DistinctAD: Distinctive Audio Description Generation in Contexts

Supplementary Material

A. Analysis of AD Reconstruction with CLIP Embedding Space

As detailed in the main paper's §3.1, our Stage-I strategy, **CLIP-AD adaption**, is inspired by a preliminary AD reconstruction experiment using the CLIP text encoder and GPT-2. We begin with the question: *is the CLIP text embedding space expressive enough for embedded AD words to be reconstructed by LLMs?* If the reconstruction process is successful—meaning that LLMs can understand the textual ADs encoded by the CLIP text encoder—then the misalignment in the VLM joint feature space likely occurs because of the CLIP vision encoder, rather than between the CLIP text encoder and the LLMs. On the other hand, if the reconstruction is not successful, then the pre-trained CLIP joint embedding space is not suitable for the AD task, and both text and vision encoders need to be retrained.

To address this question, we design the AD words reconstruction pipeline illustrated in Fig. A.1. Specifically, we input the AD sentence into a frozen CLIP text encoder, modified to output tokens for each word. We implement two versions of AD reconstruction: 1) using **only a single `[CLS]` vector**, or 2) using **all word tokens** as prompts. We append a `<BOS>` tag to signal the start of reconstruction. The output embeddings are then fed into a learnable single-layer projector, transforming the CLIP word tokens into the LLM embedding space. We apply an auto-regression loss identical to (11) in the main paper, with the visual `prompt` setting as none. The projector is trained for 10 epochs on MAD-v2-Named [68] ADs, and the performance is evaluated using classical n-gram based metrics on the MAD-Eval benchmark [21]. The reconstruction results are presented in Tab. A.1. Remarkably, by merely fine-tuning a single-layer projector, AD reconstruction achieves results closely aligned with the ground truth, such as scores of **80.8** on **BLEU1** and **612.5** on **CIDEr** with all words input. Additionally, using only a single `[CLS]` vector to recover the entire AD achieves 92.2 on CIDEr, *significantly* outperforming existing AD works, which score ∼20 CIDEr. This shows that AD words (or [CLS] vector) encoded by the CLIP text encoder can be effectively understood by LLMs, suggesting that the misalignment mainly lies within the joint VLM feature space, *i.e.*, discrepancies between CLIP vision embeddings and CLIP AD embeddings.

B. Analysis of Contextual Features

In this section, we validate our primary hypothesis: *sequential clips from an extended video often share redundant scenes or characters, resulting in similar visual features within contexts,* as discussed in §3.2 of the main paper. Fig. B.2 presents the cosine similarity matrix for neighboring (contextual) movie clips (left) and their corresponding audio descriptions (ADs) (right) from four randomly selected films. The visual clip features are derived through mean pooling over T frame embeddings encoded by the CLIP Vision encoder, while the AD features are obtained from the `[CLS]` embeddings encoded by the CLIP Text encoder. From these similarity matrices, we observe two key points: (i) Movie clips generally exhibit greater similarity to each other compared to ADs, indicated by a higher proportion of red (deep) colors; (ii) Compared to ADs, neighboring (contextual) movie clips show prominent areas of similarity around the diagonals (i.e., the block diagonal structure), demonstrating that they share similar visual features due to recurring scenes and characters.

In Fig. B.2, middle column, we illustrate the similarity of neighboring movie clips using our adapted CLIP_{AD} vision encoder in Stage-I (see §3.1 of the main paper). Significant changes compared to *vanilla CLIP* visualizations are highlighted with green rectangles. Our CLIP_{AD} helps reduce redundancy among neighboring video clips, as evidenced by the smaller similarity values within the green rectangles, which helps to improve the generation of distinctive ADs in our framework. This further demonstrates the effectiveness of our Stage-I strategy.

C. Detailed Formulation of CrossAttention

In this part, we provide an in-depth explanation of the Cross-Attention formulation, building upon (9) in the main paper. The query Q originates from the Perceiver output, denoted as $\mathcal{H}$, while both the key K and the value V are derived from the base matrix $\mathcal{M}$. We apply three Linear layers to transform the query, key, and value into a unified embedding space, as represented by the following equations:

$$Q = \mathcal{H}W_Q^T + b_Q, \tag{13}$$

$$K = \mathcal{M}W_K^T + b_K, \tag{14}$$

$$V = \mathcal{M}W_Q^T + b_V. \tag{15}$$

Subsequently, the cross-attention mechanism is formulated by computing a weighted sum of the values, where the weights are determined by the similarity between the queries and keys. The softmax function ensures the normalization of the attention weights. The final cross-attention output $\widetilde{\mathcal{H}}$ is given by:

$$\widetilde{\mathcal{H}} = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{16}$$

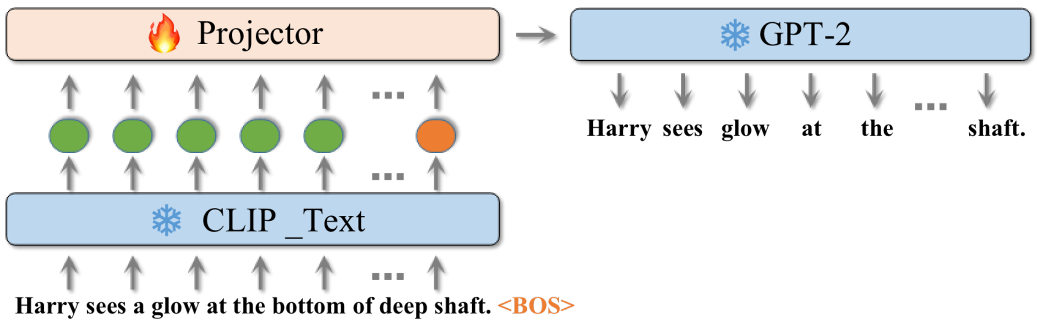

Projector
GPT-2
Harry sees glow at the ... shaft.
CLIP _Text
Harry sees a glow at the bottom of deep shaft. <BOS>

Figure A.1. Reconstructing AD words by merely fine-tuning a single-layer projector between a frozen CLIP text encoder and GPT-2.

Projector input	**(V)LM**	**LLM**	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr	SPICE
`[CLS]`	CLIP-Text	GPT-2	29.3	16.4	9.2	5.1	13.2	29.4	92.2	19.4
Words	CLIP-Text	GPT-2	80.8	74.4	68.4	63.0	47.4	82.4	612.5	66.4

Table A.1. AD reconstruction results on MAD-Eval benchmark. Only textual modality ADs in MAD-Eval are utilized for evaluation, with no movie frames involved. `[CLS]` denotes using only **one** class token vector to reconstruct the entire AD.

Ex#	$\alpha\widehat{\mathcal{H}}$	$\beta\widetilde{\mathcal{H}}$	$\mathcal{L}_{dist}$	CIDEr	R@5/16
A0	✗	✗	✗	25.2	52.3
B1	✓	✗	✗	26.1	54.1
B2	✗	✓	✗	26.4	54.5
B3	✓	✓	✗	26.0	55.5
C0	✗	✗	✓	27.0	55.9
C1	✓	✗	✓	26.8	55.7
C2	✗	✓	✓	27.2	55.9
C3	✓	✓	✓	**27.3**	**56.0**

Table D.2. Ablation studies with LLaMA3-8B in Stage-II.

where $\sqrt{d_k}$ acts as a scaling factor to stabilize the gradient flow during training.

D. Ablations with Strongest Settings

As a complement to the ablation study in Tab. 5 in the main paper, we further conduct Stage-II ablations using our strongest settings, *i.e.* CLIP-AD-B16 and LLaMA3-8B models. As shown in Tab. D.2, the performance remains generally consistent, leading to similar conclusions as those obtained with default CLIP-AD-B32 and GPT-2 settings.

E. Additional Qualitative Examples

Following Fig. 5 in the main paper, we present additional qualitative examples in Fig. E.4, utilizing our adapted CLIP-AD-B16 [58] in Stage-I and LLM LLaMA3-8B [4]. The movie clips are **consecutively** sampled from the following films: **(a)** *Signs* (2002), **(b)** *The Roommate* (2011), and **(c)** *How Do You Know* (2010), listed from top to bottom. For accurate retrieval and alignment, the starting time of each movie clip is indicated in the top-left corner of each clip. Additionally, we provide results from the publicly available AutoAD-Zero [82] for comparison. The numerous high-quality examples further demonstrate the superiority of our proposed method, DistinctAD.

Since complete predictions and codes are *unavailable* for many previous methods, such as AutoAD-I, AutoAD-II, AutoAD-III, and MM-Narrator, we only collect the qualitative examples presented in their original papers and perform qualitative comparisons in Fig. E.3. Training-free methods are highlighted with a blue background, while partial-fine-tuning methods are marked in orange. It is evident that training-free methods utilizing proprietary models like GPT-4 or GPT-4V often encounter hallucination issues, producing irrelevant or imaginary details. In contrast, partial-fine-tuning methods, *i.e.* AutoAD-I, AutoAD-II and DistinctAD, generate more accurate ADs close to human-annotated ground-truth. (We use past 3 *ground-truth* ADs as AutoAD-I's textual prompts.) Despite this, AutoAD-I can be negatively influenced by its contextual content, *e.g.* "nuns" mistakenly appears in **(d)**. AutoAD-II tends to generates similar AD words, *e.g.* *"furrowed brow"* for movie frames with close-up faces in **(a)** and **(d)**, whereas our DistinctAD is generally more distinctive.

F. Raw Frames of MAD

Due to copyright restrictions, MAD [68] only provides frame-level movie features extracted by CLIP [58]. However, to facilitate CLIP-AD adaptation in Stage-I, we require raw MAD movie frames to fine-tune the CLIP vision encoder. To achieve this, we collect MAD raw movies from third-party platforms such as Amazon Prime Video. Out of the 488 movies in the MAD-train list, 3 are not available online, as shown in Tab. F.2.

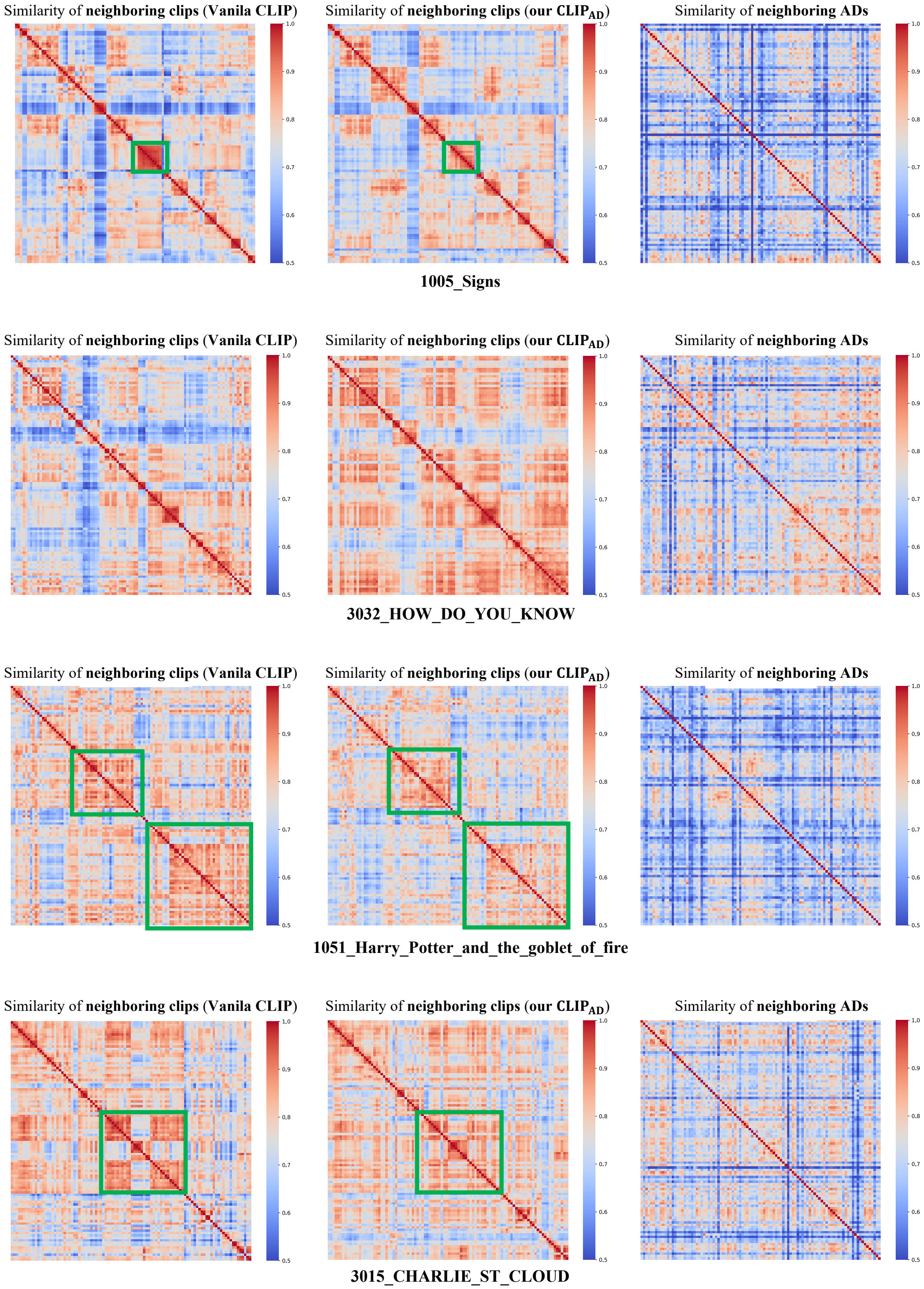
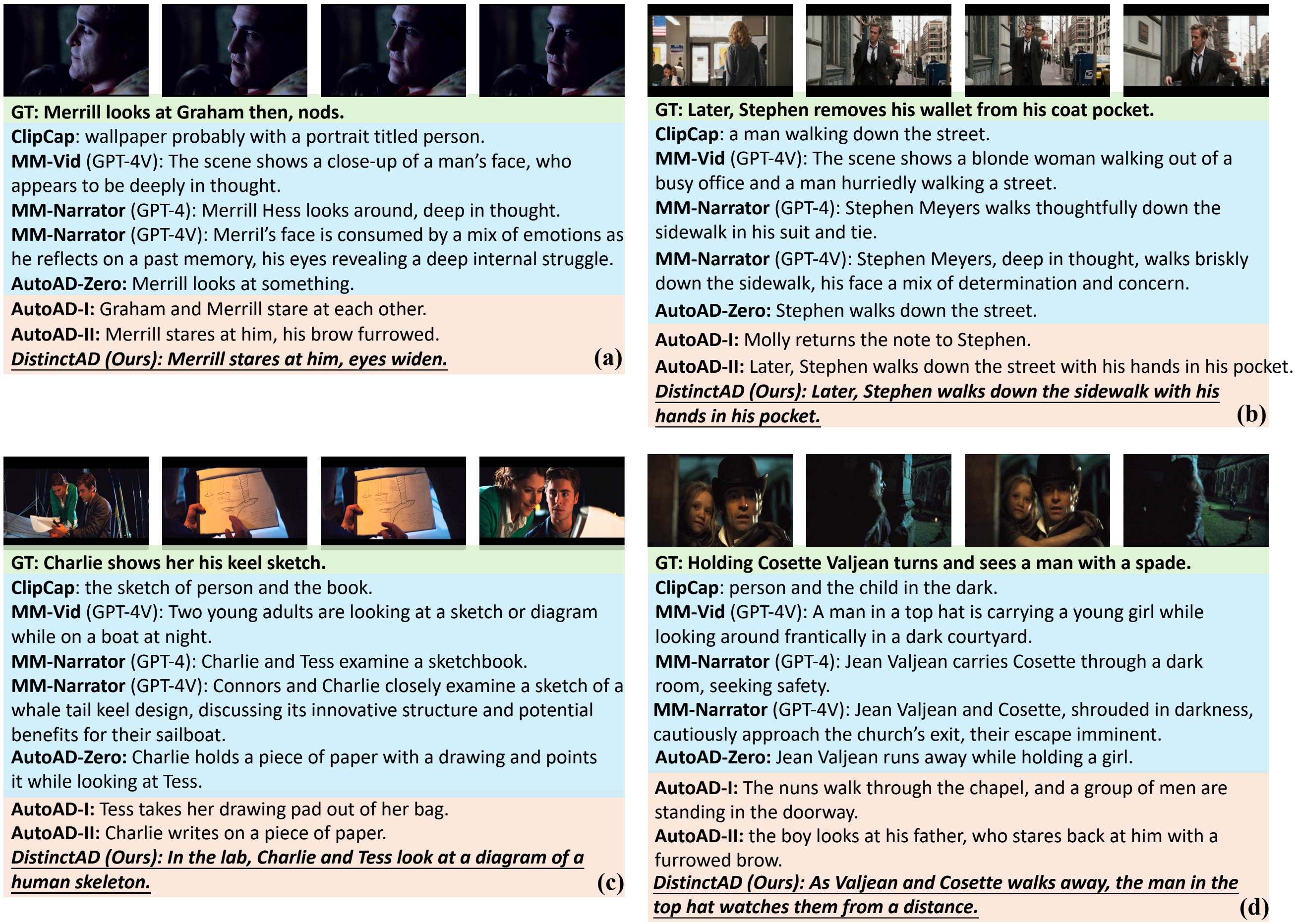


Figure B.2. Cosine similarity matrices of neighboring (contextual) movie clips using vanilla CLIP (left) and our adapted CLIP$_{\text{AD}}$ in Stage-I (middle). We also show similarity matrices of corresponding neighboring ADs (right). Movie clips are from Signs (2002), How Do You Know (2010), Harry Potter and the Goblet of Fire (2005), and Charlie St. Cloud (2010). Green boxes indicate differences between vanilla CLIP and our CLIP-AD. Zoom in for details.



GT: Merrill looks at Graham then, nods.
ClipCap: wallpaper probably with a portrait titled person.
MM-Vid (GPT-4V): The scene shows a close-up of a man's face, who appears to be deeply in thought.
MM-Narrator (GPT-4): Merrill Hess looks around, deep in thought.
MM-Narrator (GPT-4V): Merril's face is consumed by a mix of emotions as he reflects on a past memory, his eyes revealing a deep internal struggle.
AutoAD-Zero: Merrill looks at something.
AutoAD-I: Graham and Merrill stare at each other.
AutoAD-II: Merrill stares at him, his brow furrowed.
DistinctAD (Ours): Merrill stares at him, eyes widen. **(a)**

GT: Later, Stephen removes his wallet from his coat pocket.
ClipCap: a man walking down the street.
MM-Vid (GPT-4V): The scene shows a blonde woman walking out of a busy office and a man hurriedly walking a street.
MM-Narrator (GPT-4): Stephen Meyers walks thoughtfully down the sidewalk in his suit and tie.
MM-Narrator (GPT-4V): Stephen Meyers, deep in thought, walks briskly down the sidewalk, his face a mix of determination and concern.
AutoAD-Zero: Stephen walks down the street.
AutoAD-I: Molly returns the note to Stephen.
AutoAD-II: Later, Stephen walks down the street with his hands in his pocket.
DistinctAD (Ours): Later, Stephen walks down the sidewalk with his hands in his pocket. **(b)**

GT: Charlie shows her his keel sketch.
ClipCap: the sketch of person and the book.
MM-Vid (GPT-4V): Two young adults are looking at a sketch or diagram while on a boat at night.
MM-Narrator (GPT-4): Charlie and Tess examine a sketchbook.
MM-Narrator (GPT-4V): Connors and Charlie closely examine a sketch of a whale tail keel design, discussing its innovative structure and potential benefits for their sailboat.
AutoAD-Zero: Charlie holds a piece of paper with a drawing and points it while looking at Tess.
AutoAD-I: Tess takes her drawing pad out of her bag.
AutoAD-II: Charlie writes on a piece of paper.
DistinctAD (Ours): In the lab, Charlie and Tess look at a diagram of a human skeleton. **(c)**

GT: Holding Cosette Valjean turns and sees a man with a spade.
ClipCap: person and the child in the dark.
MM-Vid (GPT-4V): A man in a top hat is carrying a young girl while looking around frantically in a dark courtyard.
MM-Narrator (GPT-4): Jean Valjean carries Cosette through a dark room, seeking safety.
MM-Narrator (GPT-4V): Jean Valjean and Cosette, shrouded in darkness, cautiously approach the church's exit, their escape imminent.
AutoAD-Zero: Jean Valjean runs away while holding a girl.
AutoAD-I: The nuns walk through the chapel, and a group of men are standing in the doorway.
AutoAD-II: the boy looks at his father, who stares back at him with a furrowed brow.
DistinctAD (Ours): As Valjean and Cosette walks away, the man in the top hat watches them from a distance. **(d)**

Figure E.3. Qualitative comparisons on **single** movie clips between ClipCap, MM-Vid, MM-Narrator, AutoAD-Zero, AutoAD-I, AutoAD-II, and our DistinctAD. The movies are from **(a)** Signs (2002), **(b)** Ides of March (2011), **(c)** Charlie St. Cloud (2010), and **(d)** Les Miserables (2012). Zoom in for details.

MAD_ID	IMDB_ID	Movie Title
4797	tt0395571	*Holy Flying Circus*
4839	tt4846340	*Halo: The Fall of Reach*
5900	tt0408306	*Murdered by My Father*

Table F.2. Meta information of missing films in MAD-train.

Moreover, due to geographical differences, we may download different versions of movies, potentially leading to mismatches between movie clips and annotated timestamps. To address this, we conduct a thorough check by comparing our downloaded movies with the MAD dataset and their metadata in the IMDB database. Out of 488 movies, 9 have time durations that vary more than one minute. Details are shown in Tab. F.3.

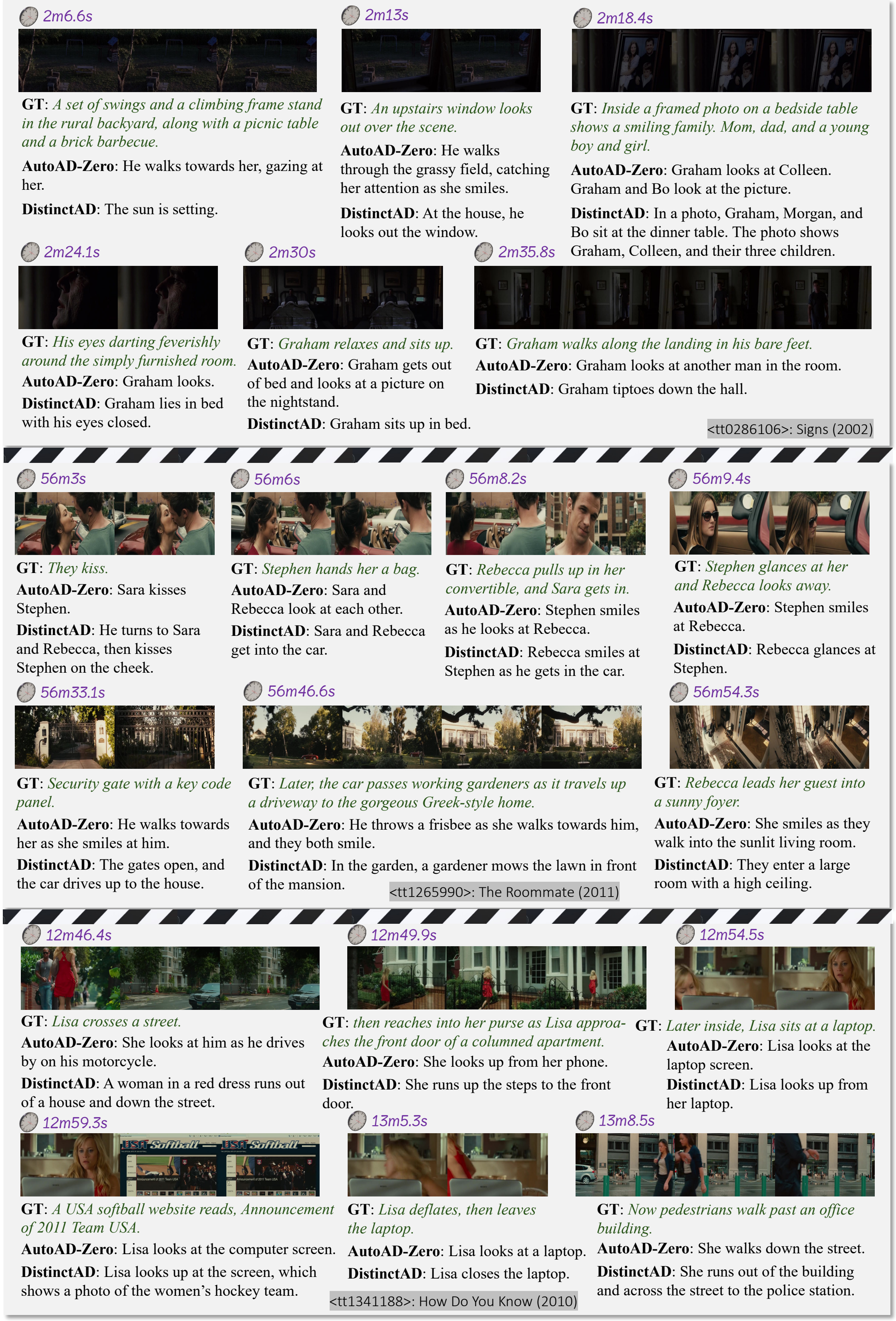
MAD_ID	IMDB_ID	MAD_Time	Our_Time	IMDB_Time
2738	tt0450232	1h 37m 26s	1h 41m 59s	1h 42m
2787	tt1136608	1h 19m 24s	1h 52m 16s	1h 52m
4017	tt5463162	1h 59m 20s	1h 57m 41s	1h 59m
4061	tt1837636	1h 28m 2s	2h 8m 12s	2h 8m
4266	tt0375735	1h 36m 8s	1h 40m 39s	1h 40m
4772	tt0424136	1h 39m 53s	1h 44m 33s	1h 44m
4902	tt0119310	1h 15m 30s	1h 11m 55s	1h 14m
5634	tt2929690	1h 40m 52s	1h 51m 50s	1h 40m
6952	tt2527338	2h 31m 52s	2h 21m 53s	2h 21m

Table F.3. Metadata for movies with duration difference exceeding 1 minute. Durations closer to the IMDB are highlighted in green.

MAD-v2-Named	# movies	# AD
MAD-Train-Features [68]	488	334,296
MAD-Train-Frames (Ours)	482	326,632

Table F.4. Statistics of our refined MAD dataset with raw frames.

According to the statistical information in Tab. F.3, we identify potential temporal misalignment noise in the existing MAD benchmark. To mitigate negative impacts during training, we exclude movies with durations that significantly differ from those in the IMDB database. The removed movie IDs are: 4017, 4902, 5634. A summary of the final employed MAD-v2-Named training dataset is provided in Tab. F.4.



2m6.6s

GT: *A set of swings and a climbing frame stand in the rural backyard, along with a picnic table and a brick barbecue.*

AutoAD-Zero: He walks towards her, gazing at her.

DistinctAD: The sun is setting.

2m13s

GT: *An upstairs window looks out over the scene.*

AutoAD-Zero: He walks through the grassy field, catching her attention as she smiles.

DistinctAD: At the house, he looks out the window.

2m18.4s

GT: *Inside a framed photo on a bedside table shows a smiling family. Mom, dad, and a young boy and girl.*

AutoAD-Zero: Graham looks at Colleen. Graham and Bo look at the picture.

DistinctAD: In a photo, Graham, Morgan, and Bo sit at the dinner table. The photo shows Graham, Colleen, and their three children.

2m24.1s

GT: *His eyes darting feverishly around the simply furnished room.*

AutoAD-Zero: Graham looks.

DistinctAD: Graham lies in bed with his eyes closed.

2m30s

GT: *Graham relaxes and sits up.*

AutoAD-Zero: Graham gets out of bed and looks at a picture on the nightstand.

DistinctAD: Graham sits up in bed.

2m35.8s

GT: *Graham walks along the landing in his bare feet.*

AutoAD-Zero: Graham looks at another man in the room.

DistinctAD: Graham tiptoes down the hall.

<tt0286106>: Signs (2002)

56m3s

GT: *They kiss.*

AutoAD-Zero: Sara kisses Stephen.

DistinctAD: He turns to Sara and Rebecca, then kisses Stephen on the cheek.

56m6s

GT: *Stephen hands her a bag.*

AutoAD-Zero: Sara and Rebecca look at each other.

DistinctAD: Sara and Rebecca get into the car.

56m8.2s

GT: *Rebecca pulls up in her convertible, and Sara gets in.*

AutoAD-Zero: Stephen smiles as he looks at Rebecca.

DistinctAD: Rebecca smiles at Stephen as he gets in the car.

56m9.4s

GT: *Stephen glances at her and Rebecca looks away.*

AutoAD-Zero: Stephen smiles at Rebecca.

DistinctAD: Rebecca glances at Stephen.

56m33.1s

GT: *Security gate with a key code panel.*

AutoAD-Zero: He walks towards her as she smiles at him.

DistinctAD: The gates open, and the car drives up to the house.

56m46.6s

GT: *Later, the car passes working gardeners as it travels up a driveway to the gorgeous Greek-style home.*

AutoAD-Zero: He throws a frisbee as she walks towards him, and they both smile.

DistinctAD: In the garden, a gardener mows the lawn in front of the mansion.

56m54.3s

GT: *Rebecca leads her guest into a sunny foyer.*

AutoAD-Zero: She smiles as they walk into the sunlit living room.

DistinctAD: They enter a large room with a high ceiling.

<tt1265990>: The Roommate (2011)

12m46.4s

GT: *Lisa crosses a street.*

AutoAD-Zero: She looks at him as he drives by on his motorcycle.

DistinctAD: A woman in a red dress runs out of a house and down the street.

12m49.9s

GT: *then reaches into her purse as Lisa approaches the front door of a columned apartment.*

AutoAD-Zero: She looks up from her phone.

DistinctAD: She runs up the steps to the front door.

12m54.5s

GT: *Later inside, Lisa sits at a laptop.*

AutoAD-Zero: Lisa looks at the laptop screen.

DistinctAD: Lisa looks up from her laptop.

12m59.3s

GT: *A USA softball website reads, Announcement of 2011 Team USA.*

AutoAD-Zero: Lisa looks at the computer screen.

DistinctAD: Lisa looks up at the screen, which shows a photo of the women's hockey team.

13m5.3s

GT: *Lisa deflates, then leaves the laptop.*

AutoAD-Zero: Lisa looks at a laptop.

DistinctAD: Lisa closes the laptop.

13m8.5s

GT: *Now pedestrians walk past an office building.*

AutoAD-Zero: She walks down the street.

DistinctAD: She runs out of the building and across the street to the police station.

<tt1341188>: How Do You Know (2010)

Figure E.4. More qualitative results on **consecutive** movie clips. Movie frames from top to bottom are taken from Signs (2002), The Roommate (2011), How Do You Know (2010), respectively. Zoom in for details.