Language-aware Visual Semantic Distillation for Video Question Answering

Bo Zou *
Tsinghua University
Beijing, China
zoub21@mails.tsinghua.edu.cn

Chao Yang *
Shanghai AI Laboratory
Shanghai, China
yangchao@pjlab.org.cn

Yu Qiao
Shanghai AI Laboratory
Shanghai, China
qiaoyu@pjlab.org.cn

Chengbin Quan
Tsinghua University
Beijing, China
quancb@tsinghua.edu.cn

Youjian Zhao †
Tsinghua University
Zhongguancun Laboratory
Beijing, China
zhaoyoujian@tsinghua.edu.cn

Abstract

Significant progress in video question answering (VideoQA) have been made thanks to thriving large image-language pretraining frameworks. Although image-language models can efficiently represent both video and language branches, they typically employ goal-free vision perception and do not interact vision with language well during the answer generation, thus omitting crucial visual cues. In this paper, we are inspired by the human recognition and learning pattern and propose VideoDistill, a framework with language-aware (i.e., goal-driven) behavior in both vision perception and answer generation. VideoDistill generates answers only from question-related visual embeddings and follows a thinking-observing-answering approach that closely resembles human behavior, distinguishing it from previous research. Specifically, we develop a language-aware gating mechanism to replace the standard cross-attention, avoiding language's direct fusion into visual representations. We incorporate this mechanism into two key components of the entire framework. The first component is a differentiable sparse sampling module, which selects frames containing the necessary dynamics and semantics relevant to the questions. The second component is a vision refinement module that merges existing spatial-temporal attention layers to ensure extracting multi-grained visual semantics associated with the questions. We conduct evaluations on various challenging video question-answering benchmarks, and VideoDistill achieves state-of-the-art performance in both general and long-form VideoQA datasets. In Addition, we verify that VideoDistill can effectively alleviate the utilization of language shortcut solutions in the EgoTaskQA dataset.

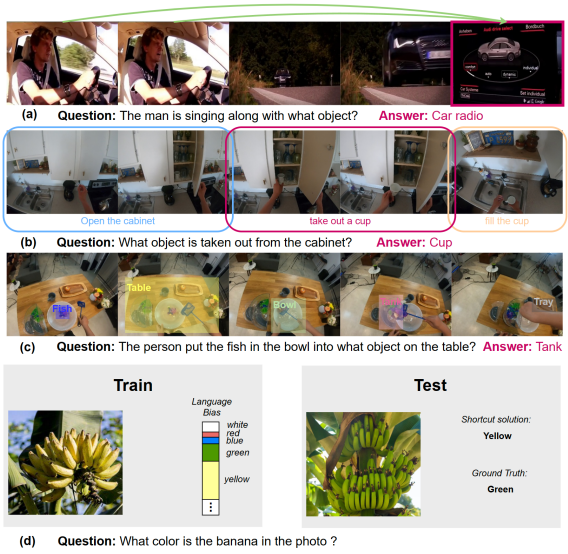

Figure 1. **Challenges of goal-free VideoQA models.** They can not efficiently handle (a) Long-term dependencies, (b) Multi-events, and (c) Multi-scale semantics in the videos. They also suffer from (d) language prior phenomenon in training question-answer pairs.

1. Introduction

In recent years, large-scale video-and-language pretraining has seen remarkable progress. Most modern video-language understanding models [4, 15, 26, 31, 35, 36, 50] independently encode uni-modality and then fuse them. Human motion perception studies [5, 17, 53] suggest that humans perceive motion and environments as goal-driven behavior. This discrepancy leads to several issues, especially in long-form video understanding.

First, goal-free video representation struggles with long-term dependencies and multi-event reasoning. While enabled by progress in cross-modal pretraining[29, 43, 46] and datasets[39, 40], current methods excel at question an-

* Equal contribution, † Corresponding author
This work was done during an internship at Shanghai AI Lab.

swering for images and short videos, but long-form videos contain many clips irrelevant or redundant to the questions, interfering with overall understanding. Encoding long videos also brings immense computational costs. Second, accurate semantic reasoning usually relies on multi-scale perception from local-spatial regions to global temporal dynamics. Goal-free methods for multi-scale visual modeling require custom submodels[9, 52, 56] for each scale or extra modalities like bounding boxes and OCR features[23, 66] but large-scale pretraining makes these approaches inefficient or infeasible. Third, goal-free video embedding requires a multimodal fusion module to synthesize questions and visual embeddings to predict answers. Incorporating questions directly can lead to shortcut solutions [1, 2, 8, 12, 19, 25, 41, 65], which means utilizing obvious clues in questions (mainly in data distribution and the relation between keywords) that exhibit more reliability in answer prediction than complicated visual reasoning, especially in the early training phases. This phenomenon is also known as language bias, which often causes major performance gaps in out-of-distribution tests. The above challenges are visualized in Figure 1.

To address these issues, we propose a language-aware (goal-driven) visual semantic distillation framework called VideoDistill. Semantic distillation means the visual encoder must embed relevant frames of questions and focus on question-related multi-scale visual semantics. Closely resembles human behavior, semantic distillation functions like keeping the goal (question) in mind and conducting meaningful and precise visual reasoning. Differing from previous VideoQA frameworks [3, 31, 32, 42, 45, 59, 60], VideoDistill can generate answers only from question-related visual embeddings (without additional textual embeddings) since the interaction of text and vision is adequate during semantic distillation. This feature also enhances the significance of visual reasoning when generating answers.

To realize semantic distillation, we first introduce Language-Aware Gate (LA-Gate), a multi-head cross-gating mechanism for cross-model interaction, which is inspired by self-gating and gated attention[14, 22, 37, 51] and is integrable into any existing transformer-based models like[4, 50]. LA-Gates compute questions' dependencies on video patch embeddings and depress or excite corresponding patches in subsequent attention layers. The proposed LA-Gate is not only the key to semantic distillation but also a new efficient modality fusion method, which is a powerful competitor of predominant Cross-Attention in VideoQA. We discuss the differences between LA-Gate and attention mechanisms in Section 3.2 and summarize three merits of LA-Gate: (1) It can alleviate the influence of language bias by avoiding the direct involvement of text embeddings. (2) It can better maintain local diversity within the video embeddings. (3) It enhances the interpretability of the modality fusion process.

VideoDistill has two LA-Gate-based modules. The first is a differentiable sparse sampling module, which uses pretrained image-language models like CLIP[43] to encode frames, then performs goal-driven frame sampling to remarkably reduce subsequent spatial-temporal attention overhead and naturally avoid long-term dependencies and multi-event reasoning by retaining only question-related frames. It also provides our framework with a good nature that is insensitive to the number of sampled frames (see section 4.3).

The second is a vision refinement module eliminating unrelated visual semantics at different perceptual levels and enhancing related multi-scale semantics to support multi-level refinement. It encodes sparsely sampled frames into question-related global embeddings for generating answers. This module brings obvious performance boosts, especially on object-related questions (see section 4.5). Our contributions are summarized as follows:

- We propose the Language-Aware Gate (LA-Gate) that enables interacting vision with language meanwhile not directly introducing language into visual representations. It can alleviate the influence of language bias, better maintain local diversity within the video embeddings, and is more interpretable compared with predominant Cross-Attention.
- Based on LA-Gate, we propose a differentiable sparse sampling module to capture question-related frames and a vision refinement module to emphasize multi-scale question-related semantics. They can benefit VideoQA models on the stability under various numbers of sampled frames and the capacity to understand multi-scale objects.
- Our model achieves new state-of-the-art performance on a wide range of downstream VideoQA tasks and text-to-video retrieval tasks.

2. Related Works

2.1. Long-form Video modeling

Long-form understanding has recently become a new research hot spot [11, 16, 31, 34, 47, 48, 54, 55, 59, 64]. Current methods develop in two directions to overcome unique challenges not encountered by previous short-term video understanding works. The first direction is to encode long-term dependencies better. LF-VILA [48] proposes hierarchical temporal window attention that begins with a small window learning the attention between adjacent frames, then gradually expands the window size to learn high-level representation. HERO [34] and MERLOT [64] predict the order of shuffled frames to understand sequential characteristics. The second direction aims to reduce the increasing computation cost of encoding longer videos. Since the common practice of randomly selecting 3 or 4 frames per video regardless of length is not suited for long videos, ClipBert [31] randomly samples a sequence of segments from a video and then aggregates their predictions. MIST [16] improves ClipBert by selecting both video segments and frame regions to be encoded and reduces computation costs further.

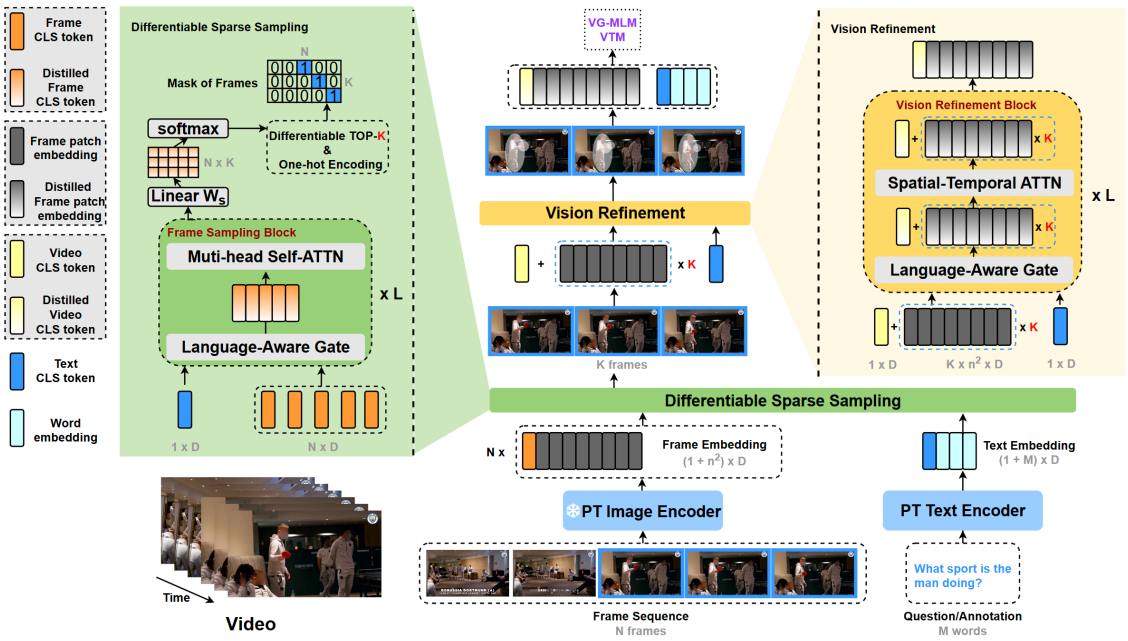

Figure 2. **Overview of VideoDistill.** VideoDistill first densely samples video frames and utilizes a pre-trained image-language encoder to extract features, then sparsely samples a small number of question-related frames by a differentiable sparse sampling module. Finally, VideoDistill uses a vision refinement module to emphasize necessary multi-scale visual semantics in selected frames.

TALLFormer [11] proposes short-term feature extraction and long-term memory mechanisms that avoid repetitive calculations during training.

Similar to the second category, we reduce the workload by decreasing inputs, but our goal-driven differentiable sparse sampling works directly on frames, the minimum composition of the video, rather than rough video segments. Kim et al. [27] propose to use a self-gating mechanism to realize frame sampling, but it is a goal-free sampling method. Also, unlike random selection [31] or simple selector based on similarity[16], our learnable sampler excavates the semantic relationships between visual candidates. In addition, sparsely sampled frames from our sampler can be considered a concise summary of the long video sequence. Therefore, the following inference can perform similarly to models on short-form inputs. Thus, VideoDistll does not make efforts in the first direction.

2.2. Video-Language Fusion in VideoQA

The answer generation of VideoQA can be formulated as the interaction of the two modalities. Some methods propose fusing language and vision in the feature extraction stage for a more interactive combination. PSAC [36] combines regular visual self-attention with visual-linguistic co-attention. HCRN [30] supports conditioning video features with linguistic clues as a context factor in every inference stage. PMT [42] adopts a pyramidal video-language interaction. MCAN[63] utilizes Guided-Attention (V2T cross-attention in this paper) to fuse language into visual feature extraction.

Unlike these techniques, our VideoDistill only keeps the question "in mind." We do not directly fuse the question information into the video embedding, thus generating "purer" visual embeddings. Since the answer prediction is only based on visual embeddings, VideoDistill can alleviate just assuming the answer from the question and follow the "look and answer" criterion [2].

Besides VideoDistill, some existing methods can also be considered as indirect fusion. Contrastive learning methods like [10, 18, 20, 43, 57, 67] pull the matched pairs closer to interact vision with language. However, they do not demonstrate the same level of performance in tasks other than retrieval [28] due to ignoring the misalignment of vision and language. The method of fine-grained contrast to mitigate the misalignment has become a recent research hot spot.

3. Methodology

We introduce VideoDistill, a pretraining framework that enables goal-driven VideoQA relying on question-related frames and their multi-scale semantics. Figure 2 gives an overview of VideoDistill. VideoDistill consists of two submodules: **Differentiable Sparse Sampling** and **Vision Refinement**. Both sub-modules are built upon our proposed **language-aware gate** (LA-Gate) to perform in a goal-driven manner.

3.1. Vision-Language Representations

VideoDistill begins with densely sampling N uniformly distributed frames and embedding them into frame represen-

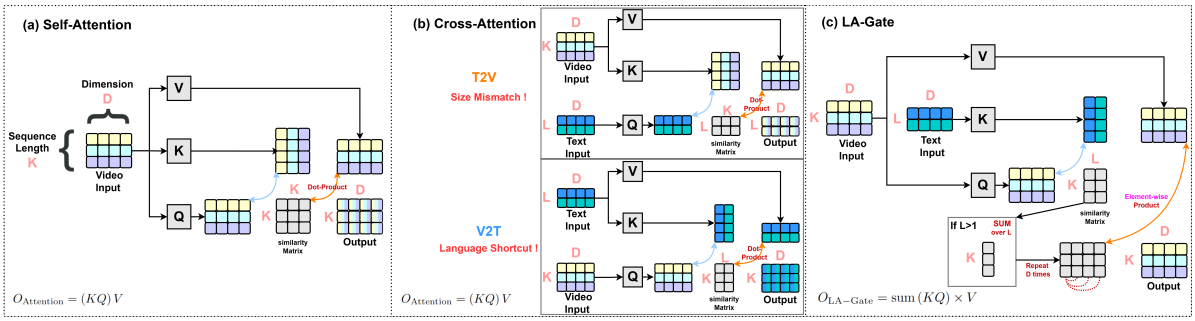

Figure 3. Illustrations of Self-Attention, Cross Attention, and our LA-Gate mechanisms

tations. We adopt the pretrained image-language encoder CLIP [43] with frozen parameters in the visual branch to reduce the computational overhead. CLIP divides N frames into $N \times n \times n$ patches, and extracts patch embeddings $v_{\text{patch}} \in \mathbb{R}^{N \times n^2 \times D}$ and CLS tokens $v_{\text{cls}} \in \mathbb{R}^{N \times D}$ that represent the global understanding of frames, where D is the dimension of representations.

For the Language branch, A text is first tokenized and then fed into the pretrained text encoder of CLIP to generate word-level embeddings $t_{\text{word}} \in \mathbb{R}^{M \times D}$ and a sentence-level token $t_{\text{cls}} \in \mathbb{R}^{D}$, where M is the length of embeddings. Note that parameters in the text encoder are updated during pretraining and downstream finetuning to mitigate the linguistic gaps between datasets.

3.2. Language-Aware Gate

We introduce a reusable unit, termed language-aware gate (LA-Gate), which takes an arbitrary visual representation v and a sentence-level language representation t_{cls} as input. LA-Gate generates v_{distill}, an output sequence of language-aware visual representation with the same dimension as v, by exciting or depressing the components of v according to their similarities with t_{cls}. For a better understanding of LA-Gate, we illustrate self-attention, cross-attention, and LA-Gate in Figure 3. We first discuss why commonly used cross-attention is not suitable for our purposes, then point out the design concept and benefits of LA-Gate.

Figure 3(b) illustrates two types of cross-attention: Text-to-Video (T2V, text as $Query$) and Video-to-Text (V2T, video as $Query$). Since outputs of the attention have the same length as $Query$ and carry information sourced from $Value$, applying cross-attention in VideoQA will have two problems. Firstly, T2V is unsuitable for the modality interaction in the middle layers as it alters the shape of visual representations. Although V2T can maintain the shape, its output will carry the info directly from the text (like reconstructing text representations into the shape of visual input).

The structure of LA-Gate is shown in Figure 3(c). There are two divergences between V2T and LA-Gate. (1) When text input length $L > 1$, we sum the similarity matrix to form a vector of importance for rows of the video input. L functions like the number of multi-heads in the attention mechanism, and each text representation will partially dictate the importance of each row in the video input. (Note that in VideoDistill, we always have $L = 1$ since the text input is t_{cls}.) (2) We expand the importance vector (repeat D times) to match the shape of the video input. Finally, we apply an element-wise product rather than a dot product on the importance matrix and the Value.

The first merit of LA-Gate is the text input only controls the importance of each visual representation, and the output does not directly involve text. Thus, the answer decoder can make decisions based on "purer" visual semantics and alleviate shortcut solutions hidden in texts.

A more significant characteristic of LA-Gate in general modality fusion usage: it can better maintain local diversity within the visual input. In Figure 3 (a) and (b), the output displays rows in mixed colors, representing the weighted sum of rows in the $Value$. Cross-attention will gather information from all rows in the $Value$ for each output token and present more global attributes than LA-Gate since LA-Gate does not blend all features. In Figure 4(c), we maintain the uniqueness of colors for LA-Gate in each output row because each row is produced by multiplying the corresponding row in the $Value$ and a scaler of importance. This characteristic makes LA-Gate fit more into our idea of information distillation because the language-aware distillation of a local area should be irrelevant to other regions.

The last benefit of LA-Gate is we can enhance the interpretability of modality fusion by incorporating LA-Gate with self-attention layers. Due to the skip connection between the input and the output, commonly used V2T cross-attention brutally adds a reconstructed text (maybe linearly projected) into the visual input to fuse vision and language. The physical meaning of the structure is difficult to explain. By contrast, LA-Gate first distills the visual input by emphasizing language-related visual semantics, and then the interaction within visual representations is performed by self-attention.

The implementation of LA-Gate is visualized in Figure 4 (we demonstrate how LA-Gate works in a vision refinement block, where v consists of frame patches from K selected frames and a video-level CLS token). Assume LA-Gate

receives $v \in \mathbb{R}^{m \times D}$, where m is the sequence length, and $t_{\text{cls}} \in \mathbb{R}^{D}$. We first produce the key from the language representation t_{cls} and queries and values from the visual representation v. Then, we calculate cosine distances between each query and the key:

$$\text{key} = \text{w}_\text{k}\left(t_{\text{cls}}\right), \text{query}_i = \text{w}_\text{q}\left(v_i\right), \text{value}_i = \text{w}_\text{v}\left(v_i\right) \quad (1)$$

$$\text{dist}_i = \text{repeat}\left(\frac{\text{query}_i \cdot \text{key}}{\|\text{query}_i\| \times \|\text{key}\|}\right), \quad (2)$$

where $v_i \in \mathbb{R}^{D}$, $i \in [1, m]$ is a feature vector in v. w_k, w_q and w_v are trainable linear projection layers, and repeat means expanding a scalar into a vector with the dimension of D. Since dist_i reflects the correlation between vision and language, we treat it as an importance coefficient of value_i. Then, we generate distilled v_i as formulated:

$$v^i_{\text{distill}} = \text{w}_\text{o}\left(\text{dist}_i \odot \text{value}_i\right), \quad (3)$$

where $\odot$ denotes element-wise production and w_o is the linear output layer. In practice, we perform a standard multi-head attention setting [49] on w_k, w_q, w_v and W_o, and concatenate outputs of each head to form v^i_{distill}. Although, we only illustrate a single attention head in Figure 2 for simplicity. Also, we apply an input skip connection on v^i_{distill} for better training stability and convergence. Finally, LA-Gate outputs $v_{\text{distill}} = \left\{v^i_{\text{distill}} | i \in [1, m]\right\}$.

3.3. Differentiable Sparse Sampling

Given N densely sampled frames, VideoDistill further adaptively picks out $K\,(K < N)$ language-related frames. To this end, we utilize stacked frame sampling blocks (FS-Blocks, L layers) which take as input v_{cls} and t_{cls} to perform a top-K selection. Since v_{cls} of frames are separately extracted, to indicate their temporal positions in the whole video, we add temporal embedding $\text{T}_\text{f} \in \{\phi_\text{T}\left(\text{f}\right) | \text{f} \in [1, \text{N}]\}$ for each of them according to their frame index. After adding temporal information, each FS-Block first distills v_{cls} conditioned on t_{cls} by LA-Gate to emphasize question-related frames, then performs an inter-frame interaction through standard multi-head attention. Figure 4 shows the architecture of FS-Blocks. We borrow the form of skip connections from the divided block in [4].

To realize the differentiable selection, we first project v^L_{cls}, the output of the L-th FS-Block with the dimension identical to v_{cls}, onto a feature space with the dimension of K by a linear layer W_s, where K is the number of frames to be selected. Then, we conduct Gumbel-Softmax sampling [24] on each $x^k \in \mathbb{R}^N, k \in [1, K]$, the row of $\text{W}_\text{s}\left(v^L_{\text{cls}}\right)$. The procedures are formulated as follows:

$$y^k_{\text{soft}} = \text{softmax}\left(\frac{x^k + gumbels}{\tau}\right), \quad (4)$$

The code will be available at https://zoubo9034.github.io/VideoDistill/

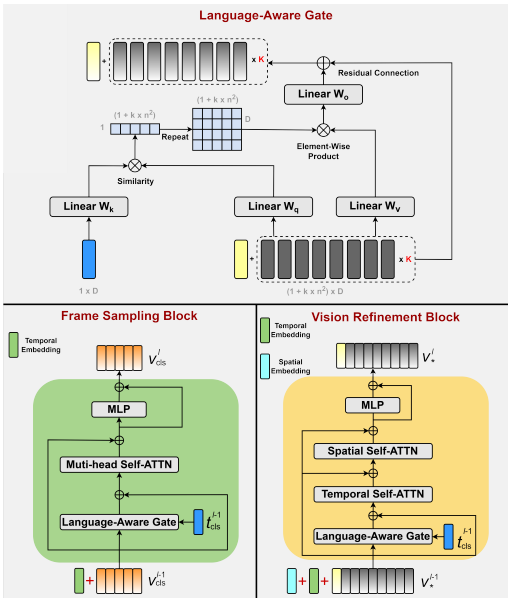

Figure 4. Architectures of Language-Aware Gate, Frame Sampling Block, and Vision Refinement Block.

$$y^k_{\text{hard}} = \text{onehot}\left(\text{argmax}\left(y^k_{\text{soft}}\right)\right), \quad (5)$$

$$mask^k = y^k_{\text{hard}} + y^k_{\text{soft}} - \text{stopgrad}\left(y^k_{\text{soft}}\right), \quad (6)$$

where $gumbels$ is a noise sampled from the Gumbel distribution with $\mu = 0$ and $\beta = 1$, y^k_{soft} is the possibility of frames to be selected as the k-th language-related frame and y^k_{hard} reflects the index of the chosen frame. Since the argmax operation has no gradient, we adopt a code-level trick in equation 6 to generate $mask = \left\{mask^k | k \in [1, K]\right\}$, which can properly pass the gradient. Finally, we apply $mask$ on v_{patch} and generate $v^K_{\text{patch}} \in \mathbb{R}^{K \times n^2 \times D}$ for the following vision refinement.

3.4. Vision Refinement

Given the selected patch embeddings v^K_{patch} and the language representation t_{cls}, this module generates video-level representation v^*_{cls}, which synthesizes multi-scale language-related visual semantics. v^*_{cls} is set as a learnable token with the dimension of D and is concatenated to v^K_{patch} as the visual input v_*. We combine the LA-Gate and any existing spatial-temporal self-attention layer [4, 6, 50] to form a vision refinement block as shown in Figure 4. When iteratively applying vision refinement blocks on v_*, LA-Gates spontaneously distill language-related visual semantics in different perceptive fields. Since the effective perceptive field gradually expands in stacked blocks [44], our module can consider multi-scale objects.

Unlike predominant pyramidal models [33, 42, 61] for multi-scale reasoning that gathers the intermediate results of each encoding stage as final representation, we empirically adopt only the v^*_{cls} from the last vision refinement block.

3.5. Pretraining Tasks

We adopt the two pretraining tasks to facilitate cross-modal interaction: Video-Text Matching (VTM) and Vision-Guided Masked Language Modeling (VG-MLM).

Video-Text Matching. To facilitate cross-modal interaction, we use VTM as the first pre-training task. VTM predicts whether input video-language pairs are matched. In practice, we randomly exchange the annotations of two input pairs (matched) in a mini-batch with a probability of 0.5. Then, we use a linear projection head on the top of v^*_{cls} to predict two-class matching logits y, and formulate the VTM loss as negative log-likelihood:

$$\mathcal{L}_{\text{VTM}} = -\mathbb{E}_{(v^*_{\text{cls}})} \log p\left(y \mid v^*_{\text{cls}}\right), \tag{7}$$

We do not adopt predominant contrastive pretraining because this paradigm requires the contrast between all possible combinations of vision and language. However, for VideoDistill, the number of possible visual embeddings of each video is proportional to the number of annotations. This leads to a quadratic growth in computational overhead (see Appendix A for more details). Nevertheless, we still apply a contrastive constraint on matched pairs (w/o exchange) to stabilize the training:

$$\mathcal{L}_{\text{CL}} = -\sum_{i \in B} \log \frac{\mathbb{I}\left(\text{matched}\right) \exp\left(s\left(v_i, t_i\right)/\tau\right)}{\sum_{j \in B} \exp\left(s\left(v_i, t_j\right)/\tau\right)}, \tag{8}$$

where B is the batch size, v_i and t_i are v^*_{cls} and t_{cls} of i-th video-language pair, $\mathbb{I}$ is a indicate function of whether the video and the annotation in the i-th pair are matched, s denotes a cosine similarity function, τ is a temperature coefficient equals 0.07 in this paper.

Vision-Guided Masked Language Modeling. The commonly used MLM [13] task aims to predict the ground truth label of the masked token according to textual context. To better map the visual and the language representations in a fine-grained manner, we propose VG-MLM, which is applied only on the matched pairs (w/o exchange) and encourages predicting the masked tokens from the visual context. In particular, we first provide VideoDistill the CLS tokens t_{cls} of unmasked annotations to sample frames and extract video representation v^*_{cls}. Then, we mask I words in the annotations and encode the masked sentences. The outcome $w^{\text{M}}_i, i \in [1, I]$, which denotes the token of i-th masked word, and v^*_{cls} are used to predict the i-th masked word. The objective of VG-MLM is formulated as follows:

$$\mathcal{L}_{\text{VG-MLM}} = -\mathbb{E}_{(t_{\text{cls}}, v^*_{\text{cls}})} \log p\left(w_i | \text{SG}\left(w^{\text{M}}_i\right), v^*_{\text{cls}}\right), \tag{9}$$

where SG means stop gradient, w_i denote the logit of i-th masked word. We stop the gradient of w^{M}_i to enhance the importance of visual clues in language modeling. Following the heuristics of Bert [13], we adopt the same masking strategy. Then we adopt a two-layer MLP on top of the combination of w^{M}_i and v^*_{cls} to generate the probability over the vocabulary, which is calculated as the negative log-likelihood loss for the masked word. The overall loss of pretraining is the combination of VTM, VG-MLM, and CL:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{VTM}} + \mathcal{L}_{\text{VG-MLM}} + \mathcal{L}_{\text{CL}}, \tag{10}$$

4. Experiments

In this section, conduct extensive experiments utilizing the pre-trained model on downstream various types of VideoQA tasks to verify the effectiveness of our proposed VideoDistill. We also transfer VideoDistill for video-text retrieval tasks to show the generalization power of the pre-trained model in Appendix C. Our pretraining set consists of three parts: (1) 3M video-caption pairs randomly sampled from generic dataset WebVid10M [4]. (2) 4.2M video-caption pairs randomly sampled from YouTube video dataset HD-VILA [59]. (3) 3.8M video-caption pairs from 1-st person view dataset EgoCLIP. We report results with $N = 100$, $K = 16$, and $L = 3$. For open-ended datasets, train a MLP classification head on the top of v^*_{cls}. For multiple-choice datasets, we choose the answer with the maximum logit over the VTM head. More details are in Appendix B.

4.1. Generic VideoQA:

We first evaluate VideoDistill on the four commonly used VideoQA datasets: **MSRVTT-QA** [58], **MSVD-QA** [58], **EgoMCQ** [38] and **MSRVTT-multiple-choice test** [62]. Details of each dataset and results are in the supplementary material (Appendix C).

4.2. Long-Form VideoQA:

We evaluate our model on two recently proposed challenging datasets for the long-form VideoQA, namely EgoTaskQA[25] and AGQA[21]. We adopt the same fine-tuning setting with MSRVTT-QA and MSVD-QA. For fairness comparison, we report the results of VideoDistill w/o large-scale pertaining. Details of each dataset are in the supplementary material.

Results. In the Table 2, our VideoDistill overwhelm previous SOTA method and bring 2.82% overall performance gains. Remarkably, in the categories of multi-agent and descriptive, we outperform previous work with 5.59% and 3.72% improvement. They show VideoDistill can better understand the storyline and temporal relations in long videos. In the category of object, we achieve 9.33% gains, which indicates VideoDistill can capture multi-scale objects. The limited performance in the category of action is due to the strong correlations between action, object, and their change. Most methods tend to over-fit this strong language bias without thorough task understanding. Although the design of VideoDistill avoids utilizing the shortcut solutions in language. In Table 1, The results in the gray color (AIO, Temp, and MIST-CLIP) are reported by [16]. If they follow the same evaluation metrics [21] as us, one of the reasons for the

Table 1. QA accuracies of state-of-the-art methods on AGQA v2 test set.

Question Types	Most Likely	PSAC	HME	HCRN	AIO [50]	Temp[ATP] [7]	MIST-CLIP	**VideoDistill†**
Object-relation	9.39	37.84	37.42	40.33	48.34	50.15	51.68	**56.32**
Relation-action	50.00	49.95	49.90	49.86	48.99	49.76	**67.18**	55.09
Object-action	50.00	50.00	49.97	49.85	49.66	46.25	**68.99**	55.12
Superlative	21.01	33.20	33.21	33.55	37.53	39.78	42.05	**43.30**
Sequencing	49.78	49.78	49.77	49.70	49.61	48.25	**67.24**	54.49
Exists	50.00	49.94	49.96	50.01	50.81	51.79	**60.33**	55.74
Duration comparison	24.27	45.21	47.03	43.84	45.36	49.59	**54.62**	49.08
Activity recognition	5.52	4.14	5.43	5.52	18.97	18.96	**19.69**	10.16
All	10.99	40.18	39.89	42.11	48.59	49.79	54.39	**55.80**

Table 2. Performances on EgoTaskQA *normal* split. †denotes training from scratch.

	Category	VisualBERT [35]	PSAC [36]	HME [15]	HGA [26]	HCRN [30]	ClipBERT [31]	**VideoDistill†**
Scope	world	39.73	40.76	41.91	38.82	44.27	42.15	**47.32**
	intent	44.51	46.19	48.92	42.12	49.77	40.94	**52.53**
	multi-agent	26.29	30.59	27.98	23.43	31.36	27.63	**36.95**
Type	descriptive	41.99	40.63	41.45	38.04	43.48	38.45	**47.20**
	predictive	30.37	31.98	35.88	25.57	36.56	31.50	**40.43**
	counterfactual	41.99	41.89	44.13	41.94	48.00	46.75	**49.64**
	explanatory	37.42	37.99	38.85	35.97	40.60	42.39	**42.53**
Semantic	action	15.02	14.75	14.99	15.08	14.92	**22.91**	16.35
	object	23.26	36.53	36.05	19.09	45.31	21.80	**54.64**
	state	59.20	61.89	63.44	55.65	68.28	54.36	**72.37**
	change	68.27	65.05	68.87	68.38	67.38	66.58	**71.47**
	all	37.93	38.90	40.16	36.77	42.20	39.87	**45.02**

performance gaps in some categories is they adopt frames with higher resolution (448×448) and more powerful encoder (CLIP-ViTB/32). Nevertheless, we still outperform current SOTA methods in Object-relation (4.64%) and Superlative (1.25%), which require multi-scale and multi-event reasoning. We also improve the overall performance with 1.41% gains.

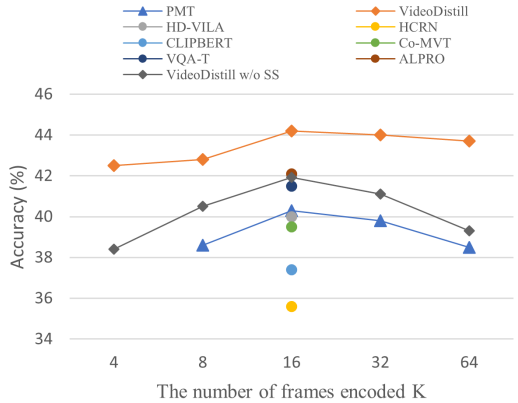

Figure 5. The impact of the number of frames.

4.3. The Impact of Differentiable Sparse Sampling

Insensitive to the number of frames. We evaluate the performances on MSRVTT-QA with variable input frames. In Figure 5, we compare the performance change of VideoDistill (w/ & w/o sparse sampling module) with PMT [42] (other methods only report results on 16 frames, which are the best results), and there are two findings. First, the differentiable sparse sampling module makes our 4-frame variation on par with other methods with many more input frames. It is intuitive because other competitors require more frames to improve the change to capture necessary information for reasoning (a disadvantage of goal-free perception). This property can dramatically reduce the computational overhead required to achieve the same performance as previous methods. Second, we notice that too many frames will damage the answering accuracy since the emergence of unrelated information disturbs the reasoning process. However, our sparse sampling module can mitigate the dropping with increasing frames and make the model more robust.

Qualitative Results. We visualize the results of our sparse sampling module on EgoTaskQA in Appendix E.

4.4. The Impact of LA-Gate:

Prevent the use of language shortcuts. We come up with two methods to quantitatively analyze this characteristic. First, we test the language-only QA performances on EgoTaskQA normal split in Table 3, which means replacing the visual inputs with a static video (We also report the performances under a Gaussian noise input in Appendix E). The more declines, the better. When compared with the language-based model Bert [13], and the previous SOTA method HCRN [30], the decline of our VideoDistill is the most significant (especially in the object category). It demonstrates that VideoDistill relies more on vision and can avoid using language bias. Second, we test VideoDistill on Ego-

Table 3. Language-only QA results on the EgoTaskQA *normal* split. **The more performance drops, the better.**

Category	BERT		HCRN (w/o vision)		**VideoDistill (w/o vision)**	
	Acc.	Change	Acc.	Change	Acc.	Change
world	36.28	-8.7%	30.17	-31.9%	18.78	-60.3%
intent	35.02	-21.3%	35.54	-28.6%	22.57	-57.0%
multi-agent	20.58	-21.7%	19.9	-36.5%	11.50	-68.7%
descriptive	34.55	-17.7%	29.97	-31.1%	21.19	-55.1%
predictive	24.75	-18.5%	18.32	-49.9%	7.42	-81.6%
counterfactual	41.3	-1.6%	41.1	-14.4%	39.70	-20.0%
explanatory	31.78	-15.1%	32.41	-20.2%	13.69	-67.8%
action	15.72	+4.6%	17.31	+16.0%	11.81	-27.8%
object	7.43	-68%	8.85	-80.5%	0	-100.0%
state	45.03	-23.9%	35.51	-48.0%	20.21	-72.1%
change	69.87	+2.3%	70.47	+4.6%	55.38	-22.5%

Table 4. Analysis of the effectiveness of each module.

Method	**Module**				**EgoTaskQA**	**MSRVTT-QA**
	SS	VR	Cross-ATT	LA-Gate		
(a) Baseline	✗	✗	-	-	29.10	23.92
(b) Uniform Sampling	✗	✓	✗	✓	41.70	40.92
(c) Soft Sampling	✗	✓	✗	✓	38.95	38.16
(d) w/o VR	✓	✗	✗	✓	33.44	26.51
(e) Cross-Attention	✓	✓	✓	✗	40.94	41.17
(f) VideoDistill	✓	✓	✗	✓	**45.02**	**44.20**

TaskQA indirect split, which is motivated by the fact [25] that during task execution, actions, objects, and their changes are often strongly correlated. It leaves the chance for the model to perform well by simply over-fitting these strong correlations (language bias) without thorough task understanding. The indirect references can avoid these correlations. In Appendix E, we show that our VideoDistill has the least absolute performance change. It indicates that VideoDistill barely utilizes language bias in questions.

A more effective fusion method than Cross-Attention. In Table 4 (e), we replace LA-Gates in our frameworks with Cross-Attention. When comparing (e) with (f), we can find that LA-Gate brings significant performance boosts on both test sets (4.08% and and 3.03%)

4.5. The Impact of Vision Refinement:

Performance boost in multi-scale object-related questions. The stacked video distillation blocks containing LA-Gate can obtain multi-scale question-related semantics. We observe huge improvements in Table 2 object category (9.33%) and Table 1 object-relation category (4.64%).

4.6. Ablation Study

We conduct ablation studies to verify the effectiveness of components and the rationality of parameter selection.

Effectiveness of each component. We design four variants and report their performances in Table 4. (a) we remove the two modules and apply evenly sampled 16 frames as input and a 3-layer transformer to fuse the vision and language. (b) we remove differentiable sparse sampling and apply evenly sampled 16 frames as input. (c) we replace the hard sampling in SS with soft sampling, which means using the weighted sum of dense frames rather than top-k selection. (d) we remove vision refinement and apply a 3-layer transformer to fuse the vision and language. (e) we replace LA-Gates in both modules with cross-attention layers.

More Ablations We validate the influence of pretraining tasks, the best number of densely sampled frames, and the reasonable number of stacked layers L in Appendix E.

4.7. Conclusion

In this paper, we study language-aware VideoQA to overcome the difficulties of long-term dependencies, multi-events, multi-scale semantics, and shortcut solutions in video understanding. In particular, we introduce a differentiable sparse sampling module that naturally avoids complicated long-term dependencies and multi-event reasoning since it only retains question-related frames and a vision refinement module that eliminates unrelated visual semantics in different perception fields and enhances multi-scale related semantics to support a multi-level refinement process. Both modules are based on our LA-Gate, which realizes cross-modal interaction meanwhile avoiding language's direct fusion into visual representations. LA-Gate is a powerful competitor of predominant Cross-Attention in VideoQA and fits our idea of semantic distillation. Therefore, we can realize high-performance VideoQA by just relying on question-related visual representations and reducing language shortcuts. Experiments on 8 VideoQA and text-to-video retrieval tasks demonstrate the effectiveness of our model.

Acknowledge

This work is supported in part by the Beijing Natural Science Foundation (No. L222024), and the National Natural Science Foundation of China (No. 62394322). One of the authors, Chao Yang, is supported by the Shanghai Post-doctoral Excellent Program (Grant No. 2022234).

References

[1] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356*, 2016. 2

[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980, 2018. 2, 3

[3] Anonymous. LLaMA-excitor: General instruction tuning via indirect feature interaction. In *Conference on Computer Vision and Pattern Recognition 2024*, 2024. 2

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 1, 2, 5, 6

[5] Dare A Baldwin, Jodie A Baird, Megan M Saylor, and M Angela Clark. Infants parse dynamic action. *Child development*, 72(3):708–717, 2001. 1

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 5

[7] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2917–2927, 2022. 7

[8] Wei-Lun Chao, Hexiang Hu, and Fei Sha. Cross-dataset adaptation for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5716–5725, 2018. 2

[9] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10638–10647, 2020. 2

[10] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 3

[11] Feng Cheng and Gedas Bertasius. Tallformer: Temporal action localization with a long-memory transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 503–521. Springer, 2022. 2, 3

[12] Corentin Dancette, Rémi Cadène, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1574–1583, 2021. 2

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 6, 7

[14] Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. Gated-attention readers for text comprehension. *arXiv preprint arXiv:1606.01549*, 2016. 2

[15] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007, 2019. 1, 7

[16] Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. *arXiv preprint arXiv:2212.09522*, 2022. 2, 3, 6

[17] György Gergely, Harold Bekkering, and Ildikó Király. Rational imitation in preverbal infants. *Nature*, 415(6873):755–755, 2002. 1

[18] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5006–5015, 2022. 3

[19] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 2

[20] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 3

[21] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297, 2021. 6

[22] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 2

[23] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9992–10002, 2020. 2

[24] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 5

[25] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *arXiv preprint arXiv:2210.03929*, 2022. 2, 6, 8

[26] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11109–11116, 2020. 1, 7

[27] Hyounghun Kim, Zineng Tang, and Mohit Bansal. Dense-caption matching and frame-selection gating for temporal

localization in videoqa. *arXiv preprint arXiv:2005.06409*, 2020. 3

[28] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 3

[29] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 1

[30] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9972–9981, 2020. 3, 7

[31] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 1, 2, 3, 7

[32] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4953–4963, 2022. 2

[33] Kun Li, Dan Guo, and Meng Wang. Proposal-free video grounding with contextual pyramid network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1902–1910, 2021. 5

[34] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 2

[35] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 1, 7

[36] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8658–8665, 2019. 1, 3, 7

[37] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 510–519, 2019. 2

[38] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *arXiv preprint arXiv:2206.01670*, 2022. 6

[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[40] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 1

[41] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710, 2021. 2

[42] Min Peng, Chongyang Wang, Yu Shi, and Xiang-Dong Zhou. Efficient end-to-end video question answering with pyramidal multimodal transformer. *arXiv preprint arXiv:2302.02136*, 2023. 2, 3, 5, 7

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4

[44] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021. 5

[45] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16877–16887, 2021. 2

[46] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 1

[47] Mykhailo Shvets, Wei Liu, and Alexander C Berg. Leveraging long-range temporal relationships between proposals for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9756–9764, 2019. 2

[48] Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. Long-form video-language pre-training with multimodal temporal contrastive learning. *arXiv preprint arXiv:2210.06031*, 2022. 2

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[50] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*, 2022. 1, 2, 5, 7

[51] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel

attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542, 2020. 2

[52] Su-Jing Wang, Ying He, Jingting Li, and Xiaolan Fu. Mesnet: A convolutional neural network for spotting multi-scale micro-expression intervals in long videos. *IEEE Transactions on Image Processing*, 30:3956–3969, 2021. 2

[53] Amanda L Woodward. Infants selectively encode the goal object of an actor's reach. *Cognition*, 69(1):1–34, 1998. 1

[54] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1884–1894, 2021. 2

[55] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019. 2

[56] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022. 2

[57] Xing Wu, Chaochen Gao, Zijia Lin, Zhongyuan Wang, Jizhong Han, and Songlin Hu. Rap: Redundancy-aware video-language pre-training for text-video retrieval. *arXiv preprint arXiv:2210.06881*, 2022. 3

[58] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 6

[59] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045, 2022. 2, 6

[60] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021. 2

[61] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 591–600, 2020. 5

[62] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018. 6

[63] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[64] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021. 2

[65] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5014–5022, 2016. 2

[66] Xinsong Zhang, Yan Zeng, Jipeng Zhang, and Hang Li. Toward building general foundation models for language, vision, and vision-language understanding tasks. *arXiv preprint arXiv:2301.05065*, 2023. 2

[67] Bo Zou, Chao Yang, Chengbin Quan, and Youjian Zhao. Spaceclip: A vision-language pretraining framework with spatial reconstruction on text. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 519–528, 2023. 3