HERO: Hierarchical EncodeR for Video+Language Omni-representation Pre-training

Linjie Li*, Yen-Chun Chen*, Yu Cheng, Zhe Gan, Licheng Yu, Jingjing Liu

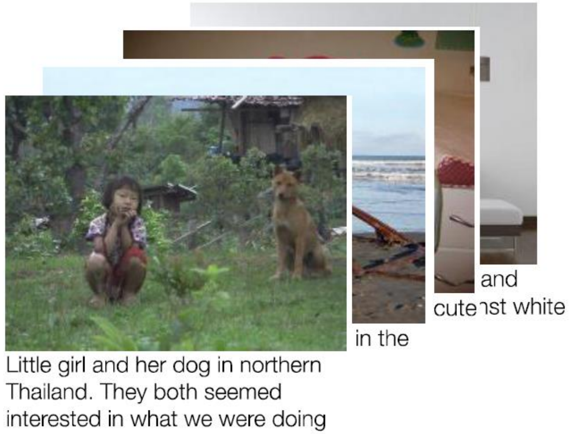
Microsoft Dynamics 365 AI Research

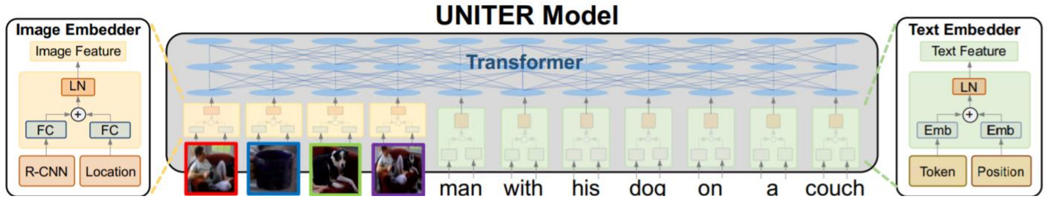
* Equal Contribution

Vision + Language Pre-training

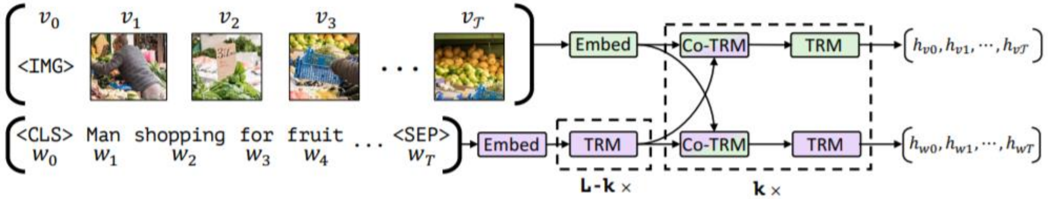
Vision: Image

Language: Textual Descriptions

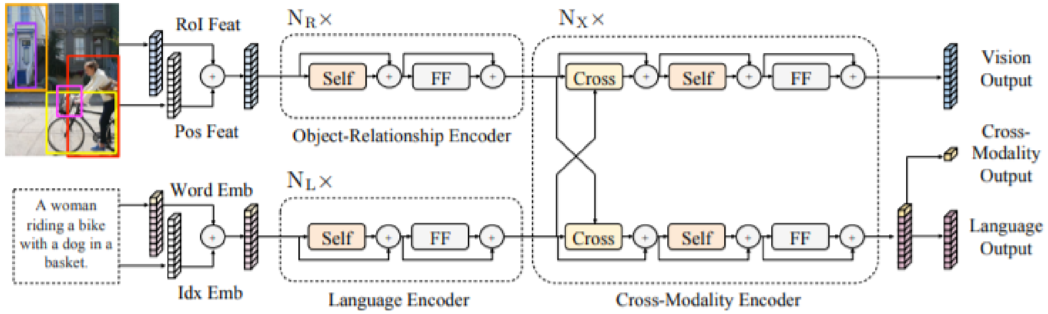




UNITER [Chen et al. 2019]



VilBERT [Lu et al. NeurIPs 2019]

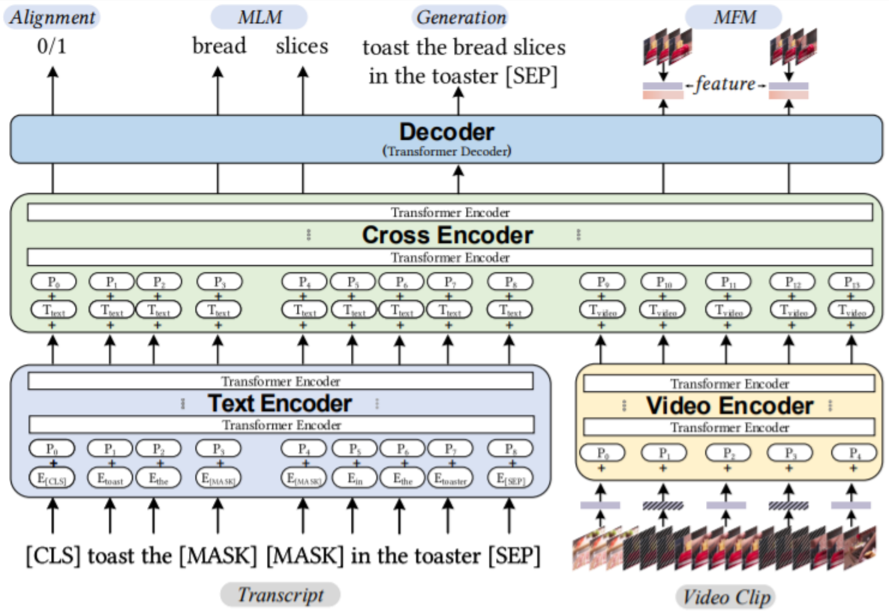


LXMERT [Tan and Bansal, EMNLP 2019]

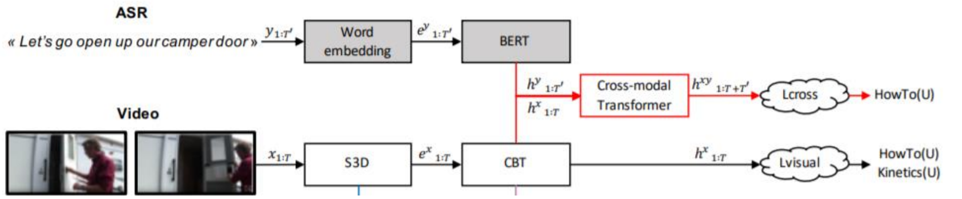
Video + Language Pre-training

Video: Sequence of image frames
Language: Subtitles/Narrations

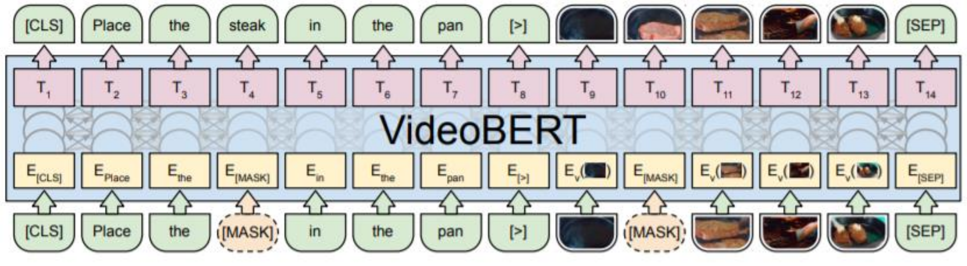
00:00:02 --> 00:00:04
That's why you won't go out with her again?
00:00:34 --> 00:00:36
- Thank God you're here. Listen to this.
- What?
00:00:66 --> 00:00:68
(Joey:) Joey doesn't share food!



UniViLM [Luo et al. 2020]



CBT [Sun et al. 2019]



VideoBERT [Sun et al. 2019]

Video + Language Pre-training

- Limitations of existing methods
  - Video + Text inputs are directly concatenated, losing the temporal alignment
  - Pre-training tasks directly borrowed from Image + Text pre-training
  - Pre-training datasets limited to narrated instructional videos from Howto100M [Miech et al. ICCV 2019]
- **HERO** (**H**ierarchical **E**ncode**R** for **O**mni-representation learning)
  - New model architecture:
    - Local temporal alignments between frames and subtitles are captured by a *Cross-modal Transformer*
    - Global temporal context are modeled by a *Temporal Transformer*
  - New Pre-training tasks: *Video-Subtitle Matching* and *Frame Order Modeling*
  - Diverse Pre-training Datasets: Howto100M and TV dataset [Lei at al. ACL 2018]
    - We further collect two downstream datasets based on Howto100M

HERO: Hierarchical EncodeR for Omni-representation learning



00:00:02 --> 00:00:04
That's why you won't go out with her again?
00:00:34 --> 00:00:36
- Thank God you're here. Listen to this.
- What?

00:00:66 --> 00:00:68
(Joey:) Joey doesn't share food!

HERO: Hierarchical EncodeR for Omni-representation learning

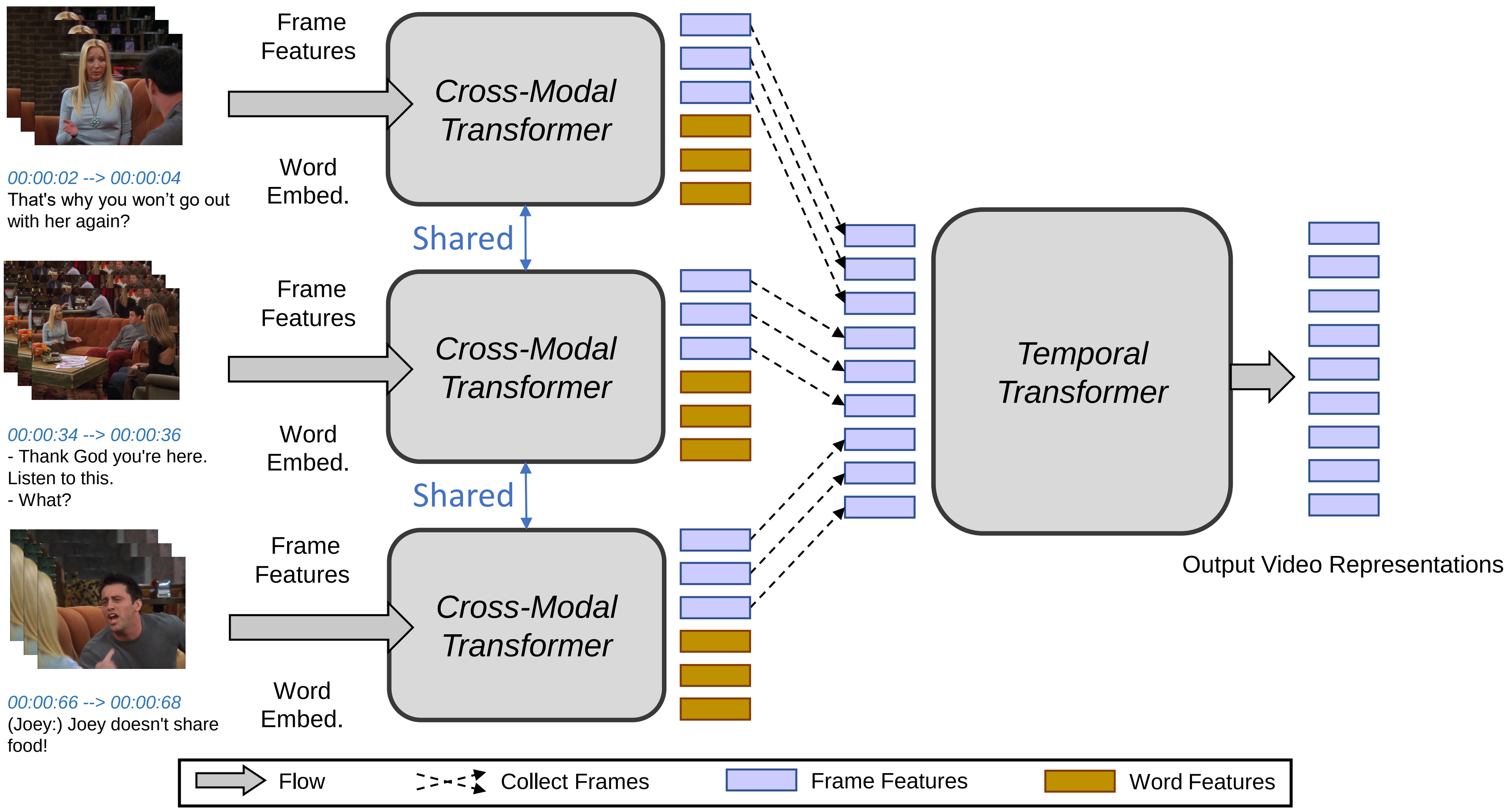
00:00:02 --> 00:00:04
That's why you won’t go out with her again?

00:00:34 --> 00:00:36
- Thank God you're here. Listen to this.
- What?

00:00:66 --> 00:00:68
(Joey:) Joey doesn't share food!

- Temporally align subtitle sentences with frames
- Frame features: 2D ResNet Features [He et al. CVPR 2016] and 3D SlowFast Features [Feichtenhofer et al. ICCV 2019]
- Subtitle sentences are tokenized and each word are embedded following RoBERTa [Liu et al. 2019]

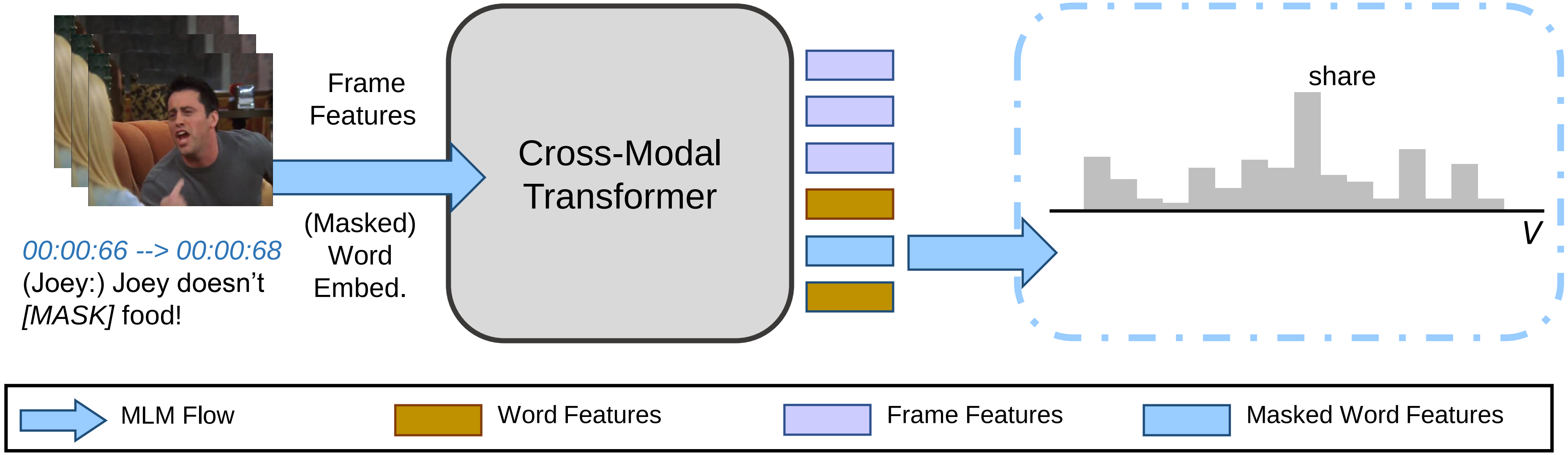
HERO: Hierarchical EncodeR for Omni-representation learning



Pre-training HERO

- Pre-training Tasks
  - Masked Language Modeling (MLM)
  - Masked Frame Modeling (MFM)
  - *Video-Subtitle Matching (VSM)*
  - *Frame Order Modeling (FOM)*

Masked Language Modeling (MLM)



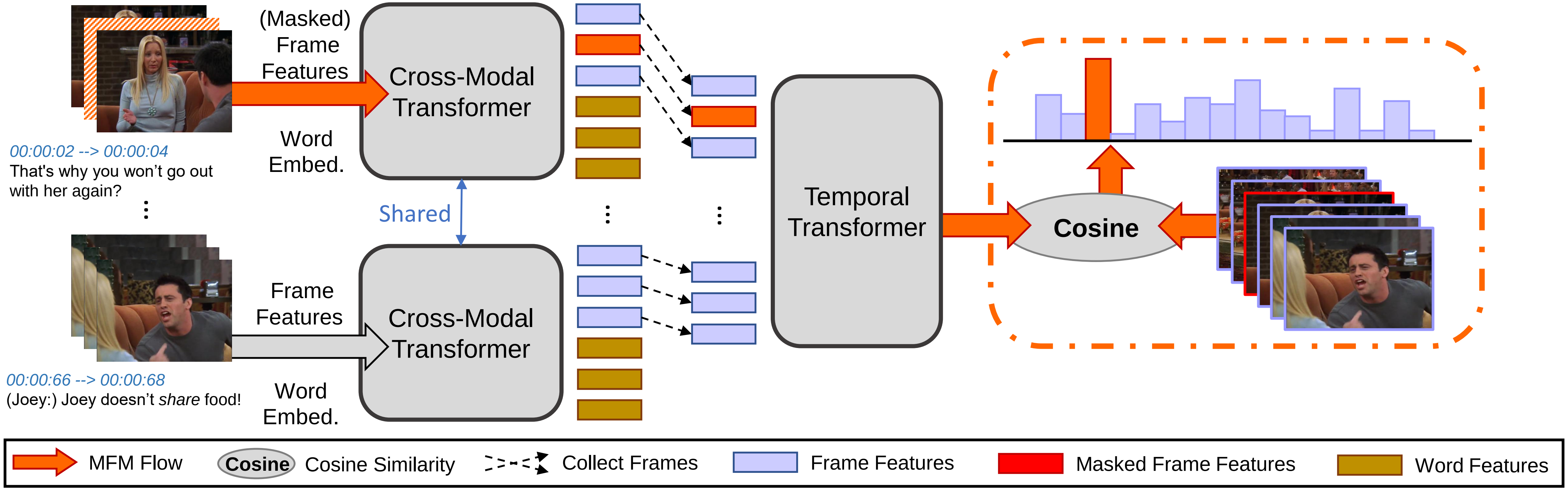
Word Tokens of Subtitle s_i: $\mathbf{w}_{s_i} = \{w_{s_i}^j\}_{j=1}^L$

Visual Frames aligned with s_i: $\mathbf{v}_{s_i} = \{v_{s_i}^j\}_{j=1}^K$

Masking Indices: $\mathbf{m} \in \mathbb{N}^M$

Loss Function of MLM: $\mathcal{L}_{\text{MLM}}(\theta) = -\mathbb{E}_D \log P_\theta(\mathbf{w}_{s_i}^{\mathbf{m}} | \mathbf{w}_{s_i}^{\backslash \mathbf{m}}, \mathbf{v}_{s_i})$

Masked Frame Modeling (MFM)



All subtitle sentences: $\mathbf{s} = \{s_i\}_{i=1}^{N_s}$

Visual Frames: $\mathbf{v} = \{v_i\}_{i=1}^{N_v}$

Masking Indices: $\mathbf{m} \in \mathbb{N}^M$

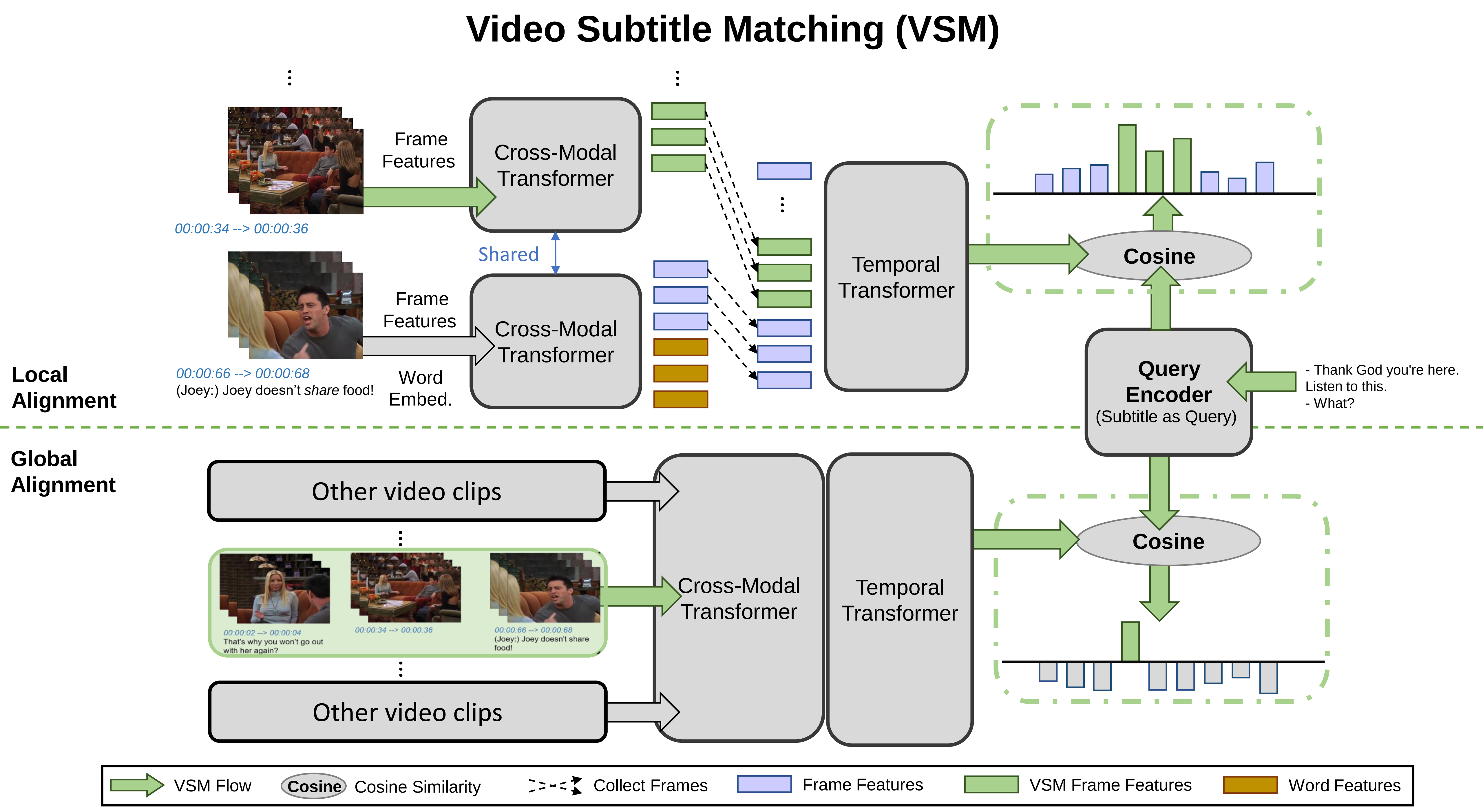
Loss Function of MFM: $\mathcal{L}_{\text{MFM}}(\theta) = \mathbb{E}_D f_\theta(\mathbf{v_m}|\mathbf{v}_{\backslash\mathbf{m}}, \mathbf{s})$

(1) Masked Frame Feature Regression (MFFR)

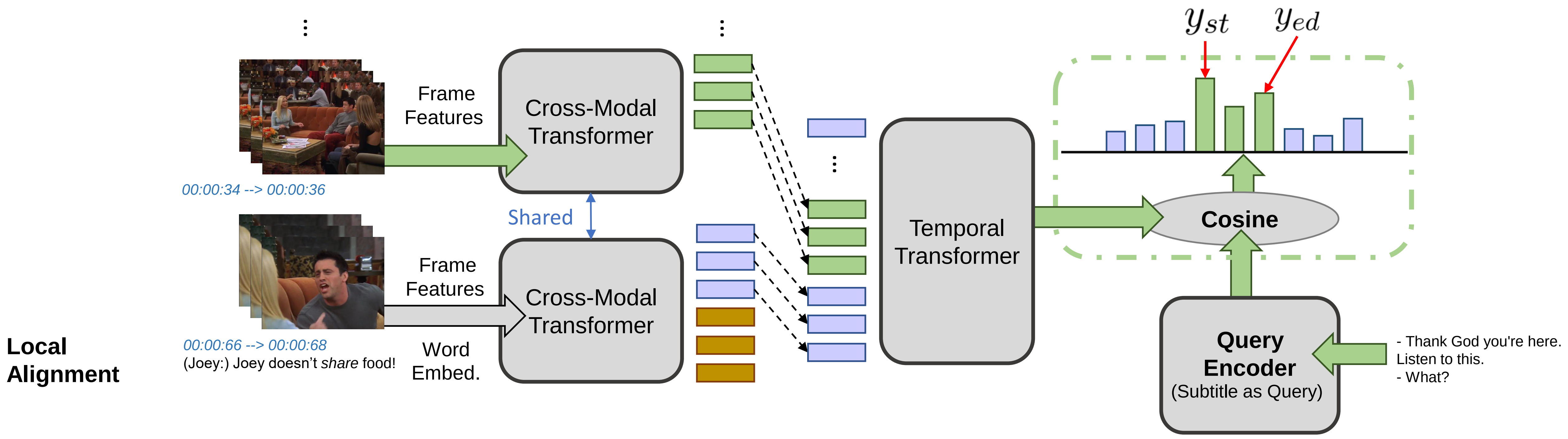
$$f_\theta(\mathbf{v_m}|\mathbf{v}_{\backslash\mathbf{m}}, \mathbf{s}) = \sum_{i=1}^{M} \|h_\theta(\mathbf{v}_\mathbf{m}^{(i)}) - r(\mathbf{v}_\mathbf{m}^{(i)})\|_2^2$$

(2) Masked Frame with Noise Contrastive Estimation (M-NCE)

$$f_\theta(\mathbf{v_m}|\mathbf{v}_{\backslash\mathbf{m}}, \mathbf{s}) = \sum_{i=1}^{M} \log \text{NCE}(g_\theta(\mathbf{v}_\mathbf{m}^{(i)})|g_\theta(\mathbf{v_{neg}}))$$

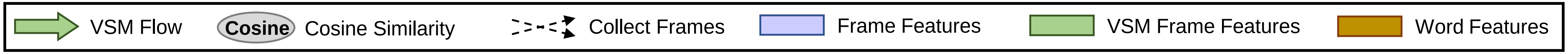

Video Subtitle Matching (VSM)
Frame Features
Cross-Modal Transformer
00:00:34 --> 00:00:36
Shared
Frame Features
Cross-Modal Transformer
Word Embed.
00:00:66 --> 00:00:68
(Joey:) Joey doesn't *share* food!
Local Alignment
Temporal Transformer
Cosine
Query Encoder
(Subtitle as Query)
- Thank God you're here.
Listen to this.
- What?
Global Alignment
Other video clips
00:00:02 --> 00:00:04
That's why you won't go out with her again?
00:00:34 --> 00:00:36
00:00:66 --> 00:00:68
(Joey:) Joey doesn't share food!
Other video clips
Cross-Modal Transformer
Temporal Transformer
Cosine
VSM Flow
Cosine
Cosine Similarity
Collect Frames
Frame Features
VSM Frame Features
Word Features

Video Subtitle Matching (VSM)



Start and end index of overlapping frames: $y_{st}, \ y_{ed}$

Loss function of local alignments: $\mathcal{L}_{local} = -\mathbb{E}_D \log(\mathbf{p}_{st}[y_{st}]) + \log(\mathbf{p}_{ed}[y_{ed}])$



Video Subtitle Matching (VSM)

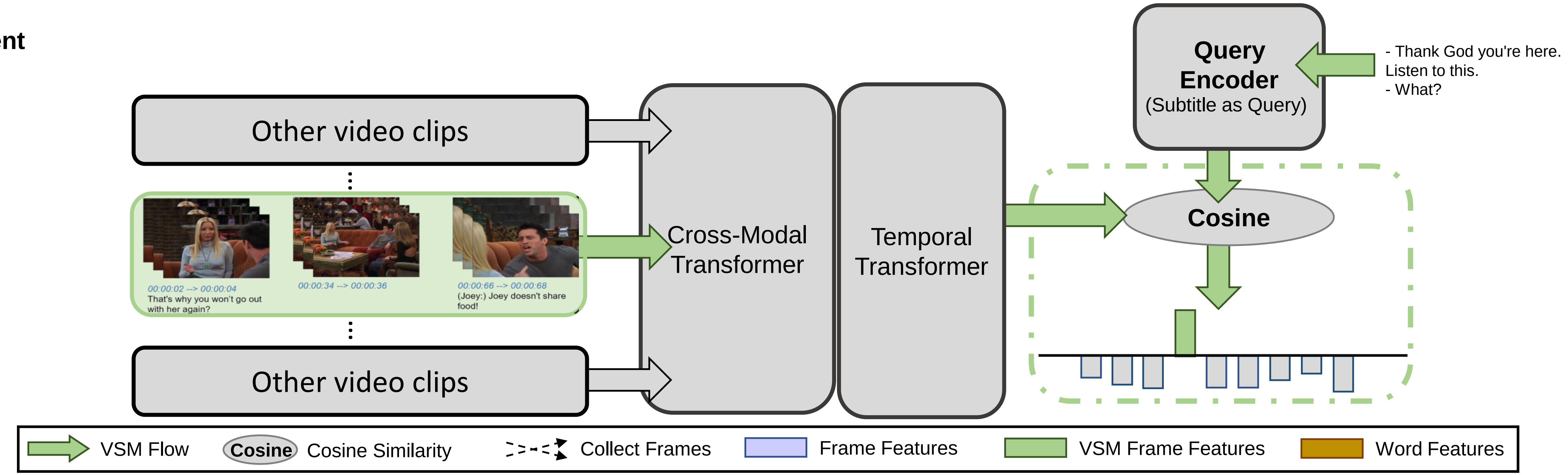
Positive and negative video-subtitle pairs: $(s_q, \mathbf{v})$, $(s_q, \hat{\mathbf{v}})$, $(\hat{s}_q, \mathbf{v})$
Similarity measure: S

Hinge loss: $\mathcal{L}_h(S_{pos}, S_{neg}) = \max(0, \delta + S_{neg} - S_{pos})$

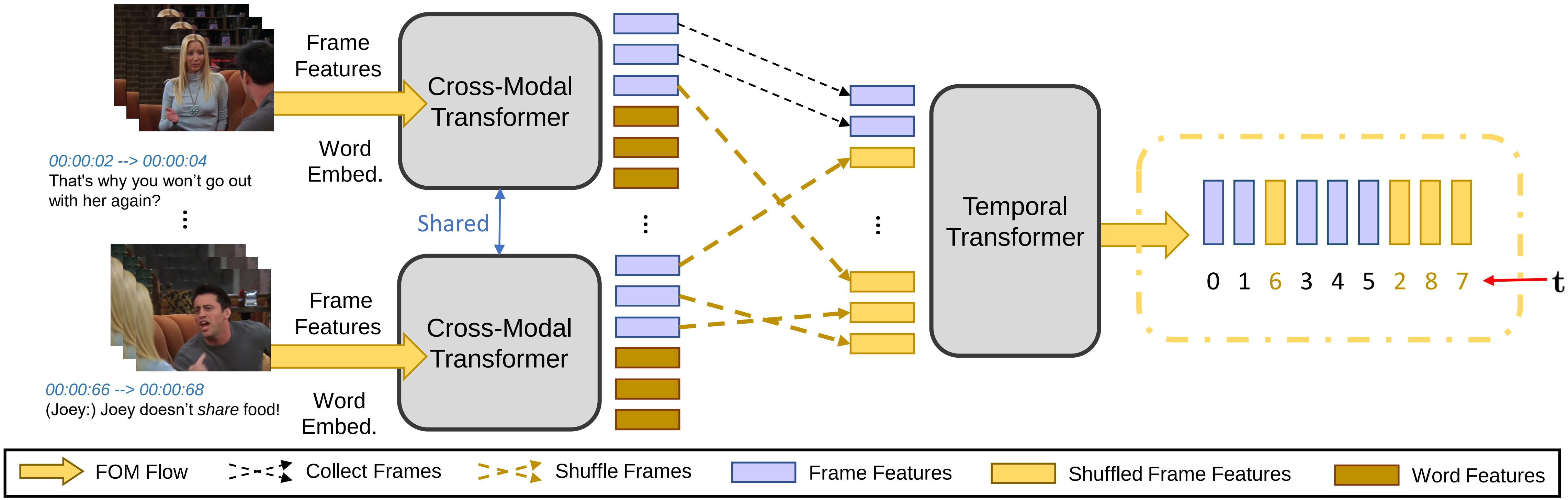
Loss function of global alignments:

$$\mathcal{L}_{global} = -\mathbb{E}_D[\mathcal{L}_h(S_{global}(s_q, \mathbf{v}), S_{global}(\hat{s}_q, \mathbf{v})) + \mathcal{L}_h(S_{global}(s_q, \mathbf{v}), S_{global}(s_q, \hat{\mathbf{v}}))]$$

Global Alignment



Frame Order Modeling (FOM)



Reorder Indices: $\mathbf{r} = \{r_i\}_{i=1}^{R} \in \mathbb{N}^R$

Original timestamp: $\mathbf{t} = \{t_i\}_{i=1}^{R}$

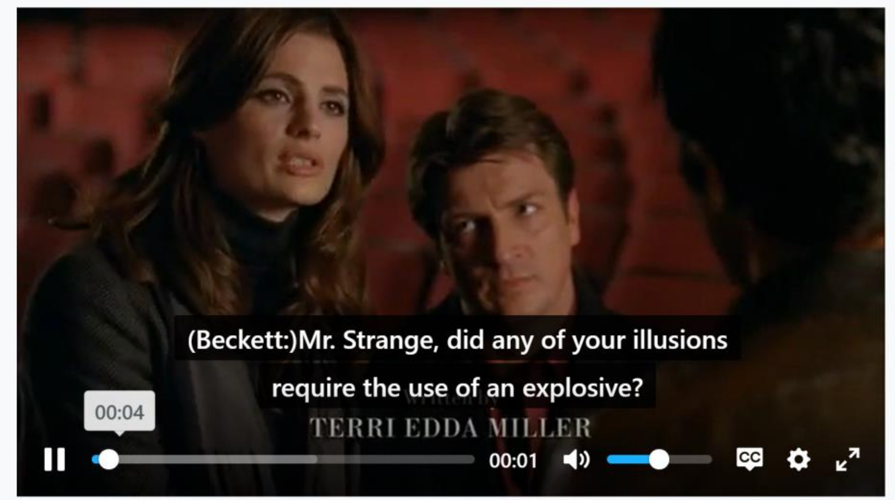
Loss Function of FOM: $\mathcal{L}_{\text{FOM}} = -\mathbb{E}_D \sum_{i=1}^{R} \log \mathbf{P}[r_i, t_i]$

Pre-training HERO

- Pre-training Tasks
  - Masked Language Modeling (MLM)
  - Masked Frame Modeling (MFM)
  - *Video-Subtitle Matching (VSM)*
  - *Frame Order Modeling (FOM)*
- Pre-training Datasets
  - TV Dataset
  - Howto100M Dataset

Our Pre-training Data for Video + Language

TV Dataset



- 22K video clips from 6 popular TV shows
- Each video clip is 60-90 seconds long
- Dialogue ("character name: subtitle") is provided

Howto100M Dataset

- 1.22M instructional videos from YouTube
- Exclude videos in non-English languages and cut the rest into 60-second clips
- 660K video clips with English subtitles

Video + Language Downstream Tasks

Video: Sequence of image frames
Language: Subtitles/Narrations

00:00:02 --> 00:00:04
That's why you won't go out with her again?
00:00:34 --> 00:00:36
- Thank God you're here. Listen to this.
- What?
00:00:66 --> 00:00:68
(Joey:) Joey doesn't share food!

Video Captioning
Caption: Joey's dating policy: never shares food!

Text-based Video Moment Retrieval
Query: Joey's dating policy: never shares food!

Video Question Answering
Question: Why did Joey complain about his date?
Answer: She took Joey's fries

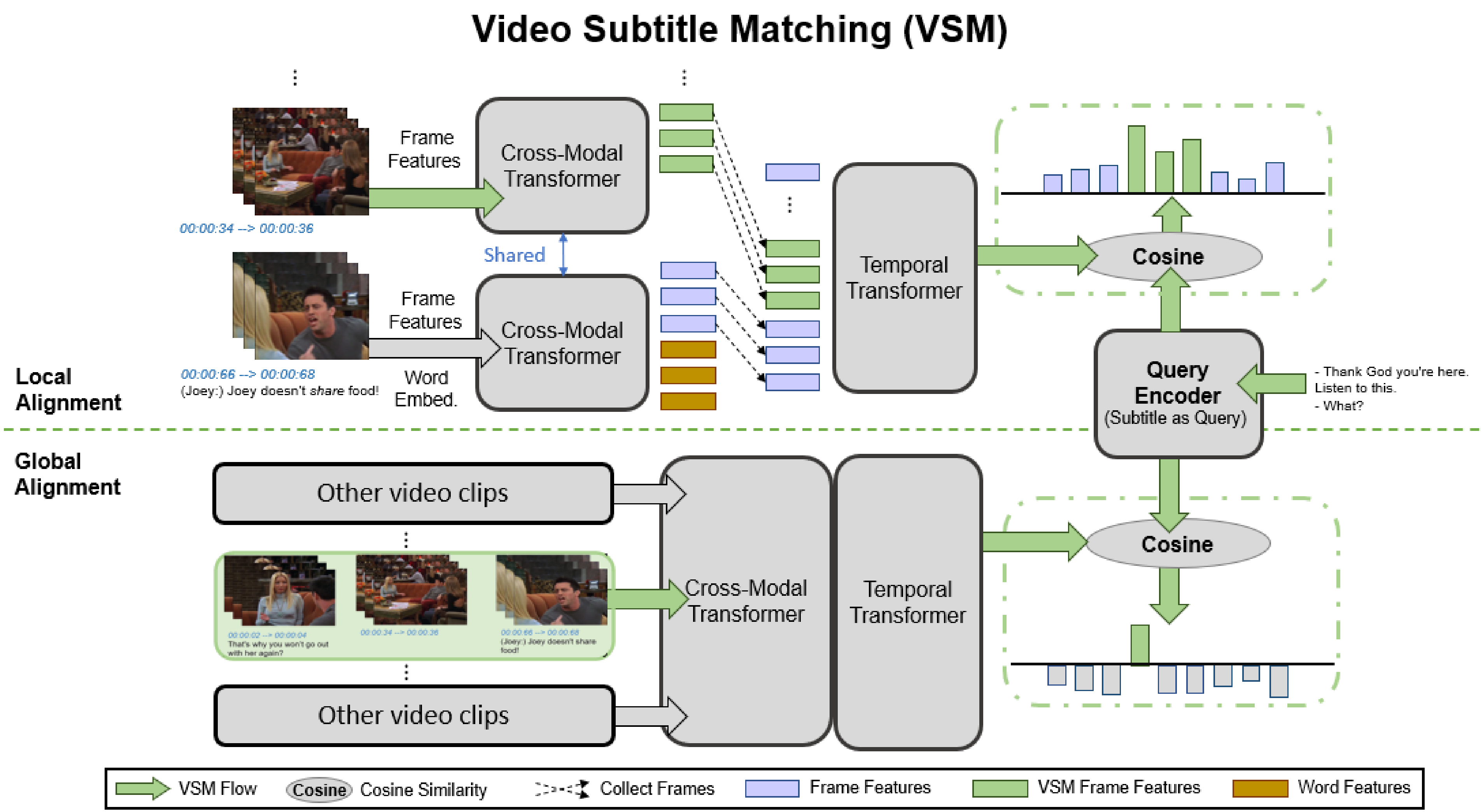
Downstream Task 1: Video Moment Retrieval

TVR [Lei et al. 2020]

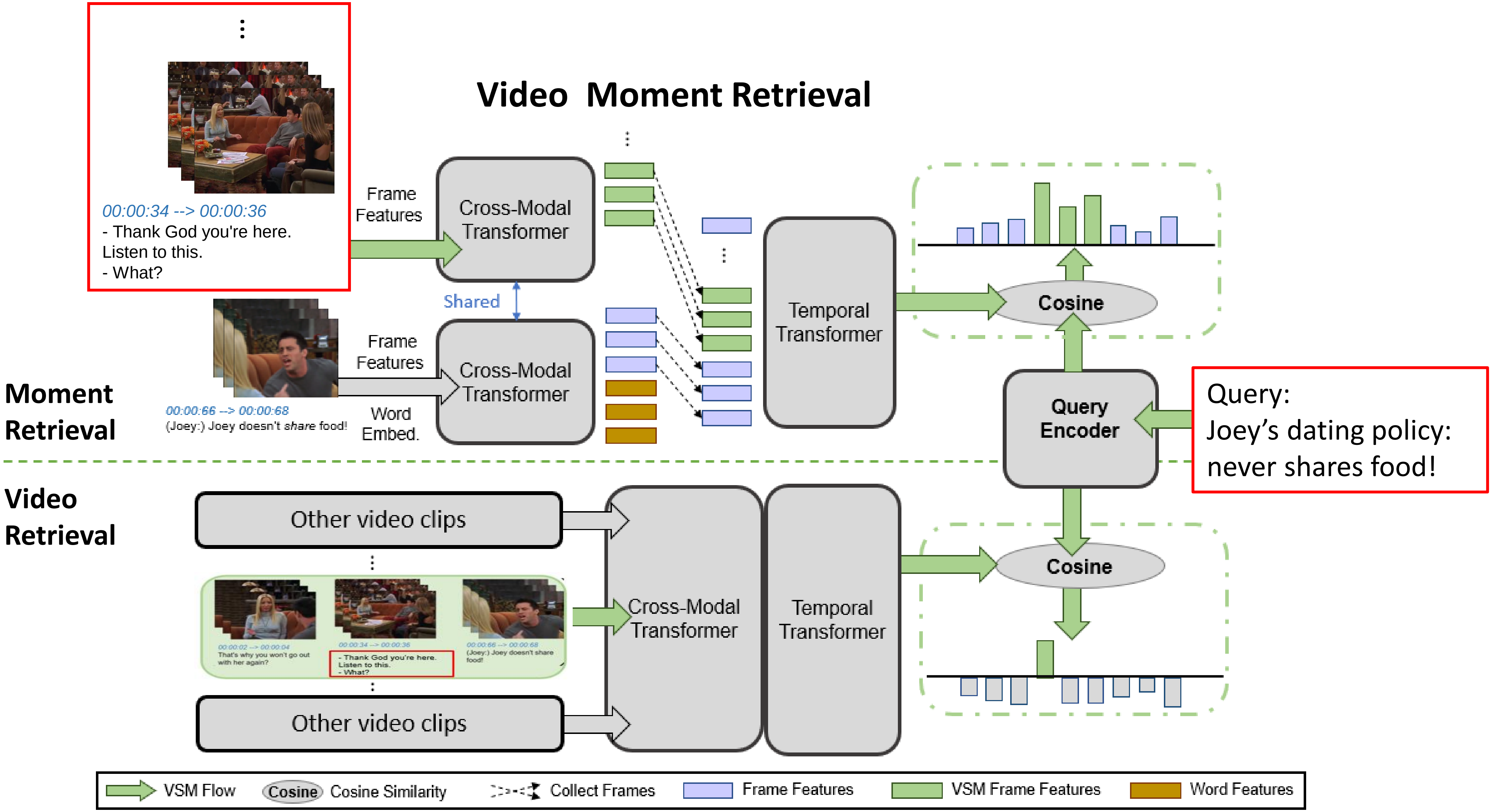
Video Moment Retrieval = Video Retrieval + Moment Retrieval

- **Subtask I: Video Retrieval**
  - From video corpus, retrieve the most relevant video clip described by the query
- **Subtask II: Moment Retrieval**
  - Given the query, localize the correct moment from the most relevant video clip
- Evaluation:
  - Average recall at K (R@K) over all queries
  - Temporal Intersection over Union (tIOU) is used to measure the performance of moment retrieval

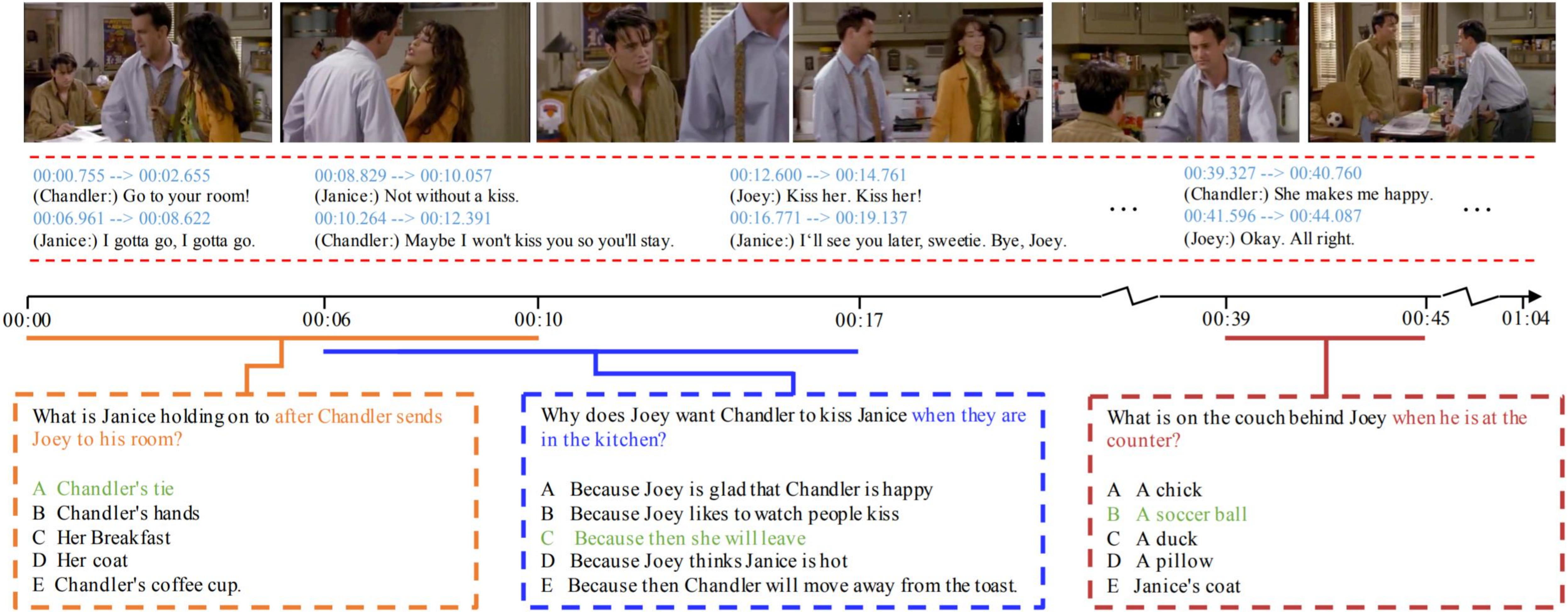
Downstream Task 1: Video Moment Retrieval



Downstream Task 1: Video Moment Retrieval

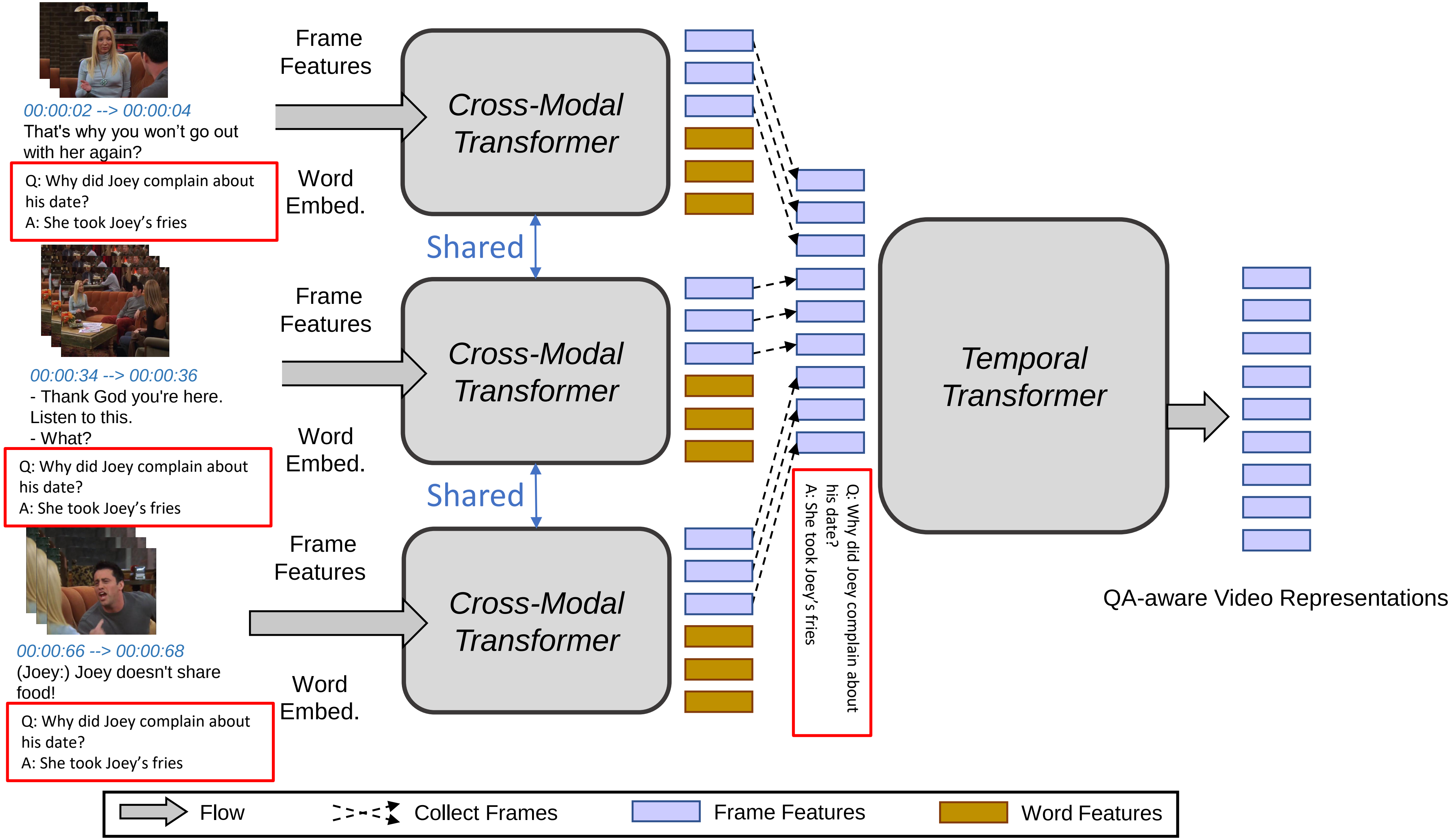


Downstream Task 2: Video Question Answering

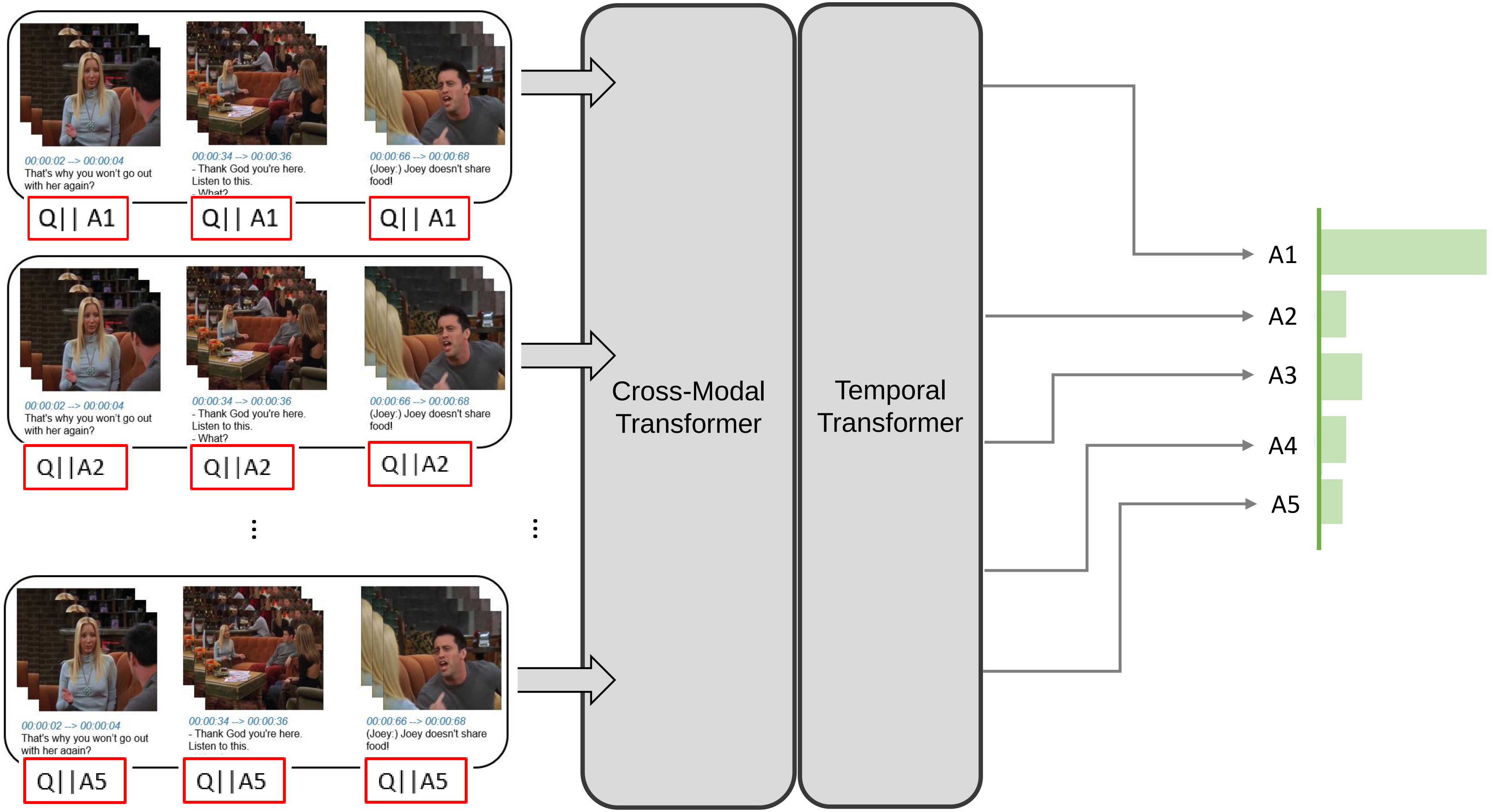


TVQA [Lei et al. EMNLP 2018]

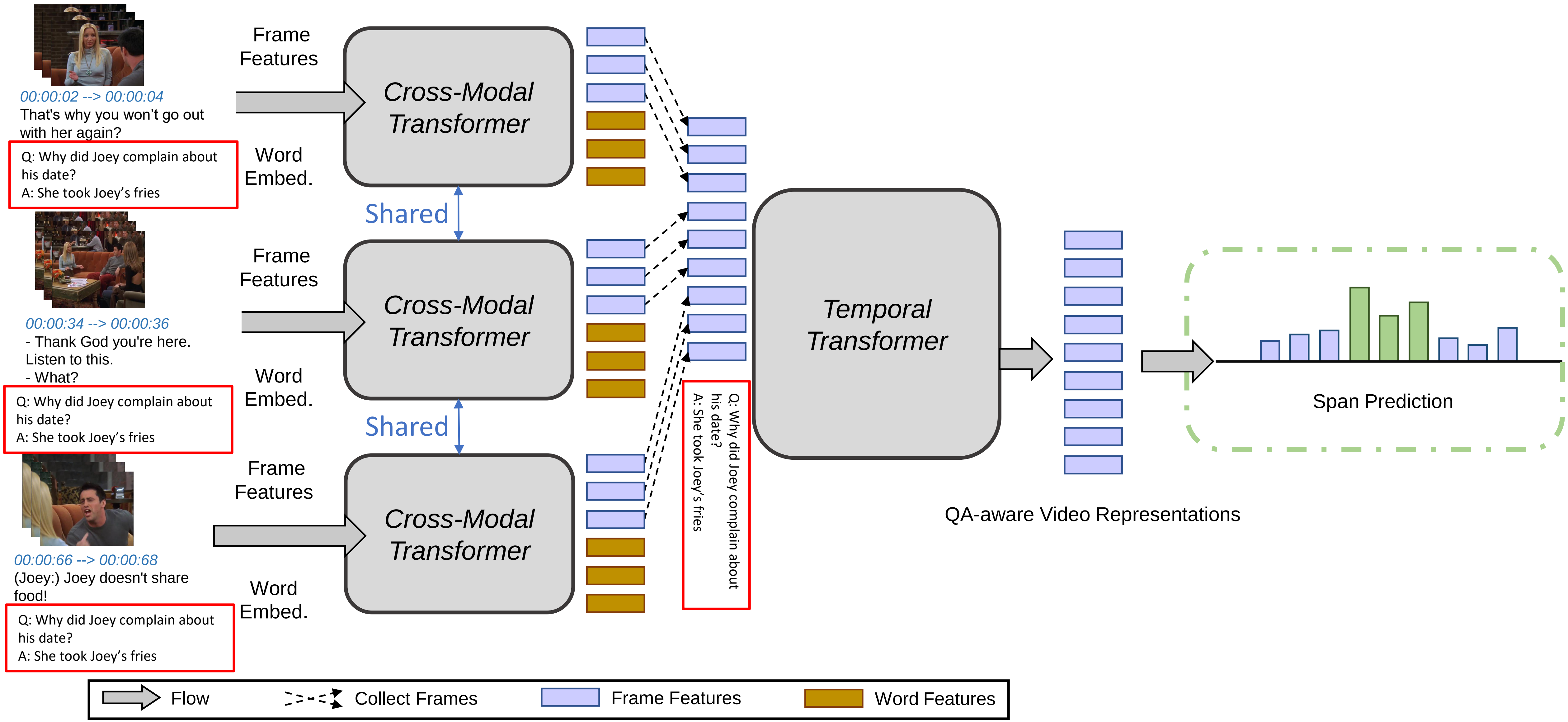
Downstream Task 2: Video Question Answering



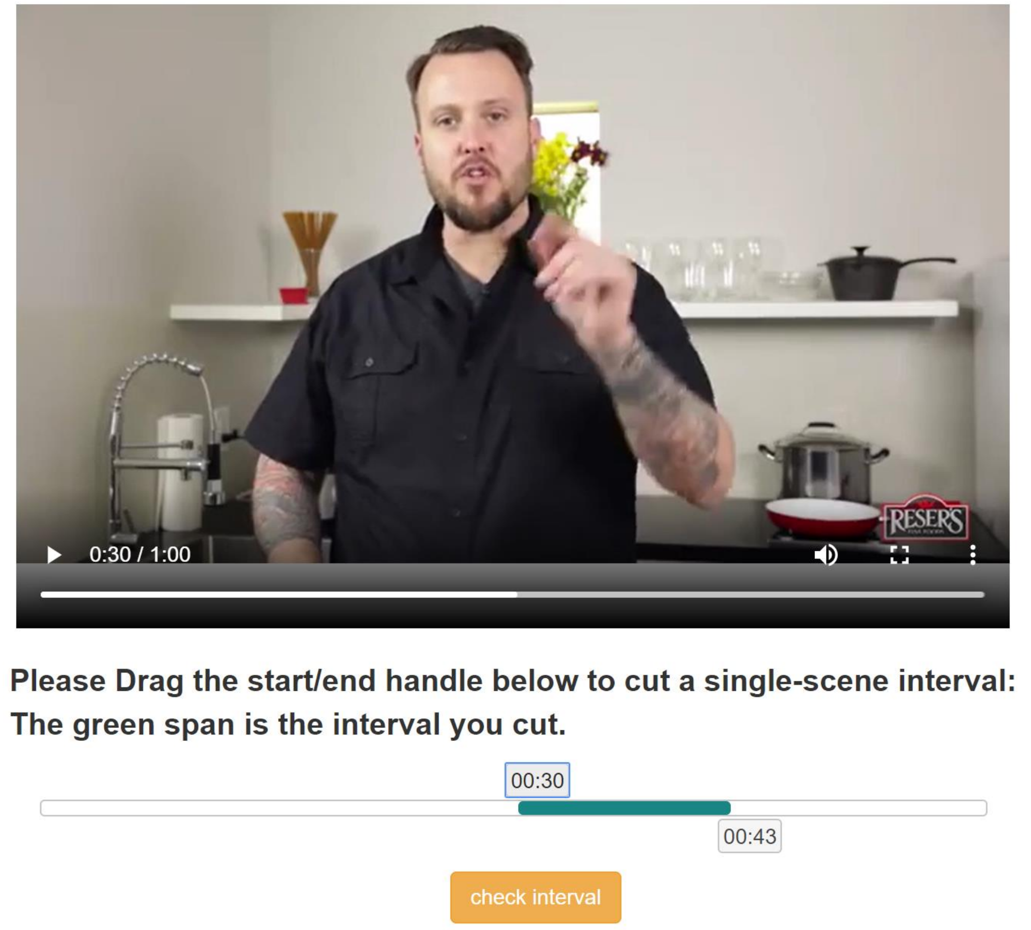
Downstream Task 2: Video Question Answering



Downstream Task 2: Video Question Answering



Downstream Data Collection



Caption:

8 to 20 words

Text-based Video Moment Retrieval

- **Howto100M-R**
  - 67K text queries are collected for 30K 60-second video clips from Howto100M
- Instructions:
  - First, select a video segment
  - Then, write a caption describing the selected segment

Downstream Data Collection

Your Question:

5 to 25 words

Correct answer

Wrong answer 1

Wrong answer 2

Wrong answer 3

Video Question Answering

- **Howto100M-QA**
  - QA collected for video segments annotated from video moment retrieval
  - On average, 2 questions per video segment
  - One correct answer and three wrong answers are written by the same annotator
  - Using adversarial matching [Zeller et al. CVPR 2019] to construct harder negative answers

Ablation Study

Pre-training Data		Pre-training Tasks	TVR			TVQA	Howto100M-R			Howto100M-QA
			R@1	R@10	R@100	Acc.	R@1	R@10	R@100	Acc.
TV	1	MLM	2.92	10.66	17.52	71.25	2.06	9.08	14.45	76.42
	2	MLM + MNCE	3.13	10.92	17.52	71.99	2.15	9.27	14.98	76.95
	3	MLM + MNCE + FOM	3.09	10.27	17.43	72.54	2.36	9.85	15.97	77.12
	4	MLM + MNCE + FOM + VSM	**4.44**	**14.69**	**22.82**	72.75	2.78	10.41	**18.77**	77.54
	5	MLM + MNCE + FOM + VSM + MFFR	**4.44**	14.29	22.37	72.75	2.73	10.12	18.05	77.54
TV & Howto100M	6	MLM + MNCE + FOM + VSM	4.34	13.97	21.78	**74.24**	**2.98**	**11.16**	17.55	**77.75**

1. Best combination: MLM + MNCE + FOM + VSM
2. QA tasks benefit from FOM
3. Retrieval tasks benefit from VSM
4. Adding more data generally give better results

Ablation Study

- Comparison with two baseline models with/without pre-training
- F-TRM
  - A flat BERT-like encoder
  - Input is a single sequence by concatenating video frames and subtitle sentences
- H-TRM
  - Replacing Cross-modal Transformer with RoBERTa to encode subtitle only
  - Max-pooled subtitle sentence embeddings is added to temporally aligned frame embeddings

Pre-training	Model	TVR			TVQA
		R@1	R@10	R@100	Acc.
No[8]	F-TRM	1.99	7.76	13.26	31.80
	H-TRM	2.97	10.65	18.68	70.09
	HERO	2.98	10.65	18.25	70.65
Yes	H-TRM	3.12	11.08	18.42	70.03
	HERO	**4.44**	**14.69**	**22.82**	**72.75**

1. Without pre-training, HERO and H-TRM outperforms F-TRM
   - Inherent temporal alignment between two modalities of videos is important
2. With pre-training, HERO outperforms H-TRM
   - Cross-modal interactions between visual frames and its local textual context is critical

Comparison with SOTA Models

Method	TVR			Howto100M-R			TVQA	Howto100M-QA
	R@1	R@10	R@100	R@1	R@10	R@100	Acc.	Acc.
XML (Lei et al., 2020)	2.70	8.93	15.34	2.06	8.96	13.27	-	-
STAGE (Lei et al., 2019)	-	-	-	-	-	-	70.50	-
HERO w/o pre-training[8]	2.98	10.65	18.42	2.17	9.38	15.65	70.65	76.89
HERO w/ pre-training	**4.34**	**13.97**	**21.78**	**2.98**	**11.16**	**17.55**	**74.24**	**77.75**

1. Compared to task-specific SOTA models, HERO outperforms with/without pre-training
2. Pre-training greatly lift HERO’s performance on downstream tasks
3. HERO achieves state-of-the-art results on all four downstream tasks

Thank You