
       especially when no characters are involved, prioritize describing concrete and specific ones.\n”
    "Do not mention characters’ mood.\n”
    "Do not hallucinate information that is not mentioned in the input.\n”
    f"Try to identify the
       following motions (with decreasing priorities): {verb_list}, and use them in the description.\n”
    "{pred_text}”
    f"Limit the length of the output within {int(duration / speed_factor + 1)} words.\n\n”

    "Output template (in JSON
       format): \"summarized_AD\": \"\".\n” # Adjust the template for single / multiple AD generation.
    "Here are some example outputs:\n”
    f"{example_sentence}”
)
```

Algorithm 3 LLM-based character information removal text prompt

```
# System prompt for LLM-based character information removal in GT ADs
sys_prompt = (
    "You are an intelligent chatbot designed for removing character information of a sentence. "
    "Here's how you can accomplish the task: "
    "You should replace all character information
       including names, roles, and jobs into pronouns (e.g., he, she, they, her, him, them). "
    "Note, objects
      , locations, and animals are not counted as character information and should be kept as-is. "
    "You should output
       the result in JSON format WITHOUT providing ANY more sentences at the beginning or at the end."
)

# User prompt for LLM-based character information removal in GT ADs
# {text_gt}: GT AD
user_prompt = (
    "Please read the sentence below that describes a video clip:\n\n"
    f"Input sentence: \"{text_gt}\"\n\n"

    "Replace all character information
       including names, roles, and jobs into pronouns (e.g., he, she, they, her, him, them).\n"
    "Note, objects
      , locations, and animals are not counted as character information and should be kept as-is.\n"

    "**Examples:**\n"
    "   - Example 1:\n"
    "     - Input sentence: \"Spicoli watches Mr. Hand pass out the schedule.\"\n"
    "     - Ouput: \"He watches him pass out the schedule.\"\n"
    "   - Example 2:\n"
    "     - Input sentence: \"Waiting
       for a reply, the inspector has a look of smug satisfaction as he combs his neat moustache.\"\n"
    "     - Output
      : \"Waiting for a reply, he has a look of smug satisfaction as he combs his neat moustache.\"\n"
    "   - Example 3:\n"
    "     - Input sentence: \"Emmerich's eyebrows twitch as he watches her.\"\n"
    "     - Output: \"His eyebrows twitch as he watches her.\"\n"
    "   - Example 4:\n"
    "     - Input sentence: \"Inside is a second pair of doors.\"\n"
    "     - Output: \"Inside is a second pair of doors.\"\n"
    "   - Example 5:\n"
    "     - Input
      sentence: \"The blonde saunters over to him in her striped pantsuit and leans in for a kiss.\"\n"
    "     - Output: \"She saunters over to him in her striped pantsuit and leans in for a kiss.\"\n"
    "..." # More examples, omitted here for simplicity

    "**Output Format:**\n"
    "{\n"
    "  \"Output\": <output>\n"
    "}\n\n"
)
```

Algorithm 4 LLM-based action sentence extraction text prompt

```
# System prompt for LLM-based action sentence extraction from GT ADs
sys_prompt = (
    "You are an intelligent chatbot designed for decompose the sentence into subsentences. "
    "Here's how you can accomplish the task: "
    "You should split (rewrite if needed
      ) the sentence into subsentences, each containing only one action phrase (i.e., verb phrase). "
    "You should output
       your answer in JSON format WITHOUT providing ANY more sentences at the beginning or at the end."
)

# User prompt for LLM-based action sentence extraction from GT ADs
# {text_gt}: GT AD after character information removal
user_prompt = (
    "Please read the sentence below that describes a video clip:\n\n"
    f"Input sentence: \"{text_gt}\"\n\n"
    "Split and rewrite the sentence into subsentences, each containing only one action (i.e.,
       verb phrase) and preserving all other information (e.g., locations, time, affections, etc.).\n"
    "Do not output repeating actions.\n"
    "**Examples:**\n"
    "   - Example 1:\n"
    "     - Input sentence: \"He watches him pass out the schedule.\"\n"
    "     - Subsentences: [\"He watches him.\", \"He passes out the schedule.\"]\n"
    "   - Example 2:\n"
    "     - Input sentence
      : \"Waiting for a reply, he has a look of smug satisfaction as he combs his neat moustache.\"\n"
    "     - Subsentences: [\"He waits
       for a reply.\", \"He has a look of smug satisfaction.\", \"He combs his neat moustache.\"]\n"
    "   - Example 3:\n"
    "     - Input sentence: \"He
      swings in front of Kingpin, then bounces off a building and kicks the criminal into the air.\"\n"
    "     - Subsentences: [\"
      He swings in front of him.\", \"He bounces off a building.\", \"He kicks him into the air.\"]\n"
    "   - Example 4:\n"
    "     - Input sentence: \"His eyebrows twitch as he watches her.\"\n"
    "     - Subsentences: [\"His eyebrows twitch.\", \"He watches her.\"]\n"
    "   - Example 5:\n"
    "     - Input sentence: \"Inside is a second pair of doors.\"\n"
    "     - Subsentences: [\"Inside is a second pair of doors.\"]\n"
    "..." # More examples, omitted here for simplicity

    "**Output Format:**\n"
    "{\n"
    "  \"Subsentences\": \n"
    "  [\n"
    "    <subsentence1>,\n"
    "    <subsentence2>,\n"
    "    <subsentence3>,\n"
    "    ...\n"
    "  ]\n"
    "}\n\n"
)
```

Algorithm 5 LLM-based action metric text prompt

```
# System prompt for LLM-based action evaluation
sys_prompt = (
    "You are an evaluation assistant designed to assess the accuracy of a description
       (Des) in capturing the action specified in a reference sentence (Ref) for a movie clip. "
    "Focus only on the presence
       of the referenced action and ignore any additional, unrelated actions in the description. "
    "Ignore any character information in the description. "
    "Avoid assumptions
      about action details beyond what is explicitly provided in either the reference or description. "
    "Output the
       result exclusively in JSON format, with a score (0 to 3) and a brief explanation describing the
       relationship between the actions in Ref and Des, without any introductory or concluding text."
)

# User prompt for LLM-based action evaluation
# {text_gt}: character-free action sentence extracted from GT AD
# {text_pred}: predicted dense description (paragraph) or AD sentence
user_prompt = (
    "You will be provided with a reference action sentence (Ref) and a description (Des) for a clip. "
    "Your task is to evaluate if the action described in Ref is explicitly or clearly implied in Des. "
    "Focus only on the presence
       of the referenced action, without considering any additional actions and character information
       that may appear in Des, and do not assume any actions beyond those stated in Ref or Des. "
    "The output should be a score (0 to 3) with a brief
       one-sentence explanation describing the relationship between the actions in Ref and Des.\n\n"

    "# Scoring Criteria:\n"
    "- **Score 0:** The action in Ref is completely unrelated to actions in Des.\n"
    "- **Score 1:** The action in Ref is loosely related to an action in Des.\n"
    "- **Score 2:** The action in Ref is similar in meaning to an action in Des.\n"
    "- **Score 3:** The action in Ref exactly matches an action in Des, using the same verb.\n\n"
    "# Examples:\n"
    "- Example 1:\n"
    "  - Ref: 'He runs across the street.'\n"
    "  - Des: 'Tom is jogging down the street.'\n"
    "  - Output: {\n"
    "      'score': 2,\n"
    "      'explanation': 'The
       action \"runs across the street\" in Ref is similar to \"jogging down the street\" in Des.'\n"
    "    }\n\n"
    "- Example 2:\n"
    "  - Ref: 'He pours wine into a glass.'\n"
    "  - Des: 'The woman drinks.'\n"
    "  - Output: {\n"
    "      'score': 1,\n"
    "      'explanation
      ': 'The action \"pours wine into a glass\" in Ref is loosely related to \"drinks\" in Des.'\n"
    "    }\n\n"
    "..." # More examples, omitted here for simplicity
    "# Output Format:\n"
    "{\n"
    "  'score': <score>,\n"
    "  'explanation': '<explanation>'\n"
    "}\n\n"
    "# Now, apply these instructions to the following texts:\n\n"
    f"   - # Reference (Ref): '{text_gt}'\n"
    f"   - # Description (Des): '{text_pred}'"
)
```

References

[1] Vlog: Video as a long document. https://github.com/showlab/VLog, 2023. 5

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016. 1, 4

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[4] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005. 1, 4

[5] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 2

[6] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 1

[7] Peng Chu, Jiang Wang, and Andre Abrantes. LLM-AD: Large language model based audio description system. *arXiv preprint arXiv:2405.00983*, 2024. 5

[8] Bo Fang, Wenhao Wu, Qiangqiang Wu, Yuxin Song, and Antoni B. Chan. DistinctAD: Distinctive audio description generation in contexts. *arXiv preprint arXiv:2411.18180*, 2024. 1, 5

[9] Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. Autoad: Movie description in context. In *CVPR*, 2023. 1, 5

[10] Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. AutoAD II: The sequel – who, when, and what in movie audio description. In *ICCV*, 2023. 5

[11] Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. AutoAD III: The prequel – back to the pixels. In *CVPR*, 2024. 1, 3, 5

[12] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023. 1

[13] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004. 1, 4

[14] Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, Ce Liu, and Lijuan Wang. MM-Vid: Advancing video understanding with GPT-4V(ision). *arXiv preprint arXiv:2310.19773*, 2023. 5

[15] Kevin Qinghong Lin, Pengchuan Zhang, Difei Gao, Xide Xia, Joya Chen, Ziteng Gao, Jinheng Xie, Xuhong Xiao, and Mike Zheng Shou. Learning video context as interleaved multimodal sequences. In *ECCV*, 2024. 5

[16] Meta. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1, 2, 4

[17] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 5

[18] David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected clip. In *ENNLP*, 2022. 5

[19] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024. 1, 2, 4

[20] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 1

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 1, 4

[22] Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of Documentation*, 2004. 1

[23] Gemma Team. Gema 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 2

[24] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, 2015. 1, 4

[25] Hanlin Wang, Zhan Tong, Kecheng Zheng, Yujun Shen, and Limin Wang. Contextual ad narration with interleaved multimodal sequence. *arXiv preprint arXiv:2403.12922*, 2024. 5

[26] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2

[27] Junyu Xie, Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. Autoad-zero: A training-free framework for zero-shot audio description. In *ACCV*, 2024. 5

[28] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 2

[29] Keunwoo Peter Yu. Videoblip, 2023. 5

[30] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang

Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 2

[31] Chaoyi Zhang, Kevin Lin, Zhengyuan Yang, Jianfeng Wang, Linjie Li, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. MM-Narrator: Narrating long-form videos with multimodal in-context learning. In *CVPR*, 2024. 5

[32] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *ENNLP*, 2023. 5

[33] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *ICLR*, 2020. 4