Advancing High-Resolution Video-Language Representation with Large-Scale Video Transcriptions

Hongwei Xue*, Tiankai Hang*, Yanhong Zeng*, Yuchong Sun*,
Bei Liu, Huan Yang, Jianlong Fu, Baining Guo
Microsoft Research Asia
{v-honxue,v-tiahang,t-yazen,v-yuchongsun,bei.liu,huayan,jianf,bainguo}@microsoft.com

Abstract

*We study joint video and language (VL) pre-training to enable cross-modality learning and benefit plentiful downstream VL tasks. Existing works either extract low-quality video features or learn limited text embedding, while neglecting that high-resolution videos and diversified semantics can significantly improve cross-modality learning. In this paper, we propose a novel **H**igh-resolution and **D**iversified **VI**deo-**LA**nguage pre-training model (HD-VILA) for many visual tasks. In particular, we collect a large dataset with two distinct properties: 1) the first high-resolution dataset including 371.5k hours of 720p videos, and 2) the most diversified dataset covering 15 popular YouTube categories. To enable VL pre-training, we jointly optimize the HD-VILA model by a hybrid Transformer that learns rich spatiotemporal features, and a multimodal Transformer that enforces interactions of the learned video features with diversified texts. Our pre-training model achieves new state-of-the-art results in **10** VL understanding tasks and **2** more novel text-to-visual generation tasks. For example, we outperform SOTA models with relative increases of 40.4% R@1 in zero-shot MSR-VTT text-to-video retrieval task, and 55.4% in high-resolution dataset LSMDC. The learned VL embedding is also effective in generating visually pleasing and semantically relevant results in text-to-visual editing and super-resolution tasks.*

1. Introduction

Recent years we have witnessed an increasing number of videos with the popularity of appealing video websites and mobile apps (*e.g.*, YouTube, TikTok). As the rapid development of smartphone cameras, device storage, and 5G networks, high-quality video creation, and diverse content sharing like travel, sports, and music become a new fashion. Therefore, the capability of video analytic and joint high-level understanding with language play a key role in many video tasks, such as video search [3,39], video recommendation [6], and video editing [38,48]. To facilitate video understanding, we study joint video and language (VL) pre-training, which is a new paradigm in both natural language processing [8] and computer vision [19,52].

Existing video-language understanding models are highly limited in either scale or scope of video-language datasets. Early datasets (*e.g.*, MSR-VTT [53], DiDeMo [2]) consist of videos and descriptions that are manually annotated by humans. The heavy and expensive annotation cost limits the scale of data. Moreover, datasets with only descriptive sentences are limited in complexity and variability that largely hinders generalization power. Recently, several datasets [3, 37] are proposed by transcriptions along with videos using ASR (automatic speech recognition), so that the data scale can be greatly enlarged. One most representative work is HowTo100M [37] which consists of million-scale instructional videos. However, there are still large gaps between these video datasets and real-scenario videos in terms of video quality and semantic diversity.

To tackle the above limitations, we propose the HD-VILA-100M dataset (*i.e.*, **H**igh-resolution and **D**iversified **VI**deo and **LA**nguage) which covers a wide range of video categories and benefits a plenty of VL tasks, such as text-to-video retrieval [39] and video QA [27]. This dataset has the following key properties: (1) Large: we have collected one of the largest video-language datasets, which consists of 100M video clip and sentence pairs from 3.3 million videos with 371.5K hours in total (2.8× video hour and 8× average sentence length than HowTo100M [37]). (2) High resolution: all the videos are 720p which is much higher quality than existing datasets that are mostly 240p or 360p. (3) Diverse and balanced: we cover a wide range of topics from the YouTube, with 15 popular categories (*e.g.*, sports, music, autos). Meanwhile, we ensure a balanced video clip

*Equal contribution in alphabetical order. This work was performed when Hongwei Xue, Tiankai Hang, Yanhong Zeng and Yuchong Sun were visiting Microsoft Research Asia as research interns. Corresponding authors: Bei Liu, Huan Yang, Jianlong Fu.

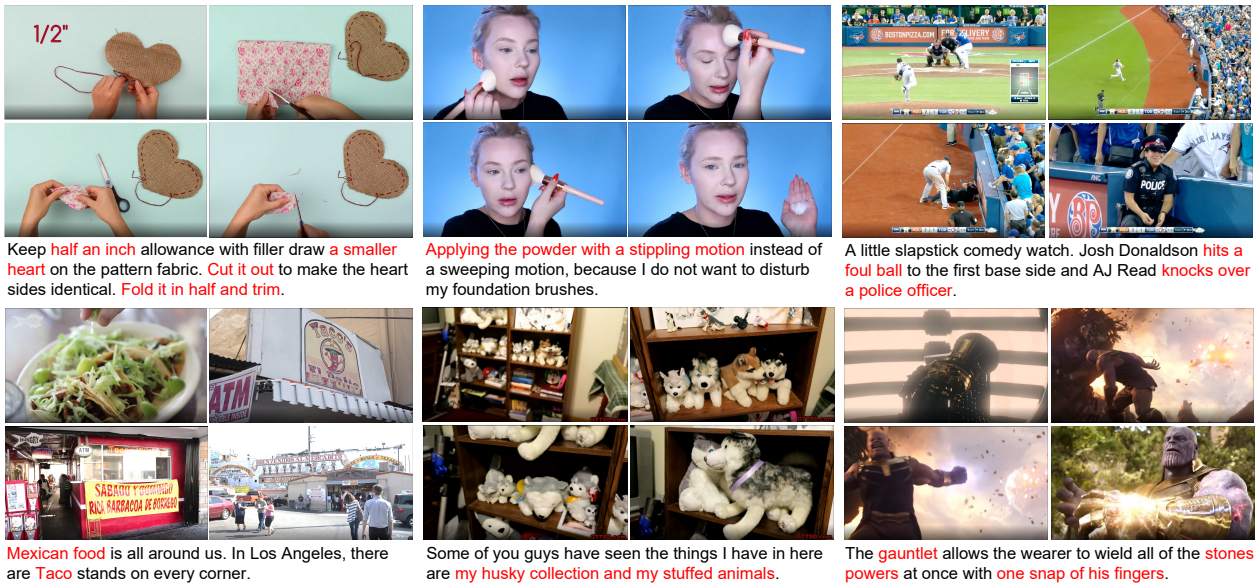

Figure 1. Examples of video clips and ASR generated transcriptions in the proposed HD-VILA-100M dataset. We present six samples (four frames for each), with diverse video categories covering HowTo & Style, People & Blog, Sports, Travel & Event, Pets & Animals, Film & Animation. Relevant words from auto-generated video transcriptions are manually highlighted in red. [Best viewed in Color]

number in each category to ease underfit problem.

To enable video-language pre-training, effective video representation is essential. Due to computational limitations (*e.g.*, memory), previous works either 1) adopt simple frame-based encoders and turn to end-to-end visual encoding and multimodal fusion [27], or 2) choose advanced spatiotemporal encoders [5, 49], while having to do visual encoding and multimodal fusion step-by-step. Few works can learn joint spatiotemporal video representation with end-to-end video-language pre-training.

In this paper, we propose to utilize **hybrid image sequence** that consists of few high-resolution (HR) frames and more low-resolution (LR) neighbor frames for multiple video learning tasks. Such a design enables end-to-end training with high-resolution spatiotemporal video representation. To achieve this goal, we tackle two questions: (1) Which HR and LR frames should be sampled? (2) How to learn spatiotemporal features with the hybrid image sequences? For the first problem, we randomly sample HR frames from a video clip to ensure the robustness of learned video features. LR frames are uniformly sampled from its surroundings considering that neighboring frames contain similar spatial information and are critical to temporal feature learning. Second, we propose to encode HR and LR frames separately while mapping HR feature to a joint embedding space with LR features by a hybrid Transformer. Such design ensures the spatiotemporal representation of videos to cover both HR and LR frames in a learnable way. The learned spatiotemporal feature is further combined with detailed spatial features, followed by a multimodal Transformer that learns to optimize video and language embedding in an end-to-end manner.

Our contributions are summarized as follows: 1) We use automatic video transcriptions to build to-date the largest high-resolution and diversified video-language dataset; 2) We propose a novel pre-training framework to learn spatiotemporal information for video representation from hybrid image sequences that consist of HR and LR frames; 3) Extensive experiments verify the effectiveness of the learned cross-modality embedding in **10** video understanding and **2** text-to-visual generation tasks. The dataset, model and code are released [1].

2. Related Work

Video Representation Video representation are typically designed with 2D/3D CNNs [5, 46, 49] or Transformers [4]. Pioneering works of VL pre-training [39, 44, 63] adopt pre-extracted video features (*e.g.*, S3D [60], I3D [5]) for video representation. While in image-language pre-training, researchers find that end-to-end training will decrease the domain gap of visual representation and improve the generalization for image-text tasks [19]. While for video representation, it is too heavy to make the video-based encoder (*e.g.*, S3D, I3D, ResNet [17], SlowFast [11]) trainable. Thus, some works [27, 57] utilize the image-based encoder (*e.g.*, ResNet [17], ViT [9]) with a sparse sampling mechanism to make the visual encoder trainable. In this paper, we explore how to make a video encoder trainable in consideration of both spatial and temporal features.

Video-Language Pre-Training Vision and language pre-training has attracted extensive attention in very recent

[1] https://github.com/microsoft/XPretrain/tree/main/hd-vila

Dataset	Domain	#Video clips	#Sentence	Avg len(sec)	Sent len	Duration(h)	Resolution
MSR-VTT [53]	open	10K	200K	15.0	9.3	40	240p
DideMo [2]	Flickr	27K	41K	6.9	8.0	87	-
LSMDC [41]	movie	118K	118K	4.8	7.0	158	1080p
YouCook II [62]	cooking	14K	14K	19.6	8.8	176	-
How2 [43]	instructional	80K	80K	90.0	20.0	2K	-
ActivityNet Caption [25]	action	100K	100K	36.0	13.5	849	-
WebVid-2M [3]	open	2.5M	2.5M	18.0	12.0	13K	360p
HowTo100M [37]	instructional	136M	136M	3.6	4.0	134.5K	240p
HD-VILA-100M (Ours)	open	103M	103M	13.4	32.5	371.5K	720p

Table 1. Statistics of HD-VILA-100M and its comparison with existing video-language datasets.

years. Aligned with the success of image-language pre-training [19, 20, 54] and applications [7, 12–14, 18, 30], video-language pre-training is showing more and more promising potentials [27, 29, 39, 44, 45, 63]. Among them, some works concentrate on specific type of downstream tasks such as video-text retrieval [3, 52] and video question answering [57]. In this paper, we explore to pre-train a generalized model on diversified and large-scale data to adapt to different video-language tasks. Video-language pre-training tasks can be mainly categorized into two types: reconstructive, contrastive. Reconstructive methods [29, 44, 45, 63] usually adopt an early fusion architecture and aim to reconstruct a masked part in the visual or textual domain. Typical pre-training tasks are masked language modeling (MLM), masked frame modeling (MFM), frame order modeling (FOM). Contrastive methods [35, 57] are inspired by contrastive learning and target to learn video-text matching. In this paper, we combine these two types of objectives for the final target.

3. Dataset

To facilitate the multimodal representation learning, we collect HD-VILA-100M, a large-scale, high-resolution, and diversified video-language dataset.

3.1. Video Collection

We choose YouTube as the video resource since it covers diverse categories of videos uploaded by different users, ranging from documentary films by professional TV channels to everyday vlogs by ordinary users. To cover more topics, we start from several official topics of YouTube videos. To ensure the high quality of videos as well as better alignment of video and transcription, we search on the YouTube website and a video analysis website [2] to find popular YouTube channels, such as BBC Earth, National Geography, etc. Videos in these channels and videos appeared in YouTube-8M [1] and YT-Temporal-180M [57] make up a list of 14 million videos. We only keep videos with subtitles and 720p resolution. We then limit the time length of each category to 30K hours to avoid long tail. We only download videos with English transcripts. Finally, we obtain 3.3 million videos in total with high-quality and distributed in 15 categories in balance (as in Figure 2).

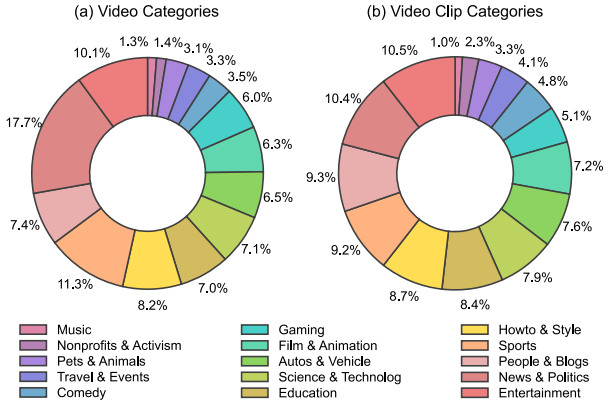

Figure 2. The distribution of categories in HD-VILA-100M dataset: (a) video, (b) video clip. [Best viewed in Color]

3.2. Video Clip Selection and Text Processing

To effectively generate video-text pairs, we use transcriptions along with the videos as the language in HD-VILA-100M. Different from traditional video-language datasets [2, 53] that use manual annotated descriptions for videos, transcriptions are available in large quantities and involve richer information. However, many subtitles in YouTube videos are generated by ASR and are usually fragmentary and lacking punctuation. To split the subtitles for complete sentences, we utilize an off-the-shelf tool [3] which shows 75.7% accuracy on its test set. Then we make video clips by aligning the sentences to corresponding clips via Dynamic Time Warping using the timestamp of the original subtitles. After processing, each pair in the HD-VILA-100M consists of a video clip about 13.4 seconds on average and a sentence with 32.5 words on average.

3.3. Data Statistics

The detailed data statistics of HD-VILA-100M are listed in Table 1. Compared with other video-language datasets, HD-VILA-100M is the largest video-language dataset in terms of duration and word number. More videos indicate richer visual information contained in HD-VILA-100M and longer sentences mean that the language includes more de-

[2] https://socialblade.com/youtube/

[3] https://github.com/ottokart/punctuator2

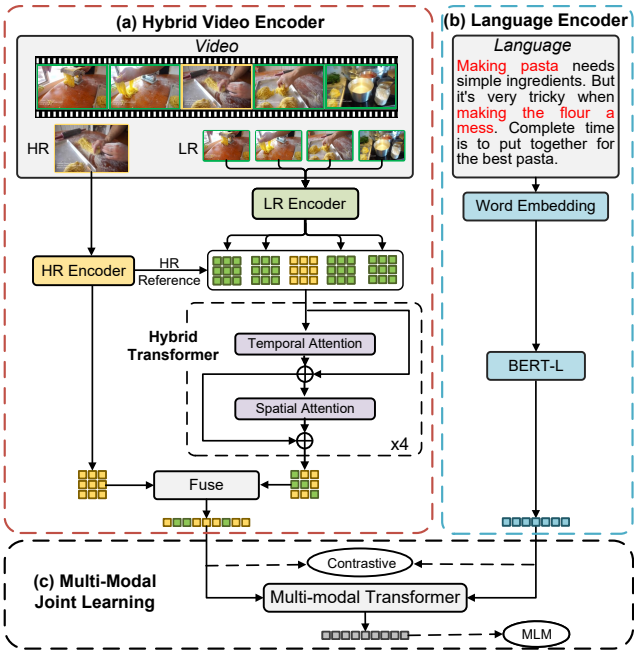

Figure 3. The framework of HD-VILA. Yellow and green colors indicate HR- and LR-related input, operation and output, respectively. Hybrid Transformer learns spatiotemporal representation from HR and LR features. [Best viewed in Color]

tailed and richer semantics. Compared with HowTo100M [37] which only includes instructional videos, HD-VILA-100M is derived from a wide range of domains and videos of each category is relatively balanced as shown in Figure 2. This merit can improve the generalization power of the pre-trained model. Moreover, all the videos in HD-VILA-100M are in 720p and the high quality ensures detailed information for video representation learning. In summary, HD-VILA-100M represents the largest, high-resolution, and diversified dataset for video and language learning.

4. Approach

Figure 3 shows the overall framework of **H**igh-resolution and **D**iversified **VI**deo-**LA**nguage (HD-VILA) model that consists of three parts: (a) hybrid video encoder, (b) language encoder, and (c) multi-modal joint learning.

4.1. Hybrid Video Encoder

Since the video clips in our dataset are long-range with 13.4 seconds on average, we adopt the strategy of sparsely sampling a sequence of segments from a video clip and then aggregating their predictions similar to ClipBERT [27]. As explained in Section 1, for each segment s_i, we randomly takes one HR frame at t-th timestep $\boldsymbol{X}_t^{s_i} \in \mathbb{R}^{3\times H\times W}$ and $2N$ surrounding LR frames $\{\boldsymbol{X}_{t+kr}^{s_i} \in \mathbb{R}^{3\times \frac{H}{4}\times \frac{W}{4}}, k \in (-N,\dots,-1,1,\dots,N)\}$ to build a **hybrid image sequence**, where r is LR frame sampling rate.

In Figure 3, the hybrid video encoder includes three parts: an HR encoder for HR frame, an LR encoder for LR neighbor frames and a **Hybrid Transformer** T_{hy} that learns spatiotemporal features by self-attention. HR encoder consists of a 4-stage ResNet F_{hr} and an adapter D_{hr}. LR encoder is a 3-stage ResNet F_{lr} to encode LR frames. Note that F_{hr} and F_{lr} are learnable to ensure both HR and LR frames can be encoded in the same space before feeding into Hybrid Transformer. We extract hybrid spatiotemporal feature $\mathcal{V}_{\text{hy}}$ of segment s_i as the output of T_{hy}. In addition, we use the HR frame feature extracted by stage 3 of F_{hr} (denoting as F_{hr}^3) as HR input of T_{hy}:

$$\mathcal{V}_{\text{hy}} = T_{\text{hy}}(F_{\text{lr}}(\boldsymbol{X}_{t-Nr}^{s_i}), ..., \phi(F_{\text{hr}}^3(\boldsymbol{X}_t^{s_i})), ...), \quad (1)$$

where ϕ is an interpolate operation to align feature size. In T_{hy}, we adopt Divided Space-Time Attention to encode spatiotemporal information similar to [4]. We extract detailed spatial feature $\mathcal{V}_{\text{hr}}$ of segment s_i as the output of E_{hr} by:

$$\mathcal{V}_{\text{hr}} = D_{\text{hr}}(F_{\text{hr}}(\boldsymbol{X}_t^{s_i})), \quad (2)$$

To adapt the output of the HR encoder to the hybrid spatiotemporal feature $\mathcal{V}_{\text{hy}}$, D_{hr} consists of a convolution layer to adjust the output feature channel, as well as a 2×2 max-pooling layer for down-sampling. The segment features is the fusion of $\mathcal{V}_{\text{hr}}$ and $\mathcal{V}_{\text{hy}}$ by a linear layer:

$$\mathcal{V} = \boldsymbol{Linear}([\mathcal{V}_{\text{hr}}, \mathcal{V}_{\text{hy}}]). \quad (3)$$

4.2. Language Encoder and Multi-Modality Joint Embedding Learning

For both language encoder and multi-modality joint embedding learning, we use self-attention to model the relationship of both uni-modality and multi-modality. More specifically, we adopt a 24-layer, 1024-dimensional Transformer, mirroring the BERT-large and initialize it with pre-trained BERT-large parameters. We use the first 12 layers as language-only Transformer and the last 12 layers as multi-modal Transformer. Language-only Transformer extracts language representation which is concatenated with video features of a segment as the input of multi-modal Transformer. We add learnable 1D and 2D position embedding to language and vision tokens, respectively. Such a modal-independent design has two advantages. Firstly, it enables to provide powerful embedding for a single-modal input in downstream tasks. For example, the vision-aware language-only embedding could be used for language-guided video generation tasks. Secondly, the two-stream architecture improves the calculation efficiency of similarity between video and language to linear complexity in some specific downstream tasks, such as video-language retrieval.

4.3. Pre-Training Tasks

We adopt two pre-training tasks in HD-VILA: video-language matching to enhance cross-modal matching and masked language modeling (MLM) to encourage the mapping between visual and language tokens in fine-grained

Method	Acc
ST-VQA [21]	30.9
Co-Memory [16]	32.0
AMU [50]	32.5
Heterogeneous Mem [10]	33.0
HCRN [26]	35.6
ClipBERT PT [27]	37.4
Ours	**40.0**

(a) MRSVTT-QA test set.

Method	Acc
CT-SAN [56]	66.4
MLB [24]	76.1
JSFusion [55]	83.4
ActBERT PT [63]	85.7
ClipBERT PT [27]	88.2
VideoClip PT [52]	92.1
Ours	**97.1**

(b) MRSVTT multiple-choice test.

Method	Action	Trans	Frame
ST-VQA [21]	60.8	67.1	49.3
Co-Memory [16]	68.2	74.3	51.5
PSAC [31]	70.4	76.9	55.7
HCRN [26]	75.0	81.4	55.9
QueST [22]	75.9	81.0	59.7
ClipBERT PT [27]	82.8	87.8	60.3
Ours	**84.3**	**90.0**	**60.5**

(c) TGIF-QA test set.

Table 2. Comparison of HD-VILA with state-of-the-art methods on video question answering tasks. (a) Results of ST-VQA and Co-Memory are implemented by [10]. (b) Results of CT-SAN and MLB are implemented by [55].

level. In particular, since the matching between video and language is somewhat weak compared with the video description dataset, we apply contrastive video-language matching to take advantage of large data.

Contrastive Video-Language Matching To align the feature space of video and language, we use a contrastive loss to maximize the similarity of a video clip and a sentence. Specifically, we treat matched pairs in a batch as positives, and all other pairwise combinations as negatives:

$$\begin{aligned}\mathcal{L}_{v2t} &= -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(v_i^\top t_i/\tau\right)}{\sum_{j=1}^{B}\exp\left(v_i^\top t_j/\tau\right)}\\ \mathcal{L}_{t2v} &= -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(t_i^\top v_i/\tau\right)}{\sum_{j=1}^{B}\exp\left(t_i^\top v_j/\tau\right)},\end{aligned} \tag{4}$$

where v_i and t_j are the normalized embeddings of i-th video and j-th sentence in a batch of size B and τ is the temperature. Video and sentence features are computed by our hybrid video encoder and language encoder. The mean of segment embeddings is used as the video-level embedding.

Masked Language Modeling We adopt Masked Language Modeling (MLM) to better build the mapping between visual and language domain. MLM aims to predict the ground-truth labels of masked text tokens from the contextualized tokens:

$$\mathcal{L}_{\text{MLM}} = -\mathbb{E}_{(\mathcal{W},\mathcal{V})}\log p\left(w_i \mid \mathcal{W}_{\setminus i}, \mathcal{V}\right), \tag{5}$$

where $\mathcal{W}$ denotes the text embedding token set, $\mathcal{V}$ denotes the visual token set, and w_i denotes the masked token. $(\mathcal{W}, \mathcal{V})$ is sampled from the distribution of text-video pairs. We adopt the same masking strategy as in BERT and use an MLP as the MLM head to output logits over vocabulary, which is then computed as the negative log-likelihood loss for the masked token. We aggregate the logits of different segments to derive a consensus, so that MLM is able to be calculated in video-level as we adopt in the approach.

5. Experiments

In this section, we conduct extensive experiments to evaluate the proposed HD-VILA pre-training model.

5.1. Pre-training Details

Inspired by the idea of "align before fuse" [28], we adopt a two-stage fashion for pre-training on HD-VILA-100M dataset. In the first stage, we perform the contrastive video-language matching task to learn cross-modality alignment. In the second stage, MLM task is performed to facilitate cross-modal understanding. For video encoder, we use ResNet-50 for F_{hr} and F_{lr}, and a 4-layer Transformer with 16 heads and 1024 hidden size for T_{hy}. We empirically divide a video clip into two segments and sample seven frames for each segment. In this setting, the two segments can cover about 6s video content, which are adequate to model the video clips in our dataset. Besides, we randomly crop 640×1024 areas for the middle HR frames, and select aligned 160×256 areas for LR neighboring frames. The size of resultant feature map before feeding into the multimodal Transformer is 10×16. For language, we follow BERT [8] to adopt the WordPiece tokenizer to split a sentence into word tokens with a max length of 50.

In pre-training, we use AdamW optimizer [34] with an initial learning rate of 5e-5 and a fixed weight decay of 1e-3. We also employ a linear decay learning rate schedule with a warm-up strategy. We train our model with 128 NVIDIA Tesla V100 GPUs for stage one and 32 for stage two. The batch size is set to 1024 and the contrastive similarity is calculated on gathered features from all GPUs. We train 7 epochs for stage one and 4 epochs for stage two empirically. We freeze the encoders during the second stage and keep the same batch size for both stages. In downstream tasks, we keep the same model configuration if not otherwise specified. We exclude the YouTube Ids in the downstream tasks from our collected HD-VILA-100M during pre-training.

5.2. Video Question and Answering

Datasets **(a) MSRVTT-QA** [50] is created based on video and captions in MSR-VTT [53]. Given a question in a complete sentence, the model selects an answer from a pre-defined set. **(b) MSRVTT multiple-choice test** [55] is a multiple-choice task with videos as queries, and captions as answers. Each video contains five candidate captions, with only one positive match. **(c) TGIF-QA** [21] is build on GIF videos. We experiment with three TGIF-QA tasks:

Method	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓
HowTo100M [37]	14.9	40.2	52.8	9.0
CE [32]	20.9	48.8	62.4	6.0
DECEMBERT [45]	17.5	44.3	58.6	9.0
HERO [29]	16.8	43.4	57.7	-
ClipBERT [27]	22.0	46.8	59.9	6.0
VLM [51]	28.1	55.5	67.4	4.0
MMT [15]	26.6	57.1	69.6	4.0
Support Set [39]	30.1	58.5	69.3	3.0
VideoCLIP [52]	32.2	62.6	75.0	-
Ours	**35.6**	**65.3**	**78.0**	**3.0**
Zero-shot				
HT MIL-NCE [35]	9.9	24.0	32.4	29.5
Support Set [39]	8.7	23.0	31.1	31.0
VideoCLIP [52]	10.4	22.2	30.0	-
Ours	**14.6**	**34.4**	**44.1**	**15.0**

Table 3. Comparison of text-to-video retrieval in MSR-VTT [53]. We gray out some lines to highlight fair comparisons with traditional retrieval models and general pre-training models. This mark is also applicable to Table 5, 6.

Action is defined as a multiple-choice task to identify an action that has been repeated in a video. *Transition* aims to identify the state before or after another state. *FrameQA* is about open-ended questions about the given video. The task objective is identical to MSRVTT-QA. More details are in the supplementary materials.

Implementation Details For TGIF Action and Transition, we respectively concatenate five candidate answers with the question into five sequences. On top of the [CLS] token of the question, we train a two-layer MLP to predict the confidence of the five candidates with cross-entropy loss. For MSRVTT-QA and TGIF Frame, we encode the answers in a one-hot fashion, and train 2-layer MLP classifier over all answer candidates with a cross-entropy loss on-top of the [CLS] token of the question. For MSRVTT Multiple-choice, we directly choose the answer with the highest similarity. We set the max batch size to fine-tune on 8 V100 32G GPUs. More details are in the supplementary materials.

Results In Table 2, the results of HD-VILA on video QA show that our model outperforms existing methods on five tasks in all the three datasets. On MSRVTT-QA and MSRVTT multiple-choice tests, we achieve **2.6** and **5.0** absolute improvement over SOTA methods. On TGIF-QA dataset, we have **1.5**, **2.2** and **0.2** absolute improvements on *Action*, *Trans* and *Frame* tasks. The limited gain of *Frame* is due to that *Frame* focuses on single frame while hindering the advantage of our hybrid image sequence. Among all the compared methods, ClipBERT [27] and ActBERT [63] are pre-training models. We can see that pre-training with more data will marginally improve the performance. Compared with ClipBERT which is pre-trained on image-language dataset, videos provide richer information. Note that the language used in ClipBERT pre-training is more closer to downstream tasks in both content and length while the language in HD-VILA-100M has domain gap with TGIF and MSR-VTT languages. This further indicates the generalization of the video representation learned by our HD-VILA.

Method	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓
HERO [29]	2.1	-	11.4	-
S2VT [47]	11.9	33.6	-	13.0
FSE [59]	13.9	36.0	-	11.0
CE [32]	16.1	41.1	-	8.3
ClipBERT [27]	20.4	48.0	60.8	6.0
Ours	**28.8**	**57.4**	**69.1**	**4.0**

Table 4. Comparison of text-to-video retrieval on DiDeMo [2].

Method	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓
JSFusion [55]	9.1	21.2	34.1	36.0
MEE [36]	9.3	25.1	33.4	27.0
CE [32]	11.2	26.9	34.8	25.3
MMT [15]	12.9	29.9	40.1	19.3
Ours	**17.4**	**34.1**	**44.1**	**15.0**

Table 5. Comparison of text-to-video retrieval on LSMDC [41].

Method	R@1 ↑	R@5 ↑	R@50 ↑	MedR ↓
FSE [59]	18.2	44.8	89.1	7.0
CE [32]	18.2	47.7	91.4	6.0
HSE [59]	20.5	49.3	-	-
ClipBERT [27]	21.3	49.0	-	6.0
MMT [15]	28.7	61.4	94.5	3.3
Support Set [39]	29.2	61.6	94.7	3.0
Ours	**28.5**	**57.4**	**94.0**	**4.0**

Table 6. Comparison of text-to-video retrieval on ActivityNet [25].

5.3. Video-Text Retrieval

Datasets We conduct video-text retrieval experiments on four datasets. **(a) MSR-VTT** [53] contains 10K YouTube videos with 200K descriptions. We follow previous works [32, 55], training models on 9K videos, and reporting results on the 1K-A test set. **(b) DiDeMo** [2] consists of 10K Flickr videos annotated with 40K sentences. We follow [32, 59] to evaluate paragraph-to-video retrieval, where all descriptions for a video are concatenated to form a single query. **(c) LSMDC** [41] consists of 118,081 video clips sourced from 202 movies. Each video has a caption. Evaluation is conducted on a test set of 1,000 videos from movies disjoint from the train and validation sets. **(d) ActivityNet Captions** [25] contains 20K YouTube videos annotated with 100K sentences. We follow the paragraph-to-video retrieval protocols [32,59] training on 10K videos and reporting results on the val1 set with 4.9K videos.

Implementation Details We adjust the number of sampled segments and frames according to the average time of videos for each dataset. We adopt stage one model and the same training methods and objective for fine-tuning. We resize HR frame of each segment to 720p and LR frames to 180p. More details are in the supplementary materials.

Results Table 3, 4, 5, 6 show the text-to-video retrieval results of HD-VILA on four datasets. For MSR-VTT, we

Figure 4. Text-guided manipulation compared with StyleCLIP [38] and TediGAN [48]. Our model is able to handle complex descriptions and edit the inputs according to the target attributes (highlighted in red) better. All the inputs are of 1024×1024 size.

Figure 5. Text-guided super-resolution compared with pSp [40] and SR3 [42]. Our model is able to reconstruct more accurate target attributes with descriptions (*e.g.*, eyeglasses in the third case). All inputs are upsampled from 16×16 to 1024×1024.

outperform the previous works by large margins in both zero-shot and fine-tuning settings. In particular, compared with VideoCLIP [52], we have **40.4%** relatively gains of R@1 in zero-shot setting, which shows the generalization ability of our pre-trained feature. In LSMDC, we further obtain much larger relative gains with **55.4%** under fair comparison. This comes from smaller domain gap between movie videos in LSMDC and our HD-VILA-100M compared with HowTo100M in two aspects: semantic (both open domains) and resolution (both high-resolution). On DiDeMo and ActivityNet, our model also achieves better performance. The videos in these two datasets are diversified in both scale and category, and are much longer. The results shows that our model pre-trained on HD-VILA-100M with longer videos and richer semantics shows better capacity for temporal understanding. Note that there are also pre-training models that are specifically designed for video-text retrieval task by improving noise contrastive learning like SupportSet [39], or use more features other than vision and motion like MMT [15]. To make fair comparison, we gray them out in tables.

5.4. Text-to-Visual Generation

Recent studies like StyleCLIP [38] and TediGAN [48] propose to leverage cross-modal pre-training power to facilitate language-guided generation tasks, and have obtained some promising results. As shown in their work, the quality of visual generation results can reflect the quality of cross-modality embedding. Hence, in this section, we will specify how our pre-trained model can achieve this task, and verify our learned embedding by showing higher-quality visualized results compared with SOTA models.

Datasets To conduct this research, we introduce the first **F**ace-**D**escription-**V**ideo **D**ateset (**FDVD**). The dataset consists of 613 high-resolution (1024×1024) videos, resulting in 74,803 frames of human faces. The videos are collected from Ryerson audio-visual dataset [33]. We generate ten different text descriptions for each video following previous works [48]. To increase the diversity of human faces, we also leverage Multi-modal CelebA-HQ [48] for training.

Implementation Details We follow previous works [38, 48] to leverage a well pre-trained StyleGAN [23, 58, 61] as our generator, due to its superior performance. In practice, we learn several linear layers to map the vision and text embedding in HD-VILA to the latent codes w^+ used in StyleGAN. Then, images can be generated by the latent codes. To ensure the visual quality, identity preservation, and matching with descriptions of the generated results, we carefully choose a set of losses for optimization. More details are in the supplementary materials.

Text-to-Visual Editing We compare our model with the recent state-of-the-art text-guided editing models, StyleCLIP [38] and TediGAN [48] in Figure 4. The results show that our model is able to edit the target attributes of inputs according to text descriptions. For example, in the first case in Figure 4, our model turns the hair to wavy hair and also wears lipstick on the lips, where StyleCLIP and TediGAN fail to wear lipstick on the face. Some video cases will be presented in supplementary materials.

Text-to-Visual Super-Resolution We further compare our model with SOTA super-resolution methods SR3 [42] and pSp [40]. We generate 1024×1024 images from their 16×16 LR counterparts. Note that this task is extremely challenging due to such low-resolution inputs. As shown in the second case of Figure 5, SR3 [42] and pSp [40] can not reconstruct high-quality faces by only using visual information. Compared with them, our model is able to accurately reconstruct the lipstick and the straight hair with the help of text description, thanks to the pre-trained models.

5.5. Ablation Studies

In this section, we conduct ablation studies to further verify the effectiveness of the new HD-VILA-100M dataset, and the proposed hybrid video encoder.

(1) Diversity of HD-VILA-100M. We sample two video subsets from HD-VILA-100M with two million clip-text pairs for each. One subset only includes "HowTo" type, while the other consists of diversified and balanced categories sampled from the full dataset. As shown in Table 7, compared with the "HowTo" dataset with limited semantics, our diversified pre-training dataset (indicated as "Ours-720p") helps to achieve higher performance in the MSR-VTT retrieval task, with relative **66.7%** R@1 gains. We choose MSR-VTT zero-shot retrieval task for this ablation study, as it is the most widely-used evaluation task in video-language pre-training. We also make fair comparison with HowTo100M [37]. we have tried our best to collect HowTo100M at 720p, in which 69% videos are originally at 720p, and 31% are at 240p (w/o HR source) and upsampled to 720p by applying the most commonly used bicubic interpolation. We select MSR-VTT retrieval which is the most widely-used benchmark for pre-training evaluation. We report the comparison in Table 8. We compare pre-training on two datasets for the same steps (145K) and fine-tuning with the same setting. HD-VILA-100M pre-trained model surpasses HowTo100M by a large margin. This shows the advantage of HD-VILA-100M.

(2) High-resolution of HD-VILA-100M. We downsample "Ours-720p" subset into lower resolutions ("Ours-360p"), and observed a significant drop with **29.1%** relative decreases of R@1. Such evaluations demonstrate the superiority of the diversified categories and higher resolution of the proposed dataset.

Type	Size	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓
HowTo	720p	3.3	8.2	13.5	113.0
Ours	360p	3.9	11.0	18.3	67.0
Ours	720p*	4.5	13.0	20.2	62.0
Ours	720p	**5.5**	**13.1**	**20.5**	**58.0**

Table 7. Ablation study on two subsets of pre-training data. We report results of zero-shot MSR-VTT retrieval. 720p* indicates bi-cubic upsampled frames (360p to 720p).

Dataset	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓
HowTo100M	19.6	49.0	61.9	6.0
Ours	**30.0**	**58.1**	**72.3**	**4.0**

Table 8. Comparison of pre-training datasets on MSR-VTT retrieval with the same steps.

#HR	#LR	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓
1	0	16.3	40.0	53.3	9.0
0	10	26.7	57.0	69.5	4.0
1	6	33.0	64.4	76.2	3.0
1	10	**35.6**	**65.3**	**78.0**	**3.0**
1	14	33.7	64.1	76.2	**3.0**

Table 9. Ablation study on frame selection. We report results of MSR-VTT retrieval, where #HR/#LR are the numbers of high/low-resolution frames.

(3) Numbers of HR/LR frames. As the number of high/low-resolution frames used for video modeling often plays a key role in video pre-training, we adjust frame numbers and fine-tune the pre-training model in different settings. As shown in Table 9, high-resolution frames lead to significant increases compared with the setting only using low-resolution inputs. In particular, the setting of 1-HR & 10-LR achieves the best performance, compared with 0-HR & 10-LR ("0" indicates that one branch is removed), and 1-HR & 0-LR, which demonstrates the rationality of jointly modeling spatial and temporal features in our approach.

6. Conclusion

In this paper, we propose to learn high-resolution and diversified video-language multi-modal representation by pre-training on large-scale video-language pairs. To empower pre-training, we introduce a new dataset **HD-VILA-100M** which is the largest high-resolution and diversified video-language dataset. To more efficiently employ the richer information in videos, we propose a novel pre-training model HD-VILA that learns spatiotemporal information using HR and LR frames as a **hybrid** image sequence with a hybrid Transformer. Experiments on **12** video-language understanding and text-to-visual generation tasks show the capability of HD-VILA-100M dataset and the effectiveness of our model.

Acknowledgement We would like to thank the insightful discussion, valuable suggestions and kind help from Prof. Jiebo Luo, Prof. Ruihua Song, Prof. Limin Wang, Houwen Peng, and Dongdong Chen.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 3

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812, 2017. 1, 3, 6

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 1, 3

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, July 2021. 2, 4

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 2

[6] Bisheng Chen, Jingdong Wang, Qinghua Huang, and Tao Mei. Personalized video recommendation through tripartite graph propagation. In *ACM MM*, pages 1133–1136, 2012. 1

[7] Shizhe Chen, Bei Liu, Jianlong Fu, Ruihua Song, Qin Jin, Pingping Lin, Xiaoyu Qi, Chunting Wang, and Jin Zhou. Neural storyboard artist: Visualizing stories with coherent image sequences. In *ACM MM*, pages 2236–2244, 2019. 3

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, pages 4171–4186, 2019. 1, 5

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2

[10] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *CVPR*, pages 1999–2007, 2019. 5

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *CVPR*, pages 6202–6211, 2019. 2

[12] Jianlong Fu, Tao Mei, Kuiyuan Yang, Hanqing Lu, and Yong Rui. Tagging personal photos with transfer deep learning. In *WWW*, pages 344–354, 2015. 3

[13] Jianlong Fu and Yong Rui. Advances in deep learning approaches for image tagging. *APSIPA*, 6, 2017. 3

[14] Jianlong Fu, Jinqiao Wang, Yong Rui, Xin-Jing Wang, Tao Mei, and Hanqing Lu. Image tag refinement with view-dependent concept representations. *T-CSVT*, 25(8):1409–1422, 2014. 3

[15] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, pages 214–229, 2020. 6, 7

[16] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *CVPR*, pages 6576–6585, 2018. 5

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[18] Yupan Huang, Hongwei Xue, Bei Liu, and Yutong Lu. Unifying multimodal transformer for bi-directional image and text generation. In *ACM MM*, pages 1138–1147, 2021. 3

[19] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *CVPR*, pages 12976–12985, 2021. 1, 2, 3

[20] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 3

[21] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, pages 2758–2766, 2017. 5

[22] Jianwen Jiang, Ziqiang Chen, Haojie Lin, Xibin Zhao, and Yue Gao. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In *AAAI*, pages 11101–11108, 2020. 5

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 7

[24] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *ICLR*, 2016. 5

[25] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017. 3, 6

[26] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *CVPR*, pages 9972–9981, 2020. 5

[27] 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 *CVPR*, pages 7331–7341, 2021. 1, 2, 3, 4, 5, 6

[28] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 5

[29] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *EMNLP*, pages 2046–2065, 2020. 3, 6

[30] Nanxing Li, Bei Liu, Zhizhong Han, Yu-Shen Liu, and Jianlong Fu. Emotion reinforced visual storytelling. In *ICMR*, pages 297–305, 2019. 3

[31] 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 *AAAI*, pages 8658–8665, 2019. 5

[32] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *BMVC*, 2019. 6

[33] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, page e0196391, 2018. 7

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[35] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, pages 9879–9889, 2020. 3, 6

[36] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018. 6

[37] 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 *ICCV*, pages 2630–2640, 2019. 1, 3, 4, 6, 8

[38] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, pages 2085–2094, 2021. 1, 7, 8

[39] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F. Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *ICLR*, 2021. 1, 2, 3, 6, 7

[40] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, pages 2287–2296, 2021. 7, 8

[41] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Joseph Pal, H. Larochelle, Aaron C. Courville, and Bernt Schiele. Movie description. *IJCV*, pages 94–120, 2016. 3, 6

[42] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *arXiv preprint arXiv:2104.07636*, 2021. 7, 8

[43] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: A large-scale dataset for multimodal language understanding. In *NeurIPS*, 2018. 3

[44] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, pages 7464–7473, 2019. 2, 3

[45] Zineng Tang, Jie Lei, and Mohit Bansal. Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In *NAACL*, pages 2415–2426, 2021. 3, 6

[46] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015. 2

[47] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond J Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. In *HLT-NAACL*, 2015. 6

[48] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *CVPR*, pages 2256–2265, 2021. 1, 7, 8

[49] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, pages 305–321, 2018. 2

[50] 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 *ACM MM*, page 1645–1653, 2017. 5

[51] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. Vlm: Task-agnostic video-language model pre-training for video understanding. *arXiv preprint arXiv:2105.09996*, 2021. 6

[52] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP*, pages 6787–6800, 2021. 1, 3, 5, 6, 7

[53] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016. 1, 3, 5, 6

[54] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. Probing intermodality: Visual parsing with self-attention for vision-and-language pre-training. 34, 2021. 3

[55] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, pages 471–487, 2018. 5, 6

[56] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. End-to-end concept word detection for video captioning, retrieval, and question answering. In *CVPR*, pages 3261–3269, 2017. 5

[57] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. In *NeurIPS*, 2021. 2, 3

[58] Yanhong Zeng, Huan Yang, Hongyang Chao, Jianbo Wang, and Jianlong Fu. Improving visual quality of image synthesis by a token-based generator with transformers. In *NeurIPS*, volume 34, 2021. 7

[59] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *ECCV*, pages 374–390, 2018. 6

[60] Da Zhang, Xiyang Dai, Xin Wang, and Yuan-Fang Wang. S3d: Single shot multi-span detector via fully 3d convolutional network. In *BMVC*, 2018. 2

[61] Heliang Zheng, Jianlong Fu, Yanhong Zeng, Jiebo Luo, and Zheng-Jun Zha. Learning semantic-aware normalization for generative adversarial networks. In *NeurIPS*, 2020. 7

[62] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 3

[63] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, pages 8746–8755, 2020. 2, 3, 5, 6