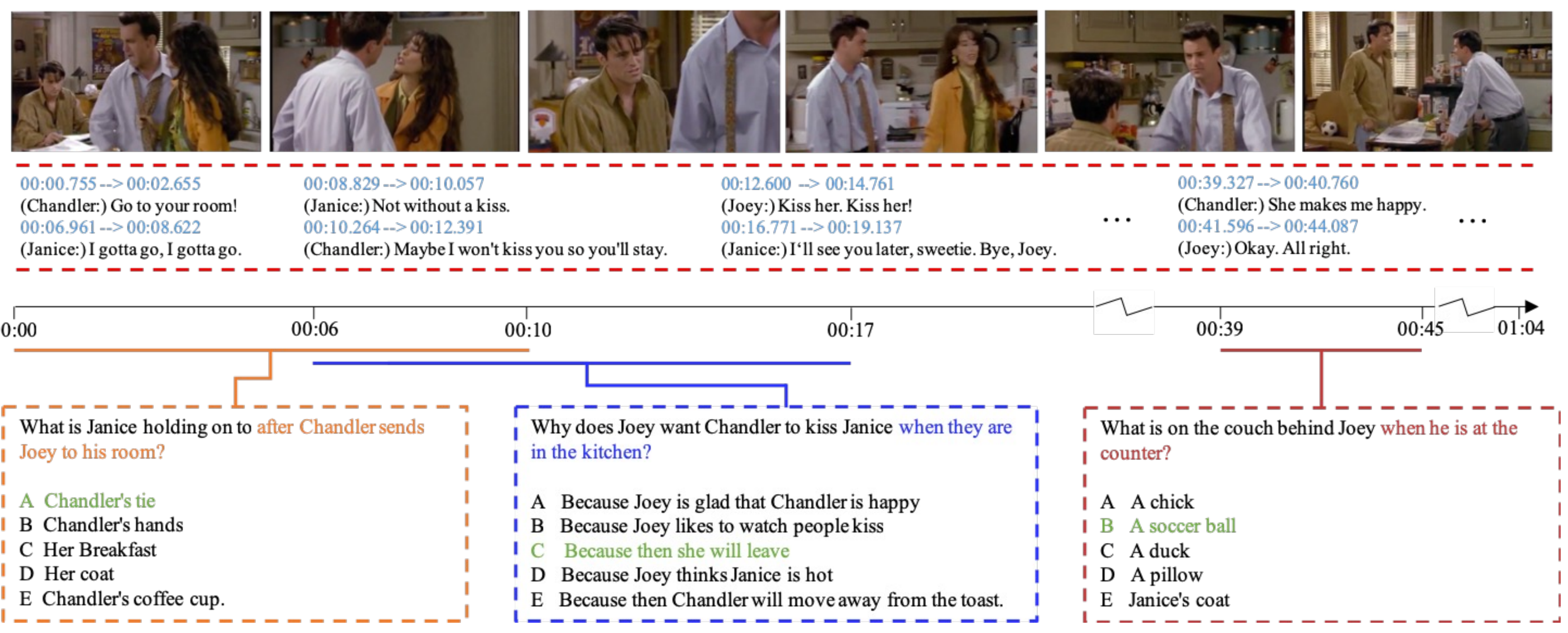
# Self-Adaptive Sampling for Efficient Video Question-Answering on Image--Text Models

Wei Han, Hui Chen, Min-Yen Kan, Soujanya Poria

*Singapore University of Technology & Design, Singapore*
*National University of Singapore, Singapore*

## Introduction

- Video Question Answering (ViQA): Given a short video, answer the question based on the video



00:00.755 --> 00:02.655
(Chandler:) Go to your room!
00:06.961 --> 00:08.622
(Janice:) I gotta go, I gotta go.

00:08.829 --> 00:10.057
(Janice:) Not without a kiss.
00:10.264 --> 00:12.391
(Chandler:) Maybe I won't kiss you so you'll stay.

00:12.600 --> 00:14.761
(Joey:) Kiss her. Kiss her!
00:16.771 --> 00:19.137
(Janice:) I'll see you later, sweetie. Bye, Joey.

…

00:39.327 --> 00:40.760
(Chandler:) She makes me happy.
00:41.596 --> 00:44.087
(Joey:) Okay. All right.

…

What is Janice holding on to after Chandler sends Joey to his room?

A Chandler's tie
B Chandler's hands
C Her Breakfast
D Her coat
E Chandler's coffee cup.

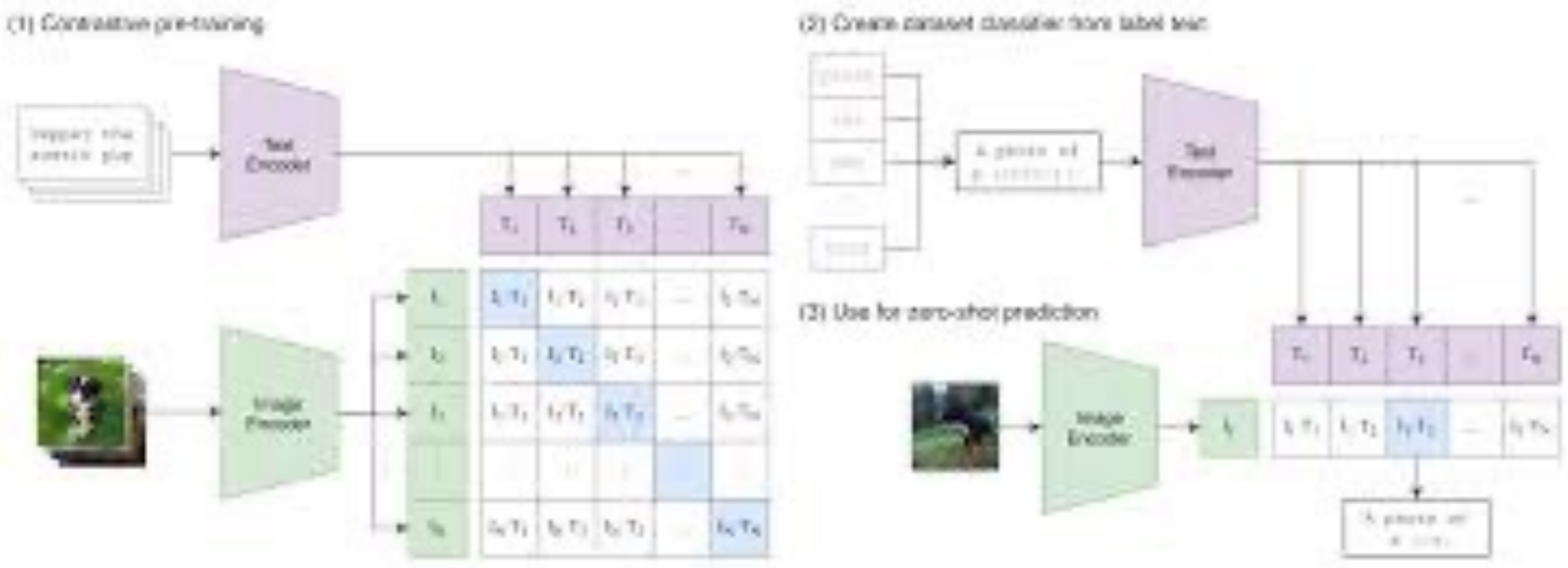
Why does Joey want Chandler to kiss Janice when they are in the kitchen?

A Because Joey is glad that Chandler is happy
B Because Joey likes to watch people kiss
C Because then she will leave
D Because Joey thinks Janice is hot
E Because then Chandler will move away from the toast.

What is on the couch behind Joey when he is at the counter?

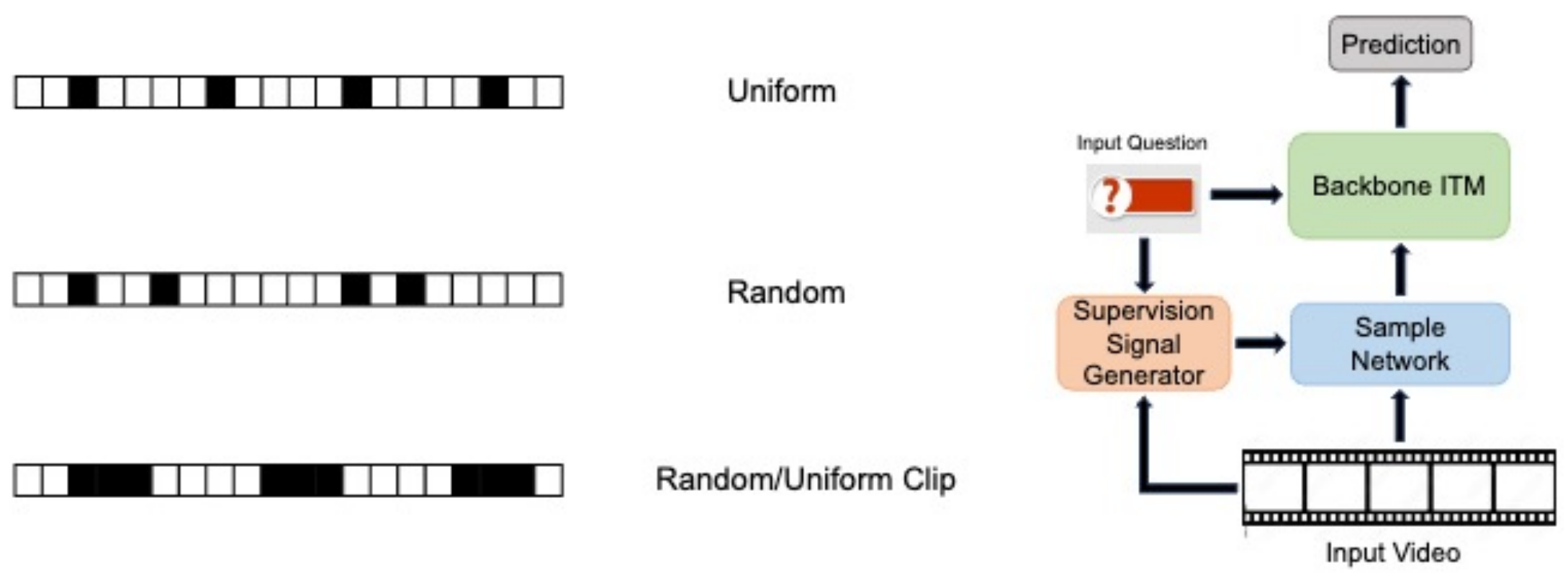
A A chick
B A soccer ball
C A duck
D A pillow
E Janice's coat

- Image—Text Models (ITMs): a subclass of visual language models (VLMs) that accept image sequences and text as input and generate text outputs, such as CLIP. To process video input, a series of frames must be first sampled from that video.


(1) Contrastive pre-training
Text Encoder
Image Encoder
(2) Create dataset classifier from label text
(3) Use for zero-shot prediction

## Related Works

- Current Sampling Strategy
  - Learning-free sampling is cost-effective but hard to reach optimal
  - Learning-based sampling can adapt to different question input, but requires additional computational cost (huge) and difficult to converge



(a) Heuristic Sampling

(b) Learning Based Sampling

Figure 2: Existing sample strategies for video–question answering tasks. In heuristic sampling, the black boxes indicate selected frames.

## Research Question

- Can we move the sampling stage offline (decouple it from the main network)?
- Can we find a simple yet effective formulation for the offline sampling?
- Is question-aware sampling always required (can we design a question-agnostic one)?

## Method

- Most Implied Frames (MIF)
  - A captioner and a scorer to calculate scores for each frame
  - Choose the frames of the highest scores as sampled ones

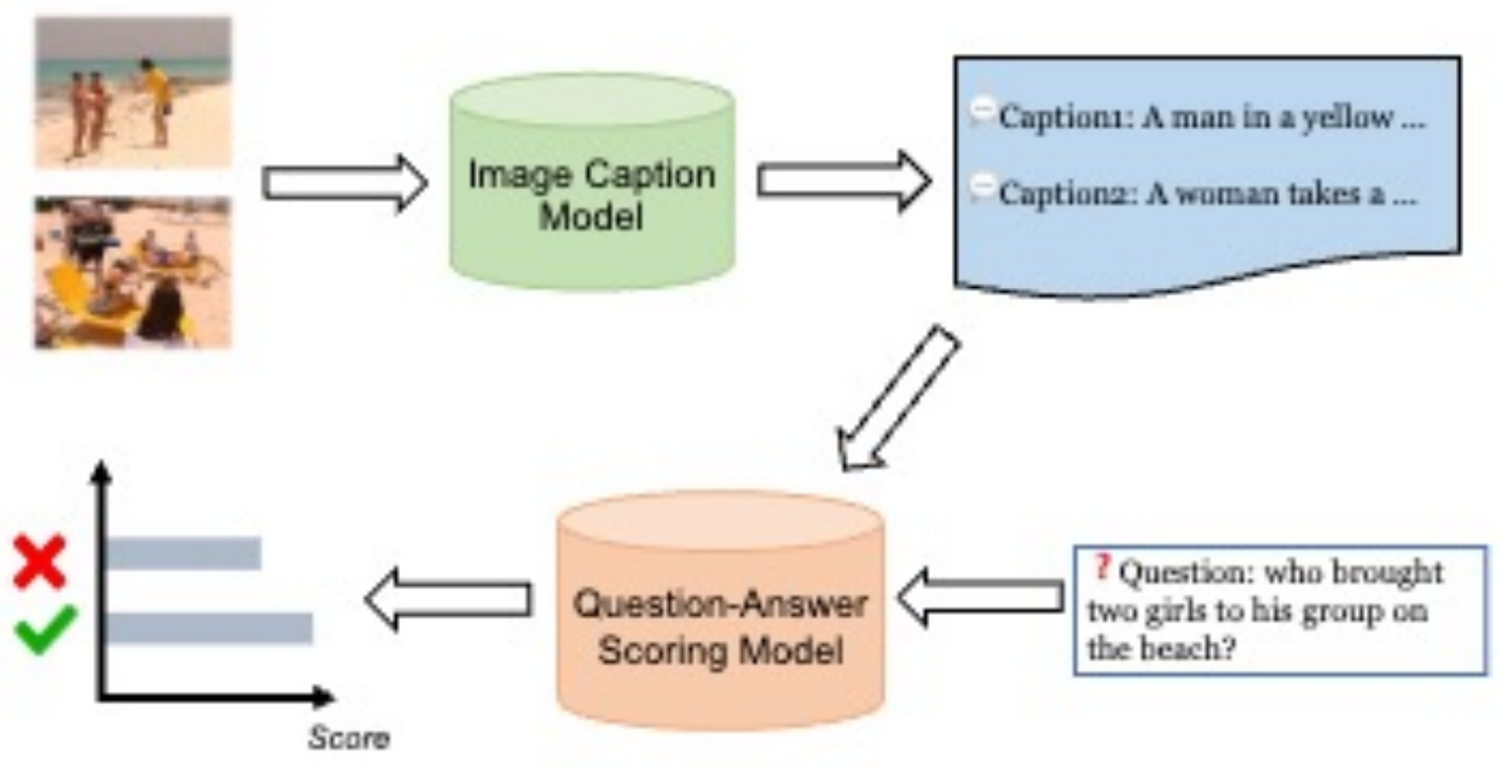


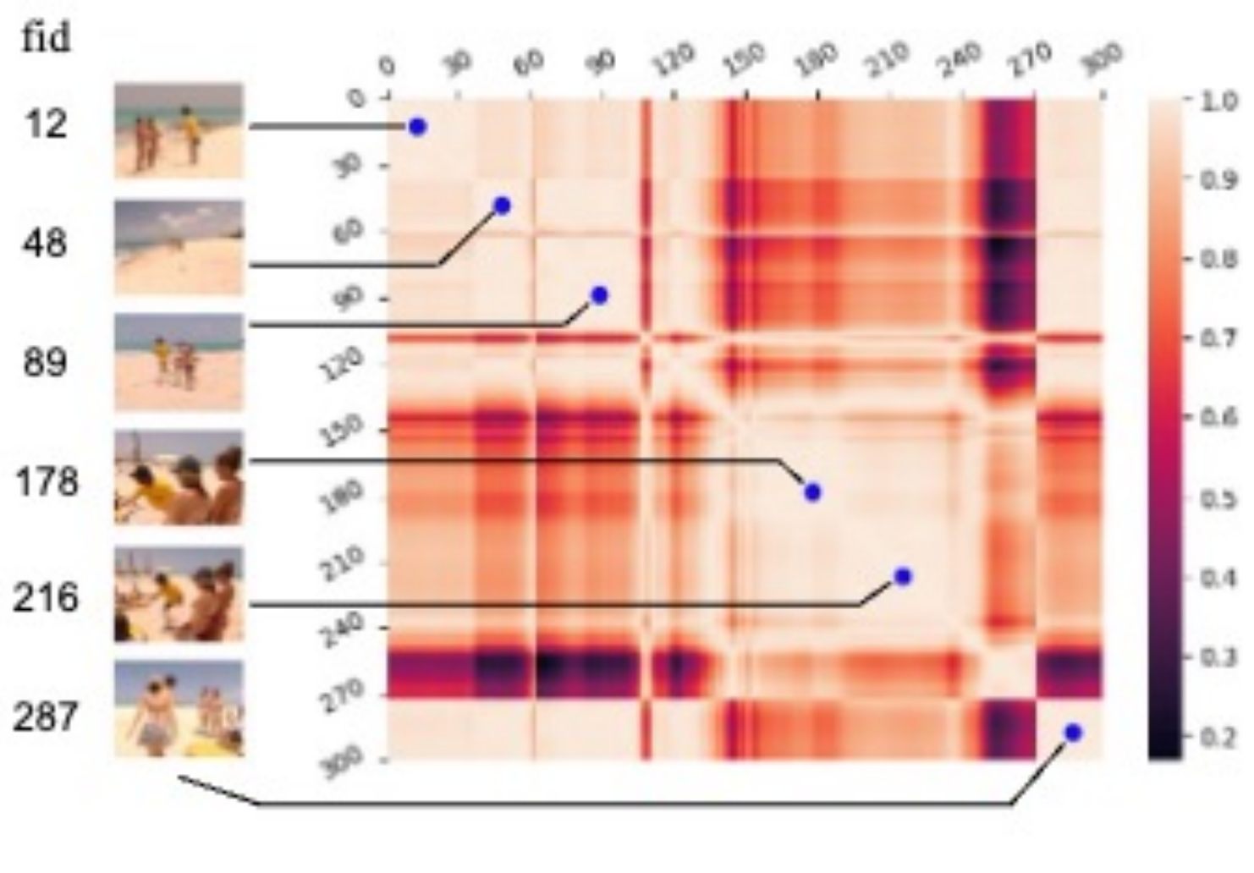
Image Caption Model
Caption1: A man in a yellow ...
Caption2: A woman takes a ...
Question-Answer Scoring Model
Question: who brought two girls to his group on the beach?
Score

| FID | Caption | Q1 | Q2 |
|---|---|---|---|
| 1 | a puppy playing with toys. | | |
| 2 | a white puppy playing with a toy. | | |
| 3 | a white puppy with black eyes and a blue ball. | ✓ | |
| 4 | a puppy that is laying down on the floor. | | |
| 5 | a puppy playing with a blue ball. | | |
| 6 | a puppy that was found in a house. | | ✓ |
| 7 | a puppy that is laying down on the floor. | | |
| 8 | a puppy that is sitting on the floor. | | ✓ |
| 9 | a puppy is sitting on the floor. | ✓ | ✓ |
| 10 | a white puppy sitting on a table. | | ✓ |
| 11 | a white puppy laying on the floor. | ✓ | ✓ |
| 12 | a puppy playing with a blue ball. | ✓ | |
| 13 | a white dog standing on top of a floor. | ✓ | ✓ |
| 14 | a white dog walking on the floor. | ✓ | |
| 15 | a small white dog playing with a ball. | | |
| 16 | a dog chewing on a toy in a cage. | | |

- Ablative study of MIF: Is question-aware sampling a must?
  - We change the scale (capability) of captioner and grader respectively
  - There is no obvious correlation between captioner-grader capability and accuracy

| $\mathcal{M}_c$ | $\mathcal{M}_g$ | MSVD | MSRVTT |
|---|---|---|---|
| ***Separate Model*** | | | |
| GIT-S | BERT-S | 46.5 | 42.3 |
| GIT-B | BERT-B | 46.7 | 42.4 |
| GIT-L | BERT-L | 46.9 | 42.1 |
| ***Unified Model*** | | | |
| BLIP2-T5-XL | | 46.6 | 42.0 |
| BLIP2-T5-XXL | | 46.2 | 42.2 |

- Most Dominant Frames (MDF)
  - Based on previous analysis, we can move one step forward by even discarding the question-aware component
  - Sampling scores are calculated on visual feature similarity



**Algorithm 1:** Most Dominant Frames (MDF)

```
Input: Video frames $V = \{v_1, v_2, ..., v_T\}$, vision
       model $\mathcal{M}$, width-adjusting rate $\lambda$
Output: Visual prefix $F = \{f_1, f_2, ..., f_N\}$
Encode frames using the vision model
  $E = \mathcal{M}(V) = \{e_1, e_2, ..., e_T\}$
Compute $dom$ score for all frames and set $W$,
  according to Eq. 4 and Eq. 5.
Init $F = \{f_{\arg\max_t dom(t)}\}$, index set
  $I = \{0, 1, ..., i - W, i + W, ..., T\}$
while $|F| < N$ and $I \neq \oslash$ do
    $t' \leftarrow \arg\max_t dom(t)$
    $F \leftarrow F \cup \{f_{t'}\}$
    $I \leftarrow I \setminus \{t''\}_{t''-t'<W}$
if $|F| < N$ then
    $\tau \leftarrow \mathtt{argtop}_N(\{dom(t)\}_{t \in T})$
    return $F \cup \{f'_t\}_{t' \in \tau}$
else
    return $F$
```

## Results

| Model | MSVD | MSRVTT | TGIF |
|---|---|---|---|
| ***GIT Backbone*** | | | |
| Base (Wang et al., 2022) | 52.2 | 41.1 | 67.5 |
| IGV (Li et al., 2022c) | 53.2 | 41.5 | 68.1 |
| VCSR (Wei et al., 2023) | 52.7 | 41.6 | 68.6 |
| MIF | 54.5 | **42.3** | 69.9 |
| MDF | **55.3** | 42.0 | **70.0** |
| ***AIO Backbone*** | | | |
| Base (Wang et al., 2023) | 46.1 | 42.7 | 64.0 |
| IGV (Li et al., 2022c) | 46.3 | 43.3 | 64.7 |
| VCSR (Wei et al., 2023) | 46.4 | 43.0 | 64.5 |
| MIF | 46.7 | **44.0** | 65.9 |
| MDF | **46.9** | 43.8 | **66.2** |

Table 3: Test set results on MSVD, MSRVTT and TGIF. Best scores are bolded.

- Both MIF and MDF achieve good performance
- MDF is competitive to MDF, showing that question-aware sampling is not necessary