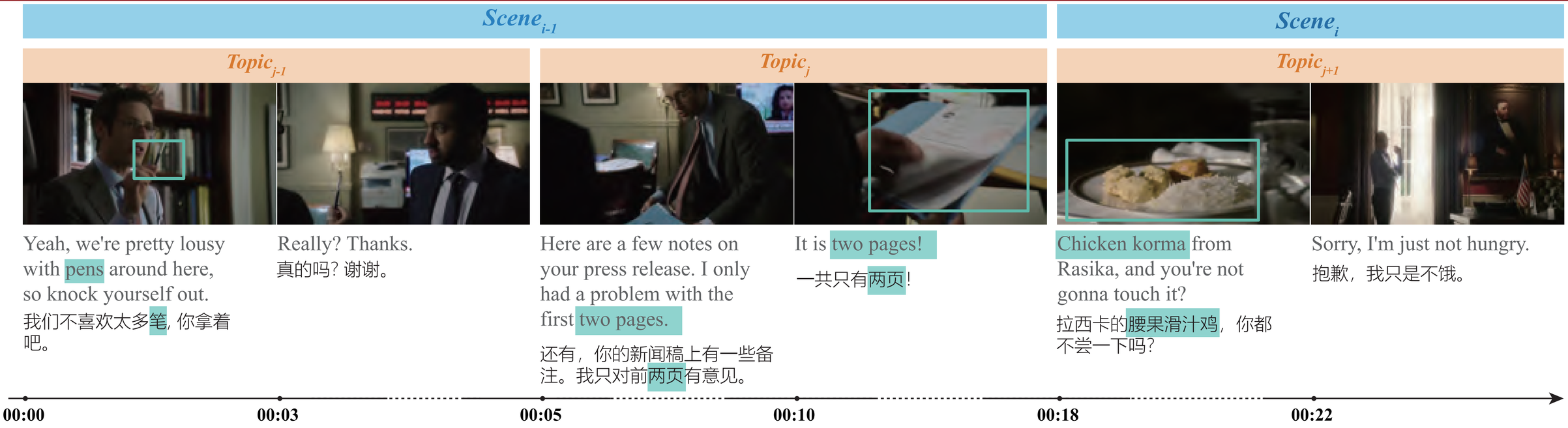
VSTAR

A Video-grounded Dialogue Dataset for Situated Semantic Understanding with Scene and Topic Transitions

Yuxuan Wang, Zilong Zheng†, Xueliang Zhao, Jinpeng Li, Yueqian Wang, Dongyan Zhao†

Peking University, Beijing Institute for General Artificial Intelligence (BIGAI)

https://VSTAR-benchmark.github.io



ONE Large-scale VSTAR Dataset

The **VSTAR** is collected from 395 TV series (8,159 episodes, 185K 90-second clips) with carefully cleaned dialogues and metadata information.

Dataset	Vision	Language	Scene	Topic	# Dialogues	# Turns	Turns/Clip	Words/Turn
VisualDialog	Image	QA	✗	✗	120K	2.4M	20.0	4.0
Twitch-FIFA	Live Video	Dialogue	✗	✗	15K	161K	10.4	6.0
AVSD	Recorded Video	Dialogue	✗	✗	11K	118K	20.0	9.5
MoiveNet†	Movies	Dialogue	✓	✗	-	421K	-	7.2
OpenViDial 2.0	Movies & TV Series	Dialogue	✗	✗	116K	5.6M	48.0	8.3
VSTAR (Ours)	TV Series	Dialogue	✓	✓	185K	4.6M	25.1	6.7

†: We compute statistics of the sub-dataset with scene boundary annotations.

Table 1. Comparisons of different multimodal dialogue datasets.

We crawled metadata for each episode from IMDb as complementary information. Each TV episode is paired with genres, keywords, and storylines.

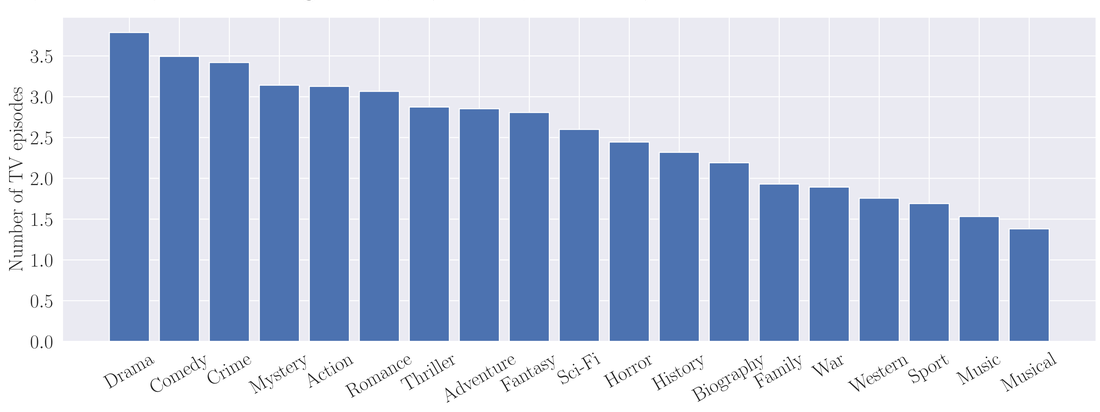


Figure 1. Distribution of **VSTAR** Genres (y-axis in log scale).

We segment the TV episode into short videos by subtitle timeline. Specifically, each short video is paired with a dialogue turn. Annotators are then asked to look through these short videos with subtitles and find if the short video is the start of a dialogue scene. To make full use of the multi-modal information, we perform dialogue topic boundary annotation and dialogue scene boundary annotation at the same time.

	Scene	Video	Source
OVSD	300	21	MiniFilm
BBC	670	11	Documentary
MovieScenes	21k	150	Movie
MovieNet	42k	318	Movie
VSTAR (Ours)	265k	8159	TV Episode

Table 2. Comparisons of dialogue scene annotation in **VSTAR**.

	Sentence	Sent/Seg	Language
DiaSeg_711	19K	5.6	English
Doc2Dial	19K	3.5	English
ZYS	12K	6.4	Chinese
VSTAR (Ours)	4.6M	9.3	English

Table 3. Comparisons of dialogue topic annotations.

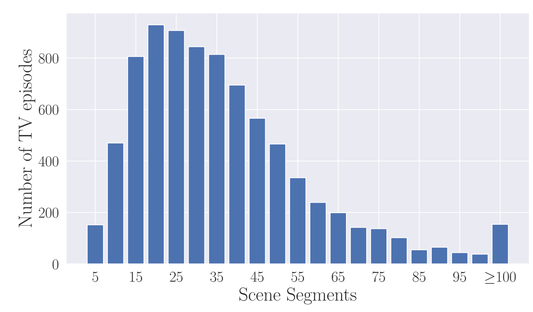


Figure 2. Distribution of # scene segments

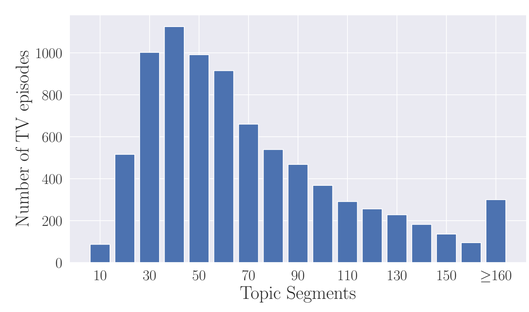


Figure 3. Distribution of # topic segments

THREE Benchmark Tasks

1. **Dialogue Scene Segmentation** A dialogue scene segment is a series of video-grounded dialogue pieces that sharing the same visual scene context. We thereby formulate the dialogue scene segmentation task as a binary classification problem: given a clip $(U, V) = \{(u_i, v_i), i = 1...K\}$, the model is asked to predict $s_i \in \{0, 1\}$ indicating the dialogue scene boundary.
2. **Dialogue Topic Segmentation** Similar to scene segmentation, we formulate dialogue topic segmentation as a turn-level classification problem. Concretely, given a video-grounded dialogue clip (U, V), we need to predict if the i-th dialogue utterance is the end of a dialogue topic.
3. **Dialogue response generation** For each dialogue clip (U, V), we set the first $N-1$ dialogue turns $\{u_1, \ldots, u_{N-1}\}$ as the dialogue context C and the last dialogue turn u_N as the gold reply r.

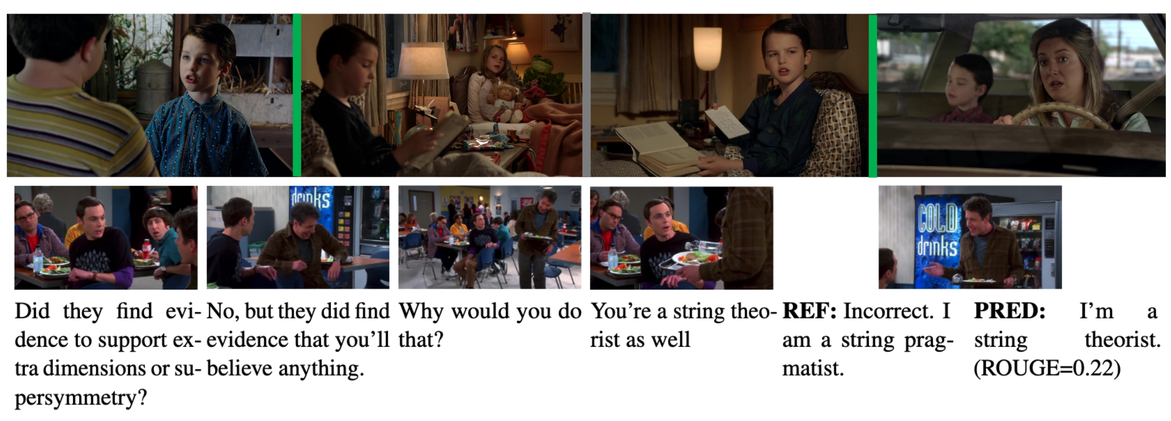


Figure 4. Cases for segmentation and generation tasks. The green bars indicate scene transition, and the grey bars indicate topic transition. **PRED** denotes our predicted response, **REF** denotes the reference human response.

TWO Transformer-based Video-Language Models

We propose a transformer-based discriminative model, namely SWST, to benchmark two segmentation tasks. For the response generation task, we develop a transformer-based generative model following encoder-decoder framework (AVDT)

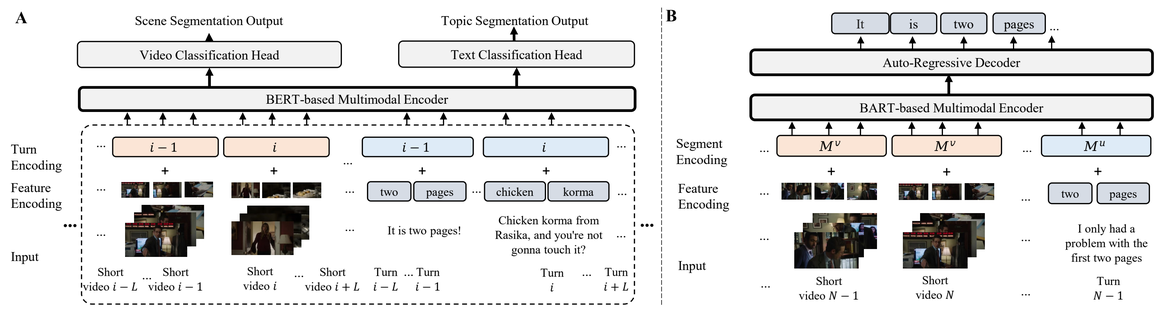


Figure 5. **Transformer-based model architecture.** (A) Sliding-window based segmentation transformer (SWST) for scene and topic boundaries identification. The dashed rectangle indicates the current sliding window for turn i. (B) Autoregressive video-grounded dialogue transformer (AVDT) for dialogue response generation.

We find several insightful results:

- SWST is able to learn scene transitions from the text-only input, which indicates that scene transitions can not be ignored in dialogues.
- Text information is very helpful to video scene segmentation, and vice versa.
- For dialogue generation, the boost from video is trivial with traditional fusion methods. But with the help of transition information, the performance improves significantly.
- Different visual backbones demonstrate similar performance and the ViT-based patch features even perform slightly better than other offline extracted features with high computational overload.