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 High-resolution and Diversified Video-Language 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., High-resolution and Diversified Video and Language) 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
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 released1.

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

\[1https://github.com/microsoft/XPretrain/tree/main/hd-vila\]
years. Aligned with the success of image-language pre-training \cite{19,20,54} and applications \cite{7,12–14,18,30}, video-language pre-training is showing more and more promising potentials \cite{27,29,44,45,63}. Among them, some works concentrate on specific type of downstream tasks such as video-text retrieval \cite{3,52} and video question answering \cite{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 \cite{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 \cite{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\footnote{https://socialblade.com/youtube/} to find popular YouTube channels, such as BBC Earth, National Geography, etc. Videos in these channels and videos appeared in YouTube-8M \cite{1} and YT-Temporal-180M \cite{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).

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 \cite{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\footnote{https://github.com/ottokart/punctuator2} 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-
4. Approach

Figure 3 shows the overall framework of High-resolution and Diversified ViDeo-LA(n)guage (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 $X_{t}^{s_i} \in \mathbb{R}^{3 \times H \times W}$ and $2N$ surrounding LR frames $\{X_{t+N}^{s_i} \in \mathbb{R}^{3 \times 4 \times \frac{H}{4}, k \in (-N, \ldots, -1, 1, \ldots, N)}\}$ 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 $V_{hy}$ 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 $T_{hy}$:

$$V_{hy} = T_{hy}(F_{hr}(X_{t}^{s_i}), \ldots, \phi(F_{hr}^{3}(X_{t}^{s_i})), \ldots),$$  \hspace{1cm} (1)

where $\phi$ 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 $V_{hr}$ of segment $s_i$ as the output of $E_{hr}$ by:

$$V_{hr} = D_{hr}(F_{hr}(X_{t}^{s_i})), \hspace{1cm} (2)$$

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

$$V = \text{Linear}([V_{hr}, V_{hy}]).$$  \hspace{1cm} (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...
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:

\[
\mathcal{L}_{v2t} = - \frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp (v_i^T t_i / \tau)}{\sum_{j=1}^{B} \exp (v_i^T t_j / \tau)}
\]

\[
\mathcal{L}_{t2v} = - \frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp (t_i^T v_i / \tau)}{\sum_{j=1}^{B} \exp (t_i^T v_j / \tau)},
\]

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 \(\tau\) 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:

\[
\mathcal{L}_{\text{MLM}} = - \mathbb{E}_{(\mathcal{W}, \mathcal{V})} \log p \left( w_i \mid \mathcal{W} \backslash \{i\}, \mathcal{V} \right),
\]

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.
downstream tasks in both content and length while the language used in ClipBERT pre-training is more closer to dataset, videos provide richer information. Note that the advantage of our hybrid image sequence. Among all the five tasks in all the three datasets. On MSRVTT-QA and TGIF-QA show that our model outperforms existing methods on 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 MSRVVT 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.

### 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 202 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...
This man has black bushy eyebrows and bangs.

This person is chubby and has rosy cheeks.

This man has beard and he is smiling.

She has wavy hair and wears red lipstick.

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 \( \times \) 1024 size.

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 Face-Description-Video Dataset (FDVD). The dataset consists of 613 high-resolution (1024 \( \times \) 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 $1024 \times 1024$ images from their $16 \times 16$ 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] cannot 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-tune 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 ↓ |
|-----|-----|-------|-------|--------|--------|
| 0   | 0   | 16.3  | 40.0  | 53.3   | 9.0    |
| 0   | 1   | 26.7  | 57.0  | 69.5   | 4.0    |
| 1   | 0   | 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.

6) 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.
