# Learning from Inside: Self-driven Siamese Sampling and Reasoning for Video Question Answering

**Weijiang Yu[1], Haoteng Zheng[1], Mengfei Li[1], Lei Ji[2], Lijun Wu[2], Nong Xiao[1], Nan Duan[2]**

**[1]School of Computer Science and Engineering, Sun Yat-sen University**
**[2]Microsoft Research Asia**

NeurIPS 2021 Presentation
Virtual Conference

December 2021

# Task Definition

- Video Question Answering

**Question**:
What happens when the boat gets hit again?

**Answers:**
1.It gets caught in a nearby tree
2.Everybody falls into the water
3.It gets damaged
4.The engine stalls
5.It is forced over a cliff

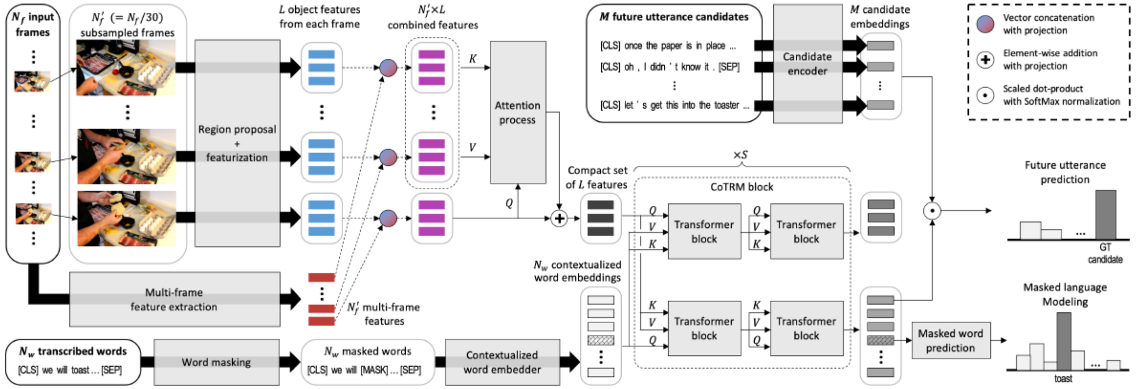
**Question:**
Why do the police conclude that Amy was murdered?

**Answers:**
1.Becausethe killer left a confession in the kitchen of their house
2.Because her murder was recorded on a surveillance camera
3.Becausethey found a body that resembled her floating
4.BecauseNick confesses to killing her
5.Because they find remnants of cleaned blood stains in the house

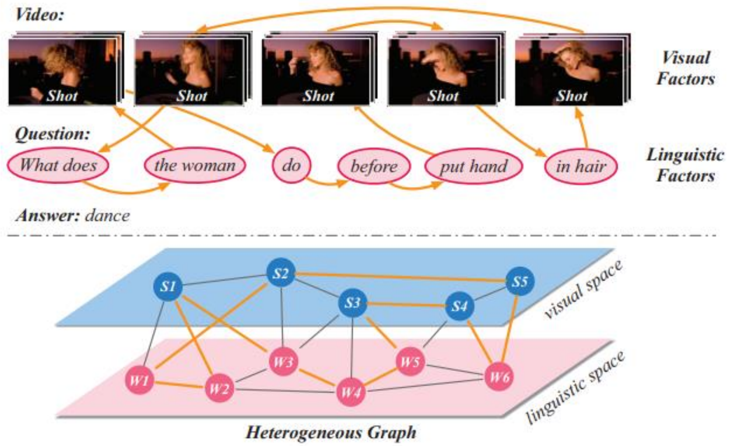
By inferring the correct answers for video-based questions, video question answering (VideoQA) has attracted increasing research attention due to its huge application potential, as a fundamental technique for vision-to-language reasoning.

# Previous Works

- Large-scale Data Pre-training & Structure Reasoning.

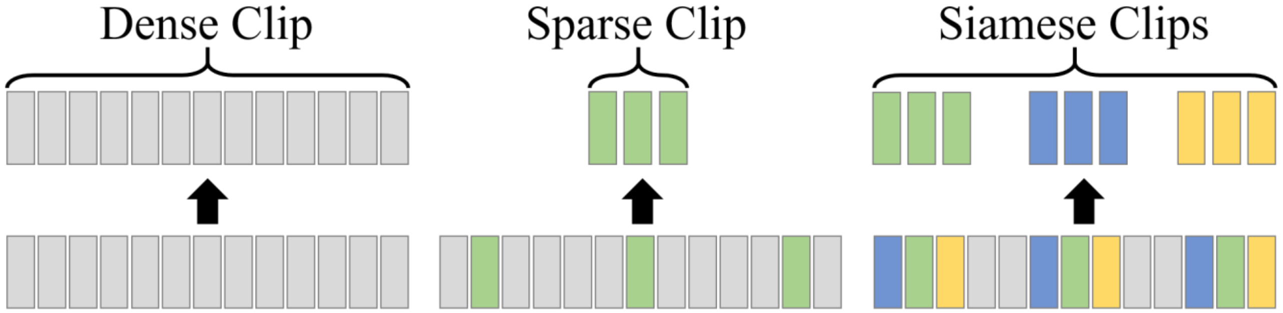

$N_f$ input frames
$N_f'$ $(= N_f/30)$ subsampled frames
Region proposal + featurization
$L$ object features from each frame
$N_f' \times L$ combined features
$K$
$V$
$Q$
Attention process
Multi-frame feature extraction
$N_f'$ multi-frame features
$N_w$ transcribed words
[CLS] we will toast ... [SEP]
Word masking
$N_w$ masked words
[CLS] we will [MASK] ... [SEP]
Contextualized word embedder
$N_w$ contextualized word embeddings
Compact set of $L$ features
$M$ future utterance candidates
[CLS] once the paper is in place ...
[CLS] oh , I didn ' t know it . [SEP]
[CLS] let ' s get this into the toaster ...
Candidate encoder
$M$ candidate embeddings
Vector concatenation with projection
Element-wise addition with projection
Scaled dot-product with SoftMax normalization
$\times S$
CoTRM block
Transformer block
Transformer block
Transformer block
Transformer block
Future utterance prediction
GT candidate
Masked language Modeling
Masked word prediction
toast

Multimodal Transformer-based Network



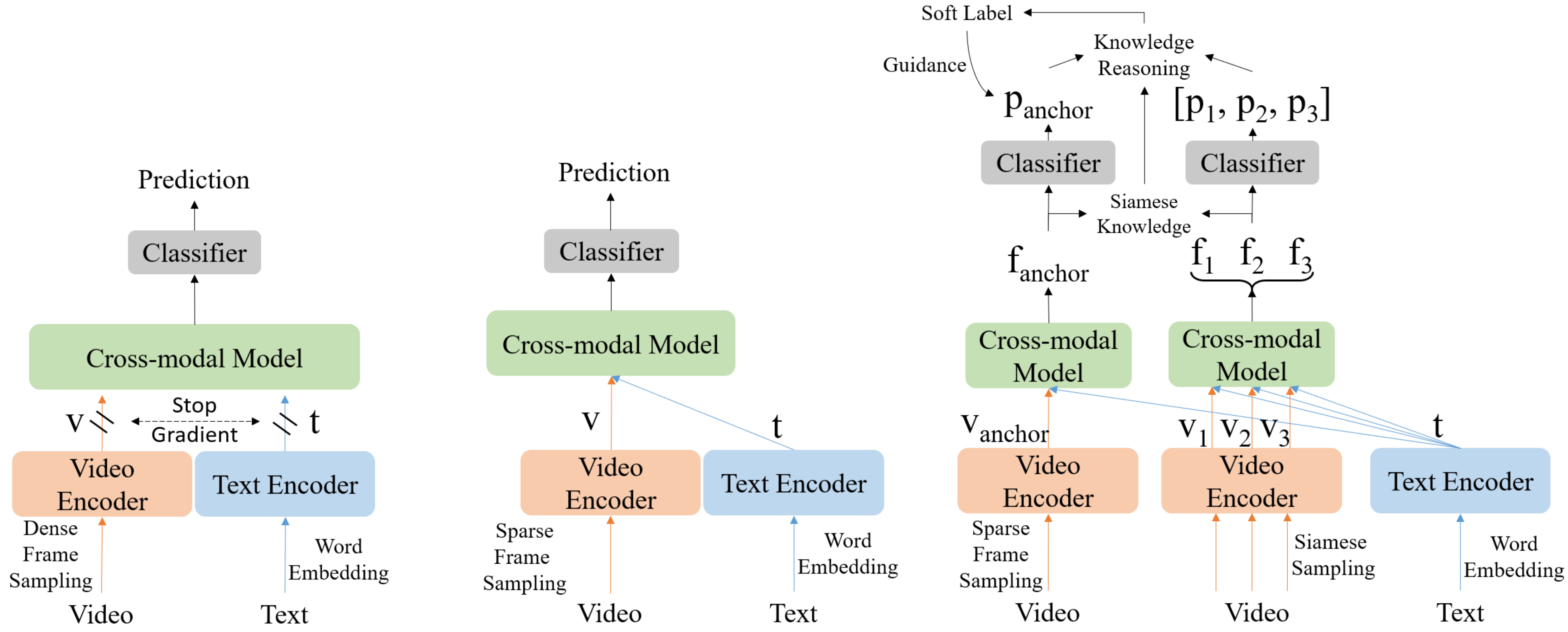
Structure Reasoning for Semantic Alignment

- Different sampling mechanisms for video frames.



(a) Dense Sampling (b) Sparse Sampling (c) Siamese Sampling

# Motivation



(a) Traditional Multimodal Learning

(b) Sparse based Multimodal Learning

(c) Self-driven based Multimodal Learning

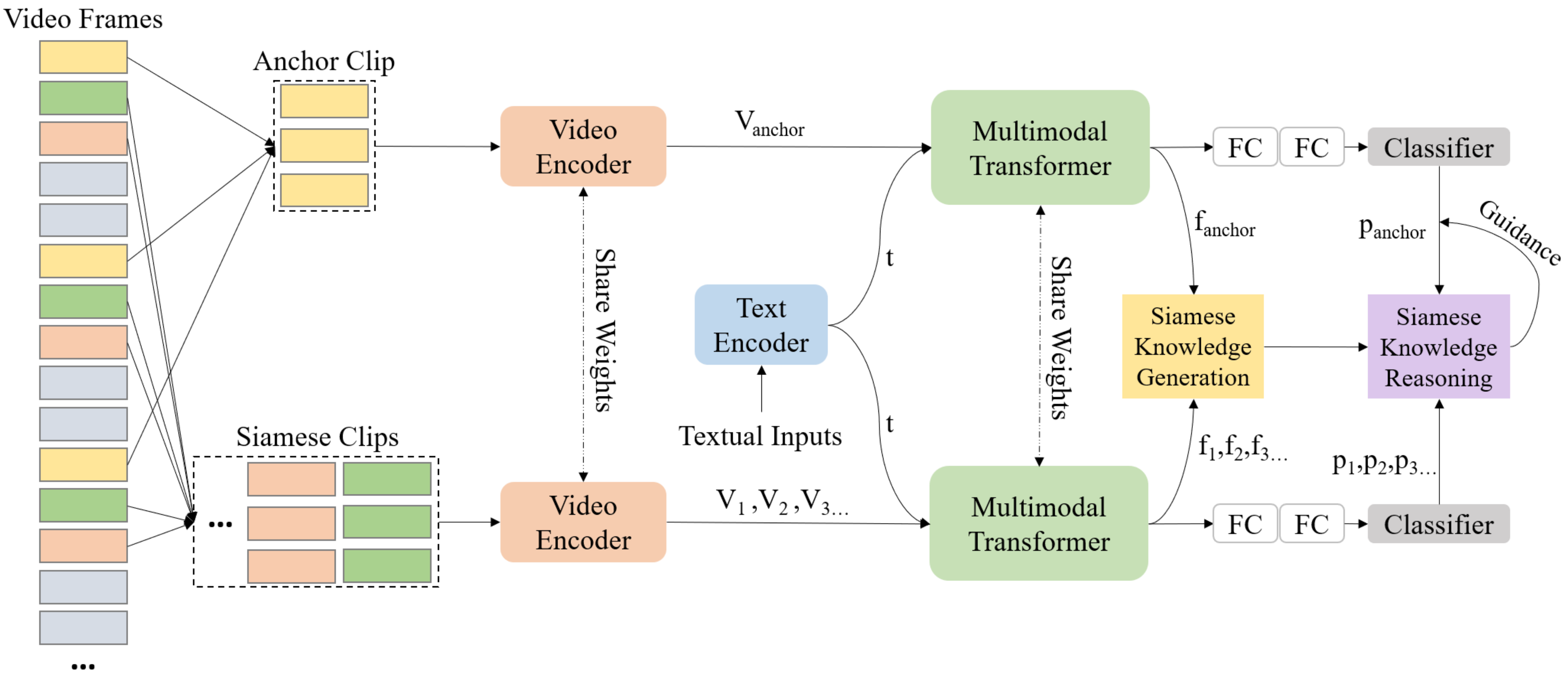
# Contributions

(i) We propose a novel end-to-end framework named SiaSamRea for learning from inside on VideoQA task, by using siamese sampling and reasoning to integrate the interdependent semantics of clips from the same video into the training process.

(ii) A novel reasoning strategy is carefully designed for building the soft guidance from the interdependent knowledge between internal clips, which consists of a siamese knowledge generation module and a Siamese knowledge reasoning module.

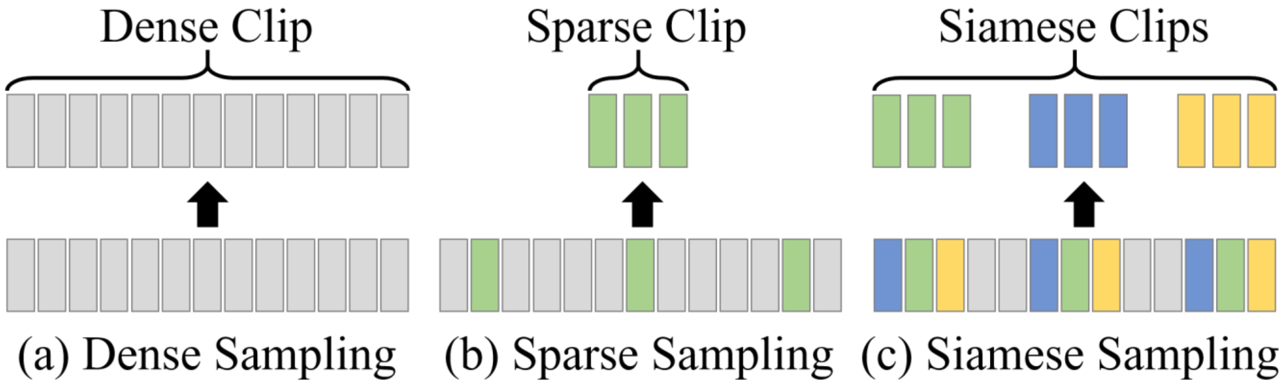
(iii) Experiments on five commonly-used VideoQA benchmarks show the superior ability of our SiaSamRea and demonstrate the effectiveness of our proposed components. Not that our method only teaches the network with interdependent knowledge during the training, which does not bring any extra burden (e.g., computation, memory and parameters) in the inference.

# Our Method – SiaSamRea (Siamese Sampling and Reasoning)



# Our Method – SiaSamRea (Siamese Sampling and Reasoning)

- Siamese Sampling



- Reasoning Strategy
  - Siamese Knowledge Generation
  - Siamese Knowledge Reasoning

# Our Method – SiaSamRea (Siamese Sampling and Reasoning)

- Reasoning Strategy
  - Siamese Knowledge Generation:

$$\mathbf{A}(i,j) = \sigma(F(\mathbf{f}_i))^\top \sigma(F(\mathbf{f}_j))$$

$$\mathbf{A} = \mathbf{A} \odot (1 - \mathbf{I}).$$

$$\hat{\mathbf{A}}(i,j) = \frac{\exp(\mathbf{A}(i,j))}{\sum_{j \neq i} \exp(\mathbf{A}(i,j))}, \ \ \forall i \in \{1, \ldots, N\}$$

- Reasoning Strategy
  - Siamese Knowledge Reasoning:

$$\hat{\mathrm{p}}_i = \sum_{j \neq i} \hat{\mathbf{A}}(i,j)\mathbf{p} = \hat{\mathbf{A}}(i)\mathbf{P}$$

$$\hat{\mathbf{P}} = \mathbf{W}\hat{\mathbf{A}}\mathbf{P}$$

$$\mathbf{Q} = \omega \mathbf{W}_1 \hat{\mathbf{A}}\mathbf{P} + (1 - \omega)\mathbf{W}_2 \mathbf{P}$$

- Optimization
  - Open-ended VideoQA:

$$\mathbf{p} = \phi_\theta(\mathbf{v}, \mathbf{q}), \mathbf{p} \in \mathbb{R}^{|\Omega|}$$

- Optimization
  - Multiple-choice VideoQA:

$$s_m = \phi_\theta(\mathbf{v}, \mathbf{q}, \mathbf{a}_m), 1 \leq m \leq M,$$

- Optimization
  - Loss Function:

$$\mathcal{L} = \alpha \mathcal{L}_{siamese} + \mathcal{L}_{gt}$$

# Experimental Results

| Methods | MSRVTT-QA | MSVD-QA |
|---|---|---|
| E-SA [38] | 29.3 | 27.6 |
| ST-TP [13] | 30.9 | 31.3 |
| AMU [38] | 32.5 | 32.0 |
| Co-mem [9] | 32.0 | 31.7 |
| HME [7] | 33.0 | 33.7 |
| LAGCN [11] | — | 34.3 |
| HGA [15] | 35.5 | 34.7 |
| QueST [14] | 34.6 | 36.1 |
| MiNOR [16] | 35.4 | 35.0 |
| TSN [40] | 35.4 | 36.7 |
| HCRN [21] | 35.6 | 36.1 |
| Clip-BERT [22] | 37.4 | — |
| SSML [2] | 35.1 | 35.1 |
| CoMVT [34] | 39.5 | 42.6 |
| SiaSamRea (Ours) | **41.6** | **45.5** |

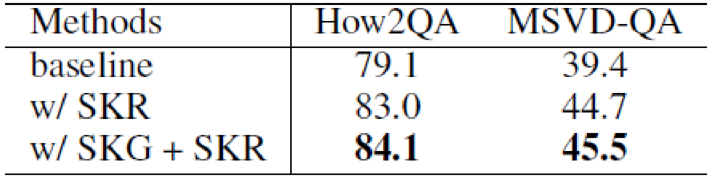
| Methods | ActivityNet-QA | How2QA |
|---|---|---|
| E-SA [48] | 31.8 | — |
| MAR-VQA [52] | 34.6 | — |
| HERO [24] | — | 74.1 |
| CoMVT [34] | 38.8 | 82.3 |
| SiaSamRea (Ours) | **39.8** | **84.1** |

TGIF-QA

| Methods | Action | Trans. | Frame | Count |
|---|---|---|---|---|
| VIS+LSTM (agg) [32] | 46.8 | 56.9 | 34.6 | 5.09 |
| VIS+LSTM (avg) [32] | 48.8 | 34.8 | 35.0 | 4.80 |
| VQA-MCB (agg) [8] | 58.9 | 24.3 | 25.7 | 5.17 |
| VQA-MCB (avg) [8] | 29.1 | 33.0 | 15.5 | 5.54 |
| CT-SAN [47] | 56.1 | 64.0 | 39.6 | 5.13 |
| ST-TP [13] | 62.9 | 69.4 | 49.5 | 4.32 |
| GR-ATT [38] | 68.8 | 73.9 | 53.0 | 4.32 |
| Co-mem [9] | 68.2 | 74.3 | 51.5 | 4.10 |
| PSAC [25] | 70.4 | 76.9 | 55.7 | 4.27 |
| STA [10] | 72.3 | 79.0 | 56.6 | 4.25 |
| MiNOR [16] | 72.7 | 80.9 | 57.1 | 4.17 |
| HME [7] | 73.9 | 77.8 | 53.8 | 4.02 |
| HCRN [21] | 75.0 | 81.4 | 55.9 | 3.82 |
| ClipBERT*[22] | 75.1 | 80.8 | 56.3 | 4.07 |
| HGA [15] | 75.4 | 81.0 | 55.1 | 4.09 |
| SiaSamRea (Ours) | **79.7** | **85.3** | **60.2** | **3.61** |

# Experimental Results

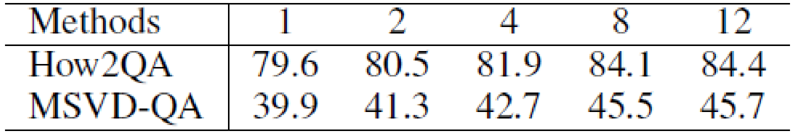
Ablation studies on How2QA and MSVDQA datasets. The SKG and SKR separately indicate siamese knowledge generation and Siamese knowledge reasoning.



| Methods | How2QA | MSVD-QA |
|---|---|---|
| baseline | 79.1 | 39.4 |
| w/ SKR | 83.0 | 44.7 |
| w/ SKG + SKR | **84.1** | **45.5** |

Table 1

The effect of the number of Siamese clips

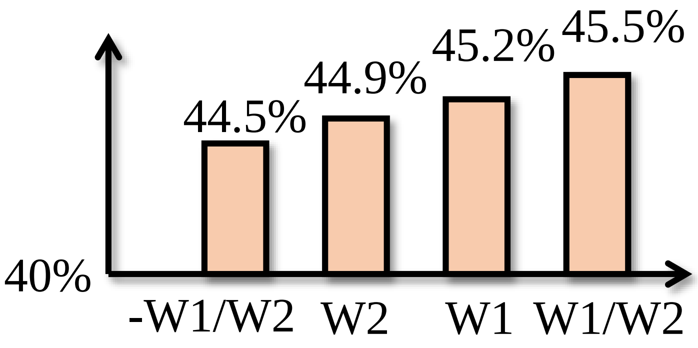


| Methods | 1 | 2 | 4 | 8 | 12 |
|---|---|---|---|---|---|
| How2QA | 79.6 | 80.5 | 81.9 | 84.1 | 84.4 |
| MSVD-QA | 39.9 | 41.3 | 42.7 | 45.5 | 45.7 |

Table 2

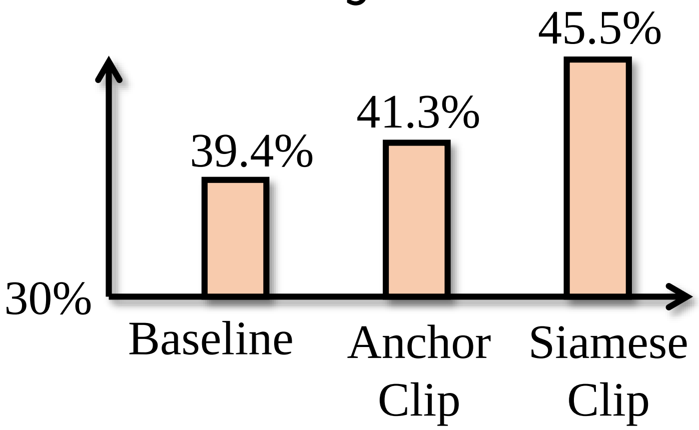
The effect for W1 and W2 in $\mathbf{Q} = \omega \mathbf{W}_1 \hat{\mathbf{A}}\mathbf{P} + (1-\omega)\mathbf{W}_2\mathbf{P}$

Figure 1

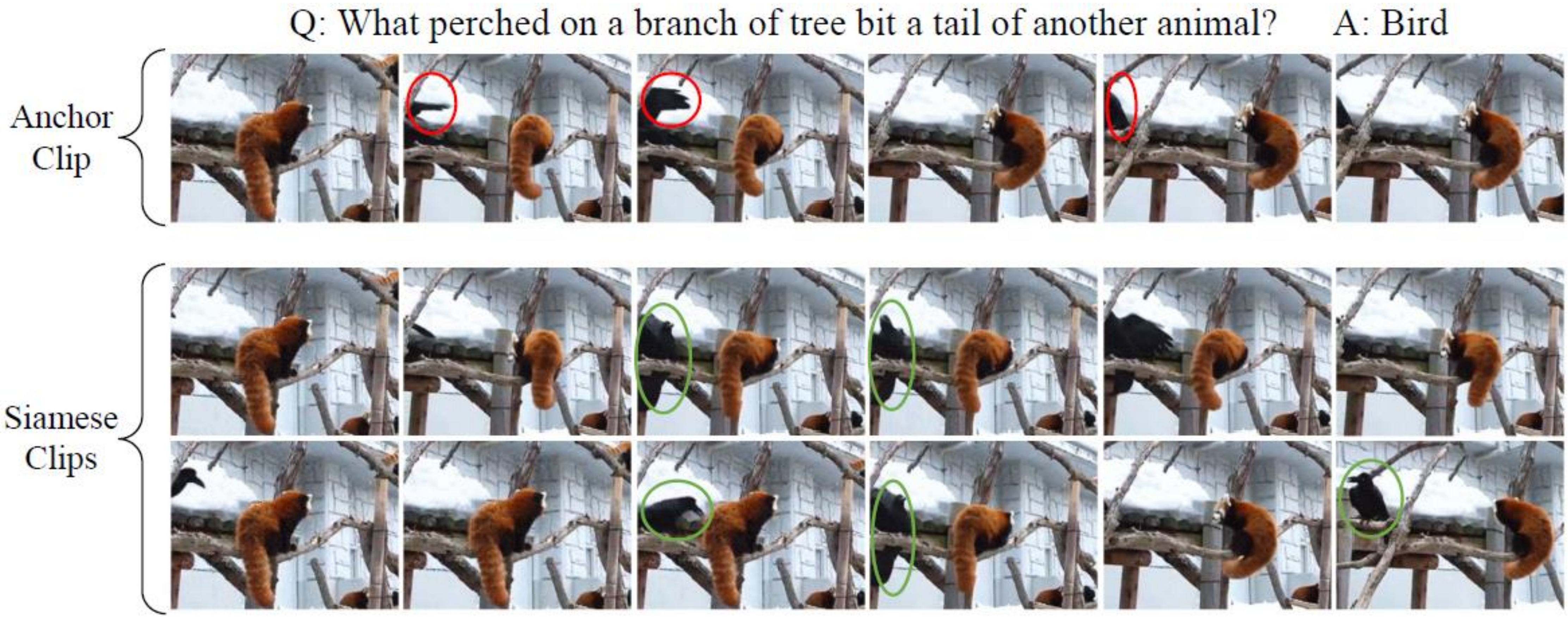


The effect for multiple copies of anchor clip

Figure 2



# Experimental Results


Q: What perched on a branch of tree bit a tail of another animal?
A: Bird
Anchor Clip
Siamese Clips

The examples to show the benefits from the siamese clips. It is hard to discriminate the visual content “bird” from ambiguous parts (red circle) in anchor clip. It is much easier to distinguish “bird” by the assistance of siamese clips that contains more complete visual content (green circle).

# Conclusion & Future Work

In this paper, we propose to endow the current multimodal reasoning paradigm with the ability of learning from inside on the VideoQA task via Siamese Sampling and Reasoning (SiaSamRea), which contains two key parts:

(1) a siamese sampling to produce some sparse clips with similar semantics in the same video;
(2) a reasoning strategy to distill the interdependent knowledge between clips into the network.

The reasoning strategy is composed of two modules:
(i) siamese knowledge generation to implicitly aggregate the inter-relationship of clips from the same video;
(ii) siamese knowledge reasoning to infer soft label by using the predicted candidates of all clips and their inter-relationship.

Our proposed SiaSamRea finally can be jointly evolved by the soft label guidance and ground truth, which is evaluated on five VideoQA datasets demonstrating state-of-the-art performance.

thank you for your listening