Machine Intelligence Research
www.mi-research.net

20(2), April 2023, 249-262
DOI: 10.1007/s11633-022-1409-1

Compositional Prompting Video-language Models to Understand Procedure in Instructional Videos

Guyue Hu[1] Bin He[2] Hanwang Zhang[1]

[1] Nanyang Technological University, Singapore 639798, Singapore

[2] The 15th Research Institute of China Electronics Technology Group Corporation, Beijing 100083, China

Abstract: Instructional videos are very useful for completing complex daily tasks, which naturally contain abundant clip-narration pairs. Existing works for procedure understanding are keen on pretraining various video-language models with these pairs and then finetuning downstream classifiers and localizers in predetermined category space. These video-language models are proficient at representing short-term actions, basic objects, and their combinations, but they are still far from understanding long-term procedures. In addition, the predetermined procedure category faces the problem of combination disaster and is inherently inapt to unseen procedures. Therefore, we propose a novel compositional prompt learning (CPL) framework to understand long-term procedures by prompting short-term video-language models and reformulating several classical procedure understanding tasks into general video-text matching problems. Specifically, the proposed CPL consists of one visual prompt and three compositional textual prompts (including the action prompt, object prompt, and procedure prompt), which could compositionally distill knowledge from short-term video-language models to facilitate long-term procedure understanding. Besides, the task reformulation enables our CPL to perform well in all zero-shot, few-shot, and fully-supervised settings. Extensive experiments on two widely-used datasets for procedure understanding demonstrate the effectiveness of the proposed approach.

Keywords: Prompt learning, video-language pretrained models, instructional videos, procedure understanding, knowledge distilling.

Citation: G. Hu, B. He, H. Zhang. Compositional prompting video-language models to understand procedure in instructional videos. *Machine Intelligence Research*, vol.20, no.2, pp.249–262, 2023. http://doi.org/10.1007/s11633-022-1409-1

1 Introduction

Instructional videos refer to videos visually and orally demonstrating how to perform a daily task, such as "preparing a particular meal" and "repairing a car", which are very popular for learning and completing complex tasks in daily life. As indicated by the educational psychologist[1], dividing a whole task into smaller segments or procedures could largely simplify the complex task and facilitate the learning process for novices. Therefore, exploring intelligent algorithms to effectively understand procedures in instructional videos has wide applications in daily life, and will largely facilitate worldwide knowledge dissemination.

Procedure understanding in instructional videos involves various kinds of tasks, such as procedure recognition, procedure segmentation, procedure localization, procedure anticipation, and procedure retrieval[2–7]. In early works, each method is designed task-specifically that only tackles corresponding issues in one specific task[4, 8]. With the booming of self-supervised learning (SSL) and unsupervised learning in the fields of computer vision and natural language processing (NLP), most of the works fall in the paradigm of pretrain-finetune[9–13]. They first pretrain various large-scale video-language models to obtain general video-text representations since instructional videos naturally contain weakly-aligned clip-narration pairs, and then finetune the learned representation to various procedure understanding tasks in predetermined category space. The pioneering work VideoBERT[10] borrows ideas from word vectors in the NLP filed to create discrete video tokens to learn video-text representation from instructional videos via the BERT model[14]. ActBERT[9] further exploits global activity information, local activity information, and text information to learn general procedure representations. Some works[11, 15] also exploit the character of modality consistency in instructional videos to learn a joint representation of visual activity and linguistic narration in large-scale video-language models via the contrastive learning technique[16].

Most of these large-scale video-language models are trained by aligning short-term video clips with corresponding narrations (clip-narration pairs), such as ActBERT[9] (4 s), MIL-NCE[11] (3.2 s). However, different

Research Article
Special Issue on Large-scale Pre-training: Data, Models, and Finetuning
Manuscript received June 30, 2022; accepted December 13, 2022; published online March 2, 2023
Recommended by Associate Editor Da-Cheng Tao

from conventional short-term actions or activities[17–25], the procedures in instructional videos usually sustain much longer (1–100 s), such as procedures in the COIN dataset[3] have an average length of 14.91 s. Although pretrained video-language models perform well in representing short-term actions (verbs), basic objects (nouns), and their combinations, they are still not sufficient to analyze long-term procedures and are far more to understand the complex instructional tasks. Besides, this paradigm requires finetuning the downstream classifiers and localizers in a predetermined category space. Procedures in instructional videos are commonly composed of one or several actions and objects, such as "apply some glue on the boards" and "insert money into the vending machine". It is obvious that the predetermined procedure category faces the serious problem of combination disaster and is also inherently inapt to transfer to unseen actions, objects, and procedures.

To solve the problems mentioned above, we propose a novel prompt learning based framework called compositional prompt learning (CPL). It prompts short-term video-language models to understand long-term procedures and reformulates a series of classical procedure understanding tasks into general video-text matching problems, including procedure classification, procedure proposal, and procedure localization. Specifically, there are one visual prompt and three textual prompts consisting of an object prompt, action prompt, and procedure prompt to compositionally mine beneficial knowledge from pretrained video-language models. They tend to align the optimization targets of various downstream procedural tasks with the pretext target of pretrained video-language models, thus facilitating long-term procedure understanding. Besides, a series of procedure understanding tasks are reformulated into general video-text matching problems, making the proposed CPL also apt to zero-shot and few-shot conditions.

In summary, the contributions of this paper are as follows:

1) The proposed compositional prompt learning framework reformulates a series of classical procedure understanding tasks into general video-text matching problems, enabling it to not only be good at fully-supervised but also fit zero-shot and few-shot settings.

2) Three compositional textual prompts among the proposed CPL could hierarchically and compositionally distill dark knowledge from pretrained video-language models.

3) The proposed CPL is capable of stirring knowledge from short-term video-language models to understand long-term procedures in complex instructional videos.

4) The proposed method achieves promising performance on a series of procedure understanding tasks including classification, proposal generation, and temporal localization on two widely-used datasets.

2 Related works

2.1 Procedure understanding

The procedure understanding tasks in instructional videos have various types and targets, including procedure localization[2], procedure segmentation[2, 7], procedure caption[6], reference resolution[26], activity anticipation[7], procedure planning[8], skill determination[27], etc. According to the usage of manual annotations, previous works for instructional procedure understanding could be roughly categorized into three groups, i.e., fully-supervised methods, weakly-supervised methods, and unsupervised methods. With the emergence of deep learning, various fully-supervised neural networks have been applied to procedure understanding, such as multi-stream bi-directional recurrent neural network (MSB-RNN) and multi-stage temporal convolutional network (MS-TCN) for action segmentation[28, 29], ordering-dependency and task-consistency methods built on SSN[30] and R-C3D[31] for temporal procedure localization[3], etc. To decrease the burden of manual annotations, some works[32] adopt the Viterbi algorithm to solve the probabilistic model of weakly supervised procedure segmentation. Action Modifier[33] learns adverb representation from instructional activity with video-level weak supervision. The unsupervised procedure understanding approaches can be further categorized into two subgroups including task-oriented methods[7, 34, 35] and general video-language representation learning methods[9–11, 16]. As for the former class, some early works learn frame-wised continuous embedding and segment the instructional activities via clustering[36]. Sener and Yao[37] proposed a generalized mallows model (GMM) to model the distribution over sub-activity permutations. Since instructional video naturally contains weakly-aligned clip-narration pairs, it is much more suitable for self-supervised modality alignment. Recently, the SSL-based pretrain-finetune paradigm has dominated this domain, such as VideoBERT[10], ActBERT[9], and MIL-NCE[11]. They first pretrain various large-scale video-language models with weakly-aligned clip-narration pairs, and then finetune the classifiers and localizers for downstream procedure understanding tasks in specific category space.

2.2 Prompt learning

Prompt learning is a rapidly emerging topic that originated from the NLP field[38], which is originally designed for probing knowledge from large-scale pre-trained language models, such as BERT[14] or GPT[39]. Various NLP tasks (e.g., understanding tasks[40] and generation tasks[41]) are reformulated as the standard fill-in-blank pretext task (i.e., cloze-test) that is widely used in pre-

training large-scale language models. Prompting learning largely bridged the gap between pretext tasks of pretraining models and real downstream tasks. It has turned out to be a huge success for zero-shot learning, few-shot learning, and open-set learning. For more information, please refer to a comprehensive survey from [42]. Inspired by the success in the NLP field, many prompt learning based works in computer vision has also emerged. CoOp[43] borrows soft prompt technique in the NLP field to the image field. CPT[44] tailored for both image and text data is capable of explicitly grounding natural language to fine-grained image regions. CoCoOp[45] introduces a conditional prompt that is sample-specific to further improve the generalization of the soft prompts. ActionCLIP[46] changes the traditional action recognition tasks into a standard video-text matching problem. DenseCLIP[47] extends the prompt learning for dense prediction tasks, such as image segmentation. Ju et al.[48] encodes temporal information with a lightweight Transformer to prompt visual-language models for efficient video understanding. In this work, we consider compositional prompts to stir pretrained short-term visual-language models to tackle long-term procedure understanding tasks.

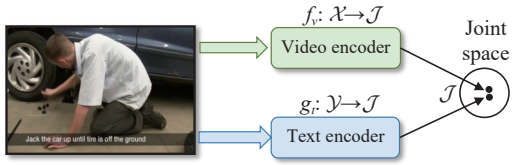
3 Methods

3.1 Preliminaries

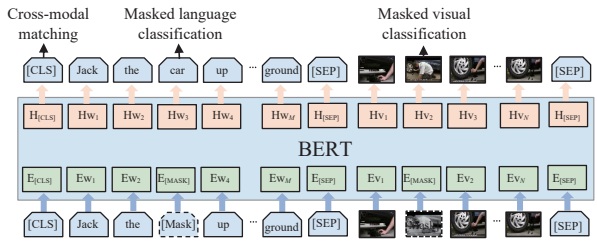
We first introduce some preliminary content about video-language pretrained models (VL-PTMs) from instructional procedure videos. Since the instructional video contains intrinsic clip-narration pairs that are semantically consistent, it is widely used for training large-scale video-language models via SSL. According to the SSL technique, the VL-PTMs from instructional videos could be categorized into two categories, i.e., contrastive learning (CL) based models[11, 15] and BERT based models[9, 10].

As shown in Fig. 1(a), given the weakly-aligned video-narration pairs $\{(x_i, y_j)\}_{i=1}^{n} \in (\mathcal{X}, \mathcal{Y})^n\}$, the contrastive learning based video-language models learn a joint distribution $P(\mathcal{X}, \mathcal{Y}; f_v, g_t)$ that embeds the video clip $x_i \in \mathcal{X}$ and text narration $y_i \in \mathcal{Y}$ in a joint semantic space $\mathcal{J} \in \mathbf{R}^d$, where $f_v \colon \mathcal{X} \to \mathcal{J}$ and $g_t \colon \mathcal{Y} \to \mathcal{J}$ are parameterized mapping functions for video clips and text narrations. The models f_v, g_t are commonly pretrained with the noise contrastive estimation (NCE) loss (1) or its multiple instances learning enhanced version (i.e., MIL-NCE (2))[11].

$$\max_{f_v, g_t} \sum_{i=1}^{n} \log \left(\frac{\mathrm{e}^{f_v(x_i)^{\mathrm{T}} g_t(y_i)}}{\mathrm{e}^{f_v(x_i)^{\mathrm{T}} g_t(y_i)} + \sum\limits_{(x', y') \sim \mathcal{N}_i} \mathrm{e}^{f_v(x')^{\mathrm{T}} g_t(y')}} \right) \tag{1}$$



(a) CL-based VL-PTMs



(b) BERT-based VL-PTMs

Fig. 1 Mechanism of typical VL-PTMs trained from large-scale untrimmed instructional videos, including paradigms based on the contrastive learning technique (a) and BERT model (b).

where $\mathcal{N}_i$ is the set of negative pairs.

$$\max_{f_v, g_t} \sum_{i=1}^{n} \log \left(\frac{\sum\limits_{(x,y) \in \mathcal{P}_i} \mathrm{e}^{f_v(x)^{\mathrm{T}} g_t(y)}}{\sum\limits_{(x,y) \in \mathcal{P}_i} \mathrm{e}^{f_v(x)^{\mathrm{T}} g_t(y)} + \sum\limits_{(x', y') \sim \mathcal{N}_i} \mathrm{e}^{f_v(x')^{\mathrm{T}} g_t(y')}} \right) \tag{2}$$

where $\mathcal{P}_i$ is the set of candidate positive pairs.

In addition, the BERT-based VL-PTMs (Fig. 1(b)) borrow the idea from pretraining language models on large corpora. Specifically, the input sequence for these models could be denoted as {[CLS], $w_1, \cdots, w_M$, [SEP], $v_1, \cdots, v_N$, [SEP]}, where $w_1, \cdots, w_M$ is the sequential embedding of text narration and $v_1, \cdots, v_N$ is the visual embedding of video clip, special token "[CLS]" and "[SEP]" denote classification and separation, respectively. The multi-modal BERT models are usually optimized with the pretext tasks of masked language classification, masked visual classification, and cross-modal matching. The masked language classification was proposed in [49], which is a standard pretext task for BERT-based models in the NLP filed[9, 10, 50]. Empirically, it randomly masks textual tokens with a probability of 15%, and replaces the masked tokens with a special token [MASK] 80% of the time, by a random textual token 10% of the time, and by the original token 10% of the time. Then, the masked language classifications are implemented by conducting a cloze test. Formally, it approximately maximizes the pseudo log-likelihood in the following:

$$L(\theta) = E_{x \sim D} \sum_{l} \log p\left(x_l \mid x_{\setminus l}; \theta\right) \tag{3}$$

where $x_{\setminus l} = (x_1, \cdots, x_{l-1}, [\mathrm{MASK}], x_{l+1}, \cdots, x_L)$ denotes masking the l-th token of input sequence $x \in D$ with previous empirical strategy, and θ is the learnable

parameter of corresponding multi-layer bidirectional transformer model[14]. The masked visual classification is nearly the same except that the input is changed from the textual token to the visual token. The cross-modal matching task is implemented by appending a linear layer upon the output of the first special token "[CLS]", which is a binary cross entropy to determine whether the inputted text and video are a positive pair. For more details about pretext tasks, we refer the readers to two surveys[51, 52].

After pretraining on abundant short-term clip-narration pairs, both the CL-based and BERT-based video-language models contain plenty of dark knowledge about basic short-term video-language concepts. More generally, the VL-PTMs generated from these two paradigms can be equally expressed as a text encoder f_v and a video encoder g_t that receptively embed textual narrations and short-term video clips into a joint semantic space $\mathcal{J}$. For simplification, we introduce and validate our framework on one representative CL-based model MIL-NCE[11] in this paper. Note that the proposed framework is not specified in this model but is broadly applicable to VL-PTMs from any mentioned paradigms above.

3.2 Compositional prompt learning

3.2.1 Overview

The pretrained video-language models perform well in representing fundamental concepts such as short-term actions (verbs), basic objects (nouns), and their combinations. However, it is far from sufficient to represent and understand long-term procedures which usually last as long as 1–100 s. Therefore, we propose a compositional prompting learning framework to stir off-the-shelf short-term video-language models to understand long-term instructional procedures.

As shown in Fig. 2, the proposed compositional prompt learning framework involves components of both the visual prompt and the compositional textual prompts. The visual prompt component is designed to temporally fuse the clip-level information from sequential short-term video clips $\{(x_i)\}_{i=1}^{t_p}$ to form procedure-level information, where t_p denotes the procedure length measured by video clip. The textual prompts consist of three compositional prompts, i.e. action prompt, object prompt, and procedure prompt. The action prompt and object prompt are respectively constructed from visible verbs and nouns in the procedure label. The procedure prompt is hierarchically constructed with the procedure label and the outputs of lower prompts (i.e., action and object prompts). Different from previous pretrain-finetune paradigm[9–11], the text encoder g_t and video encoder f_v in our CPL framework (Fig. 2) are always frozen during prompting and inference, which is very parameter-efficient. Eventually, a series of procedure understanding tasks are reformulated into general video-text matching problems, which will be introduced in detail in Section 3.2.4.

3.2.2 Textual prompts

Procedure labels usually contain a composition of one or several basic actions (verbs) and objects (nouns), such as "apply some glue on the boards", "mix and pickle", and "wind the junction to fasten the connection". Besides, the pretrained video-language models are good at aligning them with corresponding visual contents in joint semantic space $\mathcal{J}$. Therefore, we hierarchically organize the textual prompts with three parts (including action

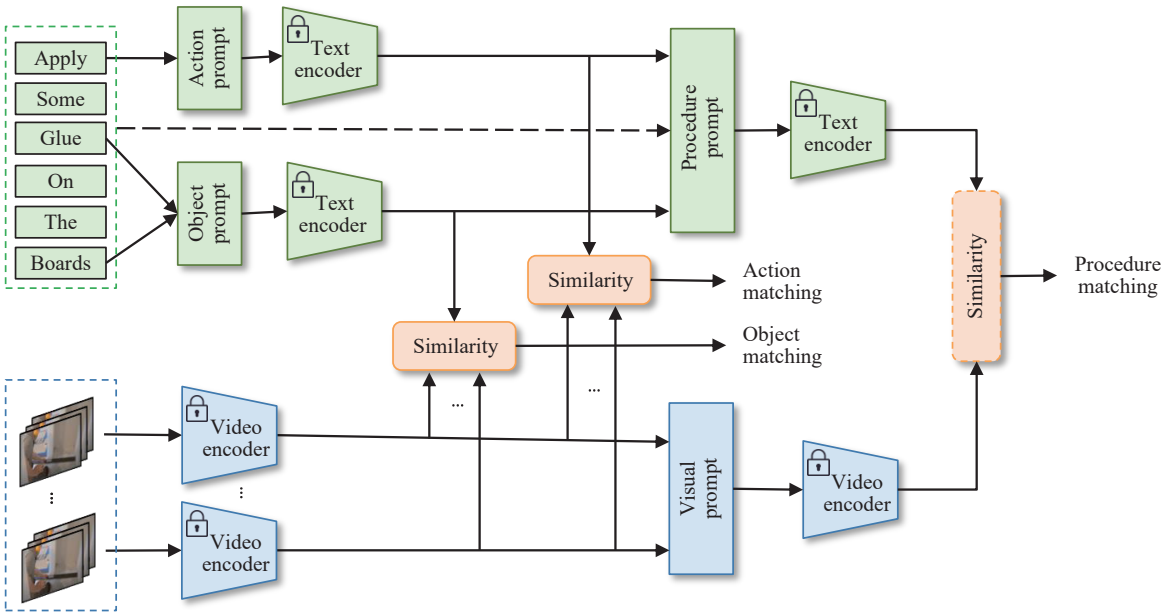


Fig. 2 Overview of the CPL framework. The video and text encoders from VL-PTMs are frozen during prompting and inference.

prompt, object prompt, and procedure prompt) and match their prompted embedding with that from video clips.

Action and object prompts. Given a procedure label y_p, we exploit the English core model from SpaCy[1] to tag its part-of-speech (PoS), thus obtaining corresponding verbs and nouns in the procedure label. After some simple manual filtering (see Section 4.2 for details), we treat the visible verbs and nouns as basic actions and objects, respectively. As a result, N_a and N_o actions and objects are obtained. Inspired by previous works CLIP[53] and CoOp[43], we design two types of text prompts, i.e., hand-crafted prompt and learnable continuous prompt.

Regarding the hand-crafted action prompt, a series of templates are constructed and applied to every action, such as {"An action of [*ActionClass*]."}, {"The video clip contains an action of [*ActionClass*]."}, {"Playing a kind of action, [*ActionClass*]."}, {"Doing action of [*ActionClass*]."}, and {"Can you recognize the action of [*ActionClass*]?"}. The [*ActionClass*] is filled with corresponding action text. The hand-crafted prompt does not introduce any learnable parameters, which intrinsically fits zero-shot learning. However, identifying suitable prompt templates needs sophisticated prompt engineering. Besides, as reported in both image and NLP domains[43, 54], different prompt templates have a prominent impact on the final performance of models. Therefore, our continuous version of the action prompt directly learns the soft prompts in an end-to-end manner. For each action class, the continuous prompt is constructed as follows:

$$template_a = \{[V_1][V_2]\cdots[V_{N_t}][ActionClass]\} \tag{4}$$

where $[V_n](n = 1, 2, \cdots, N_t)$ is a learnable vector with the same dimension as word embedding in the VL-PTMs, and the hyperparameter N_t controls the number of learnable tokens in the prompt template. Due to these learnable parameters, the continuous prompt is not applicable for zero-shot conditions.

Similarly, the hand-crafted or continuous prompt for object prompt could be also constructed in the same manner. Note that the numbers of actions and objects vary across procedures, we practically set the maximum numbers of action prompt and object prompt as $N^a_{\max}$ and $N^o_{\max}$, and pad the null classes "NullAction" and "NullObject" when needed, respectively.

Procedure prompt. The procedure prompt is built upon the procedure label and the lower action and object prompts. Given the actions y^a_i $(i = 1, 2, \cdots, N^a_{\max})$ and objects y^o_j $(j = 1, 2, \cdots, N^o_{\max})$ in a procedure label y, their outputs after corresponding prompt function and the frozen text encoder (see Fig. 2) could be obtained via (5) and (6), respectively.

$$h^a_i = g_t(prompt_a(y^a_i)) \tag{5}$$

$$h^o_j = g_t(prompt_o(y^o_j)) \tag{6}$$

where $prompt_a$ and $prompt_o$ are corresponding prompt function, and g_t is frozen text encoder. In addition, the continuous template for procedure prompt ($prompt_p$) could be designed as follows:

$$\{[V_1]\cdots[V_{N_t}][PClass]\ h^a_1\cdots h^a_{N^a_{\max}}\ h^o_1\cdots h^o_{N^o_{\max}}\}$$

where "[*PClass*]" should be filled with corresponding procedure name, $[V_n]$ $(n = 1, 2, \cdots, N_t)$ is a learnable vector (token). As for the hand-craft procedure prompt, the templates are constructed similarly to those for actions, such as {"A procedure of [*PClass*]. $h^a_1\cdots h^a_{N^a_{\max}}\ h^o_1\cdots h^o_{N^o_{\max}}$"} and {"The video clip contains a procedure of [*PClass*]. $h^a_1\cdots h^a_{N^a_{\max}}\ h^o_1\cdots h^o_{N^o_{\max}}$"}. Note that all the prompts for action, object, and procedure should be in the form of the hand-craft version when used for zero-shot conditions.

3.2.3 Visual prompt

Off-the-shelf VL-PTMs are usually pretrained with short-term clip-narration pairs that lack procedural-level temporal dynamics, inspired by ActionCLIP[46], we introduce a visual prompt function to empower the models with long-term temporal dependency. Formally, given sequential short-term video clips $\boldsymbol{x}_p = \{(x_i)\}^{t_p}_{i=1}$ of one procedure, the visual prompt function $prompt_v(f_v(\boldsymbol{x}_p))$ temporally forms a procedure level representation from clip-level information, where f_v is the short-term clip-wised video encoder branch of VL-PTMs, and t_p is the procedure length measured by the video clip (the number of video clips). So, the visual prompt simply aggregates representations from sequential video clips of one procedure to be a procedure-level representation. As shown in Fig. 3, there are two kinds of visual prompt functions used in this paper. The clip-wised average pooling (Fig. 3(a)) is the most straightforward yet very effective strategy, which simply averages pooling clip-wised representation along the temporal dimension. It will not introduce any learnable parameter and thus is applicable for zero-shot conditions. The second prompt function (Fig. 3(b)), equipped with an additional long short-term memory (LSTM) or 1D convolutional layer before average pooling, is elaborately designed to further capture temporal dynamics among clips. It introduces some additional parameters and could not be used for zero-shot settings.

3.2.4 Task reformulation

Classical procedure understanding tasks usually need task-specific classifiers or localizers, thus having different optimization targets with off-the-shelf VL-PTMs. To bridge the target gap between VL-PTMs and downstream tasks, we reformulate the form of downstream

[1] http://spacy.io

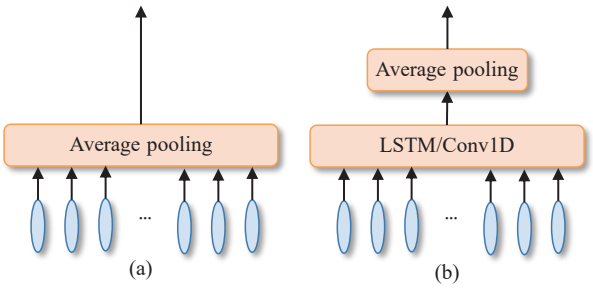


Fig. 3 Different types of visual prompts: (a) Clip-wised average prompt, and (b) average prompt enhanced with multi-layer LSTM or 1D convolutional layer for modeling temporal dynamics.

procedure understanding tasks into a general matching problem, which is also the main optimization target of VL-PTMs.

Procedure classification. Existing methods usually treat procedure classification as a close-set 1-of-N majority voting problem, while we treat it as an open-set video-text matching problem instead, as shown in Fig. 4(a). Formally, let a procedure video $x_i^p \in \boldsymbol{x}^p$ and its final semantic embedding $h_{x_i}^p = f_v(prompt_v(x_i^p))$, and let a procedure label $y_j^p \in \boldsymbol{y}^p$ and its final semantic embedding $h_{y_j}^p = g_t(prompt_{comp}(y_j^p))$. Given a procedure video $x_i^p \in \boldsymbol{x}^p$, we match it with each procedure label $y_j^p \in \boldsymbol{y}^p$ once by calculating the cosine similarity between their final semantic embeddings, then determine the best label match y_*^p as the final class prediction, i.e.,

$$y_*^p = \arg\min \langle h_{x_i}^p, h_{y_j}^p \rangle, \quad y_j^p \in \boldsymbol{y}^p \tag{7}$$

where $\langle \cdot \rangle$ denotes similarity calculation. As a result, the cosine classifier in (7) is not only capable of fully-supervised and few-shot procedure classification but is also naturally applicable for zero-shot recognition.

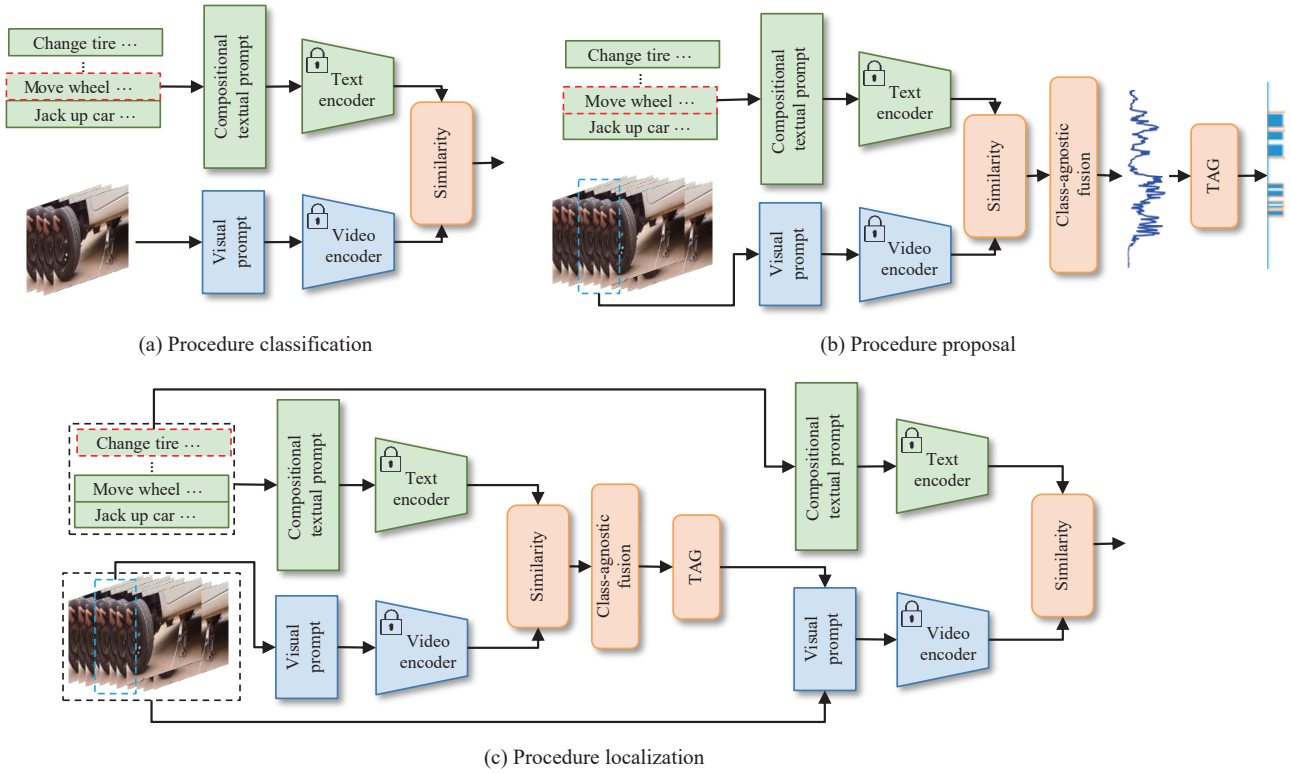
Procedure proposal. This task is aiming at generating a set of temporal proposals from an untrimmed instructional video that could overlap well with ground-truth procedures. Therefore, the generated proposals are category-independent. As shown in Fig. 4(b), given an untrimmed video $\boldsymbol{x}_i$, we first apply a temporal sliding window $w(t)$ with a length of W to calculate its time-dependent semantic embedding via the following equation:

$$h_{\boldsymbol{x}_i}(t) = f_v(prompt_v(\boldsymbol{x}_i^{w(t)})). \tag{8}$$

Then, we can obtain the similarity score $S_{ij}(t)$ between this time-dependent semantic embedding $h_{\boldsymbol{x}_i}(t)$ and the semantic embedding $h_{\boldsymbol{y}_j}$ from each procedure label $\boldsymbol{y}_j$, i.e.,

$$S_{ij}(t) = \langle h_{\boldsymbol{x}_i}(t), h_{\boldsymbol{y}_j} \rangle. \tag{9}$$

Since procedure proposals should be category-independent, we further introduce a module called class-ag-



(a) Procedure classification

(b) Procedure proposal

(c) Procedure localization

Fig. 4 Illustration of task reformulation. Three classical downstream tasks for procedure understanding are reformulated as general matching problems, including procedure classification, procedure proposal, and procedure localization.

nostic fusion (CAF) to adaptively fuse the score $S_{ij}(t)$ generated from each procedure label. Specifically, the CAF selects and averages the largest K scores for each sliding window, i.e.,

$$S_i(t) = mean(TopK(S_{ij}(t))), \quad j = 1, 2, \cdots, N_c^p \tag{10}$$

where N_c^p is the number of procedure classes. Finally, the class-agnostic score $S_i(t)$ is used to generate procedure proposals via the standard temporal actionness grouping (TAG) method[30], following the protocol in a previous work[55].

Procedure localization. As shown in Fig. 4(c), the procedure localization task can be decomposed into two stages, i.e., procedural proposal and proposal classification. Specifically, we first apply the reformulated proposal method to an untrimmed procedure video to obtain a series of procedure proposals. Then, apply the reformulated classification method to these proposals to generate the final class prediction. Both stages are based on video-text matching, so it is capable of stirring dark knowledge contained in the pretrained video-language models, which is more data-efficient and generalizable.

3.2.5 Optimization

As shown in Fig. 2, the proposed CPL contains three video-text matching components to stir hidden knowledge from VL-PTMs. The procedure matching forms the core matching loss $\mathcal{L}^p$, and the action matching and object matching form two auxiliary losses ($\mathcal{L}^a$ and $\mathcal{L}^o$), thus we have

$$\mathcal{L}_{match} = \mathcal{L}^p + \frac{(\mathcal{L}^a + \mathcal{L}^o)}{2}. \tag{11}$$

Each matching loss has a similar form as it is in classical prompt learning methods[53, 56]. Taking the procedure matching as an example, it is computed as the sum of the vision-to-text loss and text-to-vision loss, i.e.,

$$\mathcal{L}_i^p = \mathcal{L}_{v2t}^i + \mathcal{L}_{t2v}^i. \tag{12}$$

During prompting, the vision-to-text loss is constructed as follows:

$$\mathcal{L}_{v2t}^i = -\log\left(\frac{\exp(sim(h_{x_i}^p, h_{y_i}^p)/\tau)}{\sum_j \exp(sim(h_{x_i}^p, h_{y_j}^p)/\tau)}\right) \tag{13}$$

where sim is the cosine similarity function, and τ is a temperature parameter. The text-to-vision loss is similar to the vision-to-text loss except that one text label may correspond to multiple visual inputs, so it should be rewritten as follows:

$$\mathcal{L}_{t2v}^i = -\log\left(\frac{\exp(sim(h_{x_i}^p, h_{y_i}^p)/\tau)}{\sum_{j:y_j \neq y_i} \exp(sim(h_{x_j}^p, h_{y_i}^p)/\tau)}\right).$$

Note that the text set is formed from mini-batches during training while from the whole dataset during inference for memory-saving following [56].

4 Experiments

4.1 Datasets

HowTo100M dataset. The HowTo100M dataset[15] is the largest instructional video dataset, which involves broad domains in daily life, such as cooking, fitness, and gardening. It totally contains 1.22 M untrimmed YouTube videos belonging to 12 000 instructional tasks. Each video has corresponding narration from manually-entered subtitles or automatic speech recognition (ASR). The narrations and video clips are only weakly aligned for various reasons, including procedure missing, procedure reversing, unrelated narrations, etc. Since no ground-truths in terms of procedure classes or their temporal extents are available, the HowTo100M dataset is mainly used to train large-scale video-language models[9, 11, 15], including both CL-based and BERT-based VL-PTMs in Fig. 1. The example model we take for compositional prompt learning in this paper (i.e., MIL-NCE[11]) is also pretrained on this dataset.

COIN dataset. The COIN dataset[3] is currently the largest procedure understanding dataset with succinct and accurate manual procedure annotations with respect to the category and temporal extent. It contains 11 827 untrimmed YouTube videos from 180 daily tasks in 12 domains and has a large amount of 176 hours of videos. On average, each video lasts 2.36 minutes with about 3.91 procedures, and each procedure lasts about 14.91 s, thus eventually the whole dataset has 46 354 human-annotated procedures. Following the split in the original dataset paper[3], the training and testing subsets contain 9 030 and 2 797 video samples, respectively. Since the COIN dataset has clear temporal bounding boxes and succinct procedure descriptions, it could be used to evaluate many kinds of procedure understanding tasks, including procedure classification, procedure segmentation, procedure localization, procedure retrieval, etc.

CrossTask dataset. The CrossTask dataset[5] contains 4.7 K videos belonging to 83 tasks that fall into various domains, such as cooking, car maintenance, crating, and home repairs. Tasks and steps in the dataset are derived from the website wikiHow, which describes how to solve daily tasks. The primary tasks of CrossTask are fully-annotated and could be used to evaluate the temporal procedure localization task. Following the same evaluation protocols in previous works[5, 11], we report the average recall metric for the procedure localization task. It is defined by the number of procedure assignments that fall into the ground-truth extents divided by the total number of procedures.

4.2 Implementation details

We select an off-the-shelf VL-PTM from [11] as the short-term video-language model, which is pretrained on the HowTo100M dataset[15]. Following the settings in [11], the visual backbone is an S3D network[57] and the word embedding is initially from the word2vec model[58]. The token length for the textual prompt is set as 16, and the number of templates is set as 8 for all prompts, which are empirically obtained by grid search. The nouns and verbs among procedures are obtained by PoS tagging via SpaCy′s core web model for English. Inspired by previous works[15, 33], we manually filter out verbs that are not physically visible (such as "accept", "recommend") or with the VBD (past tense) tag that is not shown in the video (such as the verb "chopped" in the procedure "sprinkle some finely chopped coriander"). All the experiments are implemented with the PyTorch framework. The models are trained via the AdamW optimizer with improved weight decay handling and gradient clipping with a maximal norm of 0.1 is applied.

4.3 Results

To examine the effectiveness of the proposed framework, we conduct extensive experiments on three settings that have different levels of annotation availability, including fully-supervised, zero-shot and few-shot paradigms. The detailed results corresponding to each condition are reported in the following two sections.

4.3.1 Fully-supervised results

For the fully-supervised setting, we conduct extensive experiments for a series of procedure understanding tasks on the COIN and CrossTask datasets, including procedure classification, procedure proposal, and procedure localization.

Procedure classification. We first apply our CPL framework to the procedure classification task, which reformulates the traditional 1-of-N majority procedure voting into a matching problem between visual video clips and textual procedure descriptions. Therefore, we compare our novel framework with classical state-of-the-art action or activity recognition approaches on the COIN dataset, including TSN[59], TSM[60], STM[61], TDN[62], etc. As shown in Table 1, our CPL framework significantly outperforms the state-of-the-art action recognition method TDN[62] with large margins of 4.7% (Top1) and 8.3% (Top5). The performance gains mainly owing to two reasons: 1) Our CPL framework makes full use of the semantic information contained in the procedure description, and 2) the compositional textual prompting strategy is good at distilling the component knowledge in large-scale pretained video-language models.

Procedure proposal. In addition, we further examine the effectiveness of our CPL via the procedure proposal task, which aims to segment an instructional procedure video into category-independent procedure segments. To conduct a par-to-par comparison, we follow the setting and baselines from a previous work[63]. Specifically, the intersection over union (IoU) threshold for non-maximum suppression is set to 0.8, and the average recall (AR) is computed via multiple IoU thresholds that vary from 0.5 to 0.95 with an interval of 0.05. The final ARs with respect to different ANs (40, 60, 80) are reported in Table 2, where AN denotes the average number (AN) of proposals per video. As shown in Table 2, our CPL framework achieves state-of-the-art performance on both of COIN and CrossTask datasets. Specifically, the proposed CPL not only outperforms the weakly-supervised methods, such as Hand Detector and Temporal Prior[63, 64], but also exceeds the state-of-the-art supervised method[63] with large margins in terms of different proposal numbers per video.

Table 1 Comparisons of procedure classification performance on the COIN dataset

Methods	Accuracy (Top1)	Accuracy (Top5)
TSN[59]	36.5%	65.1%
TSM[60]	37.9%	69.3%
STM[61]	38.5%	72.7%
TDN[62]	40.2%	78.5%
CPL (Ours)	**44.9**%	**86.8**%

Procedure localization. We finally evaluate the performance of the procedure localization task, which aims to localize the temporal extent of a specific procedure and classify its category at the same time. To conduct a par-to-par comparison, following the protocols in previous works[5, 63], we also use class-wise recalls and the average recall as the evaluation metrics for this task. The obtained results are included in Table 3. We compare them with a series of state-of-the-art methods, including the approaches from Richard et al.[32], Alayrac et al.[34], Zhukov et al.[5], Miech et al.[11], and TVJE[15], LTOV[63], etc. The performance regarding class-wise recalls and the average recall both achieve a new state-of-the-art result on the CrossTask benchmark. The large performance gain could be partly attributed to the fact that our CPL method successfully mined useful semantic information that is already well represented in the video-language model MIL-NCE, which is pretrained on the large-scale untrimmed instructional dataset HowTo100M. In addition, the gain should be partly attributed to the reason that the proposed compositional prompt learning hierarchically aligns the optimization targets between the pretext task of VL-PTMs and the downstream localization task.

4.3.2 Zero-shot and few-shot results

As introduced in Section 3.2.4, the proposed prompt learning based framework also has great potential in performing zero-shot and few-shot procedure understanding. In this section, we carefully examine its performance un-

Table 2 Comparisons of procedure proposal performance on the COIN and CrossTask datasets. Average recall under conditions of different proposal numbers per video is reported, including $AN = 40$, $AN = 60$ and $AN = 80$.

	COIN			CrossTask		
	$AN = 40$	$AN = 60$	$AN = 80$	$AN = 40$	$AN = 60$	$AN = 80$
Random	0.01	0.02	0.03	0.01	0.02	0.03
Hand detector[64]	0.15	0.19	0.22	0.06	0.07	0.08
Temporal prior[63]	0.22	0.28	0.34	0.11	0.14	0.18
LTOV[63]	0.25	0.35	0.41	0.17	0.23	0.27
Linear supervision[63]	0.30	0.39	0.45	0.21	0.27	0.33
CPL (Ours)	**0.35**	**0.44**	**0.52**	**0.30**	**0.38**	**0.46**

Table 3 Comparisons of procedure localization performance on the CrossTask dataset. Recall score for each task and their average recall score are reported.

Methods	Make kimchi rice	Pickle cucumber	Make banana ice cream	Grill steak	Jack up car	Make jello shots	Change tire	Make lemonade	Add oil to car	Make latte
Richard et al.[32]	7.6	4.3	3.6	4.6	8.9	5.4	7.5	7.3	3.6	6.2
Alayrac et al.[34]	15.6	10.6	7.5	14.2	9.3	11.8	17.3	13.1	6.4	12.9
Zhukov et al.[5]	13.3	18.0	23.4	23.1	16.9	16.5	30.7	21.6	4.6	19.5
TVJE[15]	33.5	27.1	36.6	37.9	24.1	35.6	32.7	35.1	30.7	28.5
Miech et al.[11]	28.7	37.9	42.8	36.3	22.0	42.9	27.4	43.1	30.8	32.7
LTOV[63]	34.1	40.0	48.7	40.3	30.7	46.1	34.5	45.9	38.1	35.9
CPL (Ours)	**38.5**	**42.1**	**49.8**	**45.0**	**34.2**	**49.7**	**38.8**	**47.2**	**40.9**	**39.1**

Methods	Build shelves	Make taco salad	Make French toast	Make Irish coffee	Make strawberry cake	Make pancakes	Make meringue	Make fish curry	Average
Richard et al.[32]	12.3	3.8	7.4	7.2	6.7	9.6	12.3	3.1	6.7
Alayrac et al.[34]	27.2	9.2	15.7	8.6	16.3	13.0	23.2	7.4	13.3
Zhukov et al.[5]	35.3	10.0	32.3	13.8	29.5	37.6	43.0	13.3	22.4
TVJE[15]	43.2	19.8	34.7	33.6	40.4	41.6	41.9	27.4	33.6
Miech et al.[11]	42.8	27.5	34.0	33.7	44.3	48.0	46.0	33.9	36.4
LTOV[63]	50.0	35.4	38.1	42.6	42.6	45.9	51.6	37.8	41.0
CPL (Ours)	**54.2**	**36.1**	**42.7**	**44.9**	**47.8**	**51.3**	**53.1**	**39.8**	**44.2**

der different levels of annotation availability. Note that the hand-crafted version of prompt functions is used for the zero-shot condition while the continuous version is used for the few-shot condition.

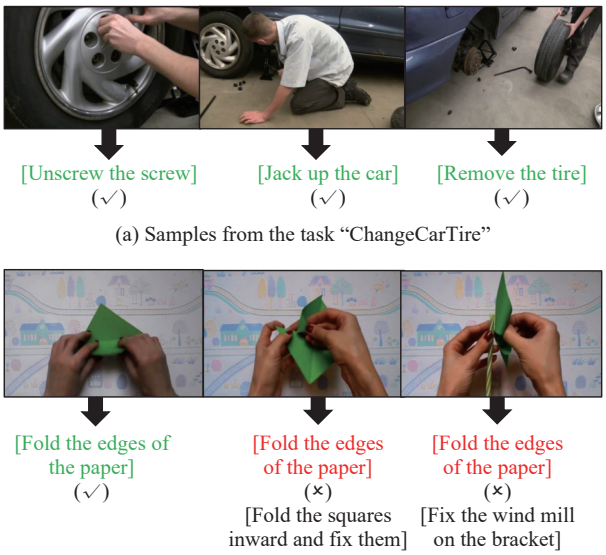
We first conduct extensive experiments for procedure classification on the COIN datasets. Specifically, we train our framework with various percentages of training data, including 0%, 1%, 2%, 5%, 10% and 100%. Thus, 0% means no supervised training that is a zero-shot setting and 100% means previous fully-supervised setting, while 1%, 2%, 5% and 10% are different levels of few-shot settings. Since our CPL framework reformulates procedure classification as a novel video-text matching problem (Section 3.2.4), we also construct a baseline method by finetuning the same VL-PTMs under classical paradigm of 1-of-N majority voting for comparing. The classification results corresponding to these settings are reported in Table 4. We can observe that our CPL achieves promising zero-shot learning ability that reaches about 22% performance of the full-supervised setting (upper bound). While the baseline finetuning method cannot perform zero-shot classification, because its downstream 1-of-N voting paradigm is different from the video-text matching paradigm of VL-PTMs in that at least one annotation is needed for alignment. Our CPL framework is capable of distilling dark knowledge from pretrained video-

Table 4 Comparing procedure classification results of the proposed CPL framework under different levels of annotation availability, including the zero-shot, few-shot and fully-supervised settings on the COIN dataset.

Methods	Settings					
	Zero-shot	Few-shot				Fully-supervised
	0%	1%	2%	5%	10%	100%
Baseline	–	0.7	2.1	6.7	11.4	39.7
CPL (Ours)	10.3	17.5	20.9	25.9	34.4	44.9

language models, making it have a zero-shot understanding of some key components of instructional procedures, such as basic actions and objects. Although the procedure descriptions are heterogeneous in large-scale auto-generated pretraining datasets (e.g., HowTo100M) and downstream human-annotated datasets (e.g., COIN, CrossTask), they are still semantically equivalent at the high-level semantic space. Besides, the proposed method is surprisingly good at few-shot learning, reaching nearly 38% and 77% of the fully-supervised performances (upper bound) with only 1% and 10% procedure annotations, respectively. While the baseline method only respectively achieves 1.7% and 29% of the upper bound performance with 1% and 10% annotations under a comparable upper bound (fully-supervised). The excellent improvements could be largely attributed to the fact that our prompt learning based framework aligns the optimization targets between the downstream classification task and pertained pretext task. We also visualize some positive and negative results from our CPL framework under the few-shot condition (10%) in Fig. 5. We can observe that the model successfully classified procedural samples in the task "ChangeCarTire" after few-shot training, partially because it may stir useful concepts (such as "wheel", "screw", "jack", "remove") from the VL-PTMs. However, it made many mistakes in task "MakePaperWindMill". This may be because the procedures in "MakePaperWindMill" are more intra-class similar and the involved concepts (such as "windmill", "bracket", and "edge") are less common in the VL-PTMs, which may need more annotations to learn and distinguish.

Similarly, we also evaluate the zero-shot and few-shot abilities of our CPL framework for procedure localization task on the CrossTask dataset, and the results are shown in Table 5. For a fair comparison, the baseline method also divides procedure localization into two stages (procedure proposal and proposal classification), and each stage is achieved by finetuning the same VL-PTMs as ours under the classical paradigm of 1-of-N majority voting instead of our matching paradigm. We can observe that our CPL method has a promising ability for zero-shot and few-shot learning. It achieves high average recalls of 26.5% (60% of upper bound performance) with only 10% annotated data while the baseline method only reaches 12.7% (31% of upper bound performance) in the same annotation condition, indicating the effectiveness and superiority of the proposed method.



(a) Samples from the task "ChangeCarTire"

(b) Samples from the task "MakePaperWindMill"

Fig. 5 Visualization of the positive and negative classification results on the COIN dataset. The correct and incorrect predictions are respectively marked in blue and red, and the ground-truth for the incorrect prediction is also listed in black.

Table 5 Comparing procedure localization results of our CPL framework on the CrossTask dataset under different levels of annotation availability, including zero-shot, few-shot and fully-supervised settings.

Methods	Settings					
	Zero-shot	Few-shot				Fully-supervised
	0%	1%	2%	5%	10%	100%
Baseline	–	0.9	2.4	5.5	12.7	40.5
CPL (Ours)	12.3	15.8	17.6	22.8	26.5	44.2

4.3.3 Analysis

In this section, we dig deeper into the proposed CPL framework to examine and analyze the influence of some key components and hyper-parameters. Specifically, we will analyze the influence of the token numbers and template numbers in the textual prompts, and the influence of prompt types in the visual prompts.

Firstly, we take the procedure classification on the COIN dataset as an example to investigate the impact of token numbers in continuous textual prompts. Specifically, we fix the template number and vary the learnable token number N_t of the textual prompt in a range [4 : 4 : 24], and the performance of procedure classification is shown in Table 6. We can observe that the accuracy improves quickly when increasing the token number at early stage since more tokens could provide more context capacity for the textual prompts. Then, the accuracy is gradually saturated and achieves a maximum value with a token number of 16, thus we choose this value as the

Table 6 Influence of the token number on procedure classification performance on the COIN dataset. †selected value.

#Tokens	4	8	12	16†	20	24
Accuracy (%)	39.2	42.4	43.6	**44.9**	44.5	44.2

default value for other experiments.

Then, we examine the impact of template number by procedure classification task on the COIN dataset. As mentioned in Section 3.2.2, the discrete prompt needs sophisticated prompt engineering, and different prompt templates have a prominent impact on the final performance of models, which has been widely reported in both fields of computer vision (CV) and NLP[43, 54]. Therefore, our CPL mainly uses continuous (learnable) prompts except for the zero-shot setting. Here, we focus on the impact of the template number of the continuous prompt. Specifically, the number of templates is varied in a range {1, 4, 8, 12, 16}, and the results are shown in Table 7. We can observe that the classification performance increases with the number of prompt templates and is saturated after 8 templates. Although the performances are the same when there are 8 and 12 templates, we eventually choose 8 as the default value in consideration of computational efficiency.

Table 7 Influence of template number on procedure classification performance on the COIN dataset. †selected value.

#Templates	1	4	8†	12	16
Accuracy (%)	39.7	43.1	**44.9**	**44.9**	44.7

As mentioned in Section 3.2.3 and Fig. 3, the visual prompt could be in the form of two categories, including the non-parameterized average pooling and parameterized LSTM/Conv1D appended by average pooling. The non-parameterized manner is applicable to all the zero-shot, few-shot, and fully-supervised settings, and the parameterized manner can only be used for the few-shot and fully-supervised settings. Here, we take the task of procedure localization on the CrossTask as an example to examine the impact of different visual prompts. As shown in Table 8, the two parameterized manners (LSTM/Conv1D + average pooling) achieve comparable performance in all settings, and both are much better than the non-parameterized manner (averaging pooling). This is because the proposed two non-parameterized methods are capable of capturing temporal dynamics among representations of video clips.

Table 8 Influence of different visual prompts on procedure localization performance on the CrossTask dataset. †selected.

Types	Settings				
	Few-shot				Fully-supervised
	1%	2%	5%	10%	100%
AveragePooling	13.2	14.9	19.5	23.8	41.7
Conv1D+AveragePooling	15.6	17.1	**23.0**	26.1	44.0
LSTM+AveragePooling†	**15.8**	**17.6**	22.8	**26.5**	**44.**2

5 Conclusions

In this work, we propose a novel prompt learning based framework to compositionally prompt pretrained short-term video-language models to efficiently perform long-term procedure understanding tasks in instructional videos. Three compositional textual prompts and one visual prompt reformulate three classical procedure understanding tasks into general video-text matching problems. The framework efficiently transfers dark knowledge from off-the-shelf video-language models to facilitate downstream tasks under the umbrella of zero-shot, few-shot, and fully-supervised settings. Eventually, our CPL achieves promising results on two widely used instructional datasets for procedure understanding tasks.

References

[1] R. J. Nadolski, P. A. Kirschner, J. J. van Merriënboer. Optimizing the number of steps in learning tasks for complex skills. *British Journal of Educational Psychology*, vol. 75, no. 2, pp. 223–237, 2005. DOI: 10.1348/000709904X22403.

[2] M. Rohrbach, S. Amin, M. Andriluka, B. Schiele. A database for fine grained activity detection of cooking activities. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Providence, USA, pp. 1194–1201, 2012. DOI: 10.1109/CVPR.2012.6247801.

[3] Y. S. Tang, D. J. Ding, Y. M. Rao, Y. Zheng, D. Y. Zhang, L. L. Zhao, J. W. Lu, J. Zhou. COIN: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 1207–1216, 2019. DOI: 10.1109/CVPR.2019.00130.

[4] Y. A. Farha, A. Richard, J, Gall. When will you do what? – Anticipating temporal occurrences of activities. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 5343–5352, 2018. DOI: 10.1109/CVPR.2018.00560.

[5] D. Zhukov, J. B. Alayrac, R. G. Cinbis, D. Fouhey, I. Laptev, J. Sivic. Cross-task CVFweakly supervised learning from instructional videos. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 3532–3540, 2019. DOI: 10.1109/CVPR.2019.00365.

[6] H. Kuehne, A. Arslan, T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, USA, pp. 780–787, 2014. DOI: 10.1109/CVPR.2014.105.

[7] L. W. Zhou, C. L. Xu, J. J. Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, New Orleans, USA, pp. 7590–7598, 2018. DOI: 10.5555/3504035.3504965.

[8] C. Y. Chang, D. A. Huang, D. F. Xu, E. Adeli, L. Fei-Fei, J. C. Niebles. Procedure planning in instructional videos. In *Proceedings of the 16th European Conference on Computer Vision*, Springer, Glasgow, UK, pp. 334–350, 2020.

DOI: 10.1007/978-3-030-58621-8_20.

[9] L. C. Zhu, Y. Yang. ActBERT: Learning global-local video-text representations. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 8743–8752, 2020. DOI: 10.1109/CVPR42600.2020.00877.

[10] C. Sun, A. Myers, C. Vondrick, K. Murphy, C. Schmid. VideoBERT: A joint model for video and language representation learning. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Seoul, Repubic of Korea, pp. 7463–7472, 2019. DOI: 10.1109/ICCV.2019.00756.

[11] A. Miech, J. B. Alayrac, L. Smaira, I. Laptev, J. Sivic, A. Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 9876–9886, 2020. DOI: 10.1109/CVPR42600.2020.00990.

[12] B. Cui, G. Y. Hu, S. Yu. DeepCollaboration: Collaborative generative and discriminative models for class incremental learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pp. 1175–1183, 2021. DOI: 10.1609/aaai.v35i2.16204.

[13] J. P. Zhang, J. M. Zhang, G. Y. Hu, Y. Chen, S. Yu. Scalenet: A convolutional network to extract multi-scale and fine-grained visual features. *IEEE Access*, vol. 7, pp. 147560–147570, 2019. DOI: 10.1109/ACCESS.2019.2946425.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, USA, pp. 6000–6010, 2017.

[15] A. Miech, D. Zhukov, J. B. Alayrac, M. Tapaswi, I. Laptev, J. Sivic. HowTo100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Repubic of Korea, pp. 2630–2640, 2019. DOI: 10.1109/ICCV.2019.00272.

[16] K. M. He, H. Q. Fan, Y. X. Wu, S. N. Xie, R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 9726–9735, 2020. DOI: 10.1109/CVPR42600.2020.00975.

[17] G. Hu, B. Cui, S. Yu. Skeleton-based action recognition with synchronous local and non-local spatio-temporal learning and frequency attention. In *Proceedings of IEEE International Conference on Multimedia and Expo*, Shanghai, China, pp. 1216–1221, 2019. DOI: 10.1109/ICME.2019.00212.

[18] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, R. Memisevic. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of IEEE International Conference on Computer Vision*, Venice, Italy, pp. 5843–5851, 2017. DOI: 10.1109/ICCV.2017.622.

[19] G. Y. Hu, B. Cui, S. Yu. Joint learning in the spatio-temporal and frequency domains for skeleton-based action recognition. *IEEE Transactions on Multimedia*, vol. 22, no. 9, pp. 2207–2220, 2020. DOI: 10.1109/TMM.2019.2953325.

[20] F. C. Heilbron, V. Escorcia, B. Ghanem, J. C. Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 961–970, 2015. DOI: 10.1109/CVPR.2015.7298698.

[21] G. Y. Hu, B. Cui, Y. He, S. Yu. Progressive relation learning for group activity recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 977–986, 2020. DOI: 10.1109/CVPR42600.2020.00106.

[22] M. S. Liu, J. Q. Gao, G. Y. Hu, G. F. Hao, T. Z. Jiang, C. Zhang, S. Yu. MonkeyTrail: A scalable video-based method for tracking macaque movement trajectory in daily living cages. *Zoological Research*, vol. 43, no. 3, pp. 343–351, 2022. DOI: 10.24272/j.issn.2095-8137.2021.353.

[23] B. X. Wu, C. G. Yang, J. P. Zhong. Research on transfer learning of vision-based gesture recognition. [Online], Available: https://dblp.org/rec/journals/corr/abs-1812-05770.html?view=bibtex, 2021.

[24] Z. W. Xu, X. J. Wu, J. Kittler. STRNet: Triple-stream spatiotemporal relation network for action recognition. [Online], Available: https://dblp.org/rec/conf/cvpr/WuGHFK20.html?view=bibtex, 2021.

[25] L. F. Wu, Q. Wang, M. Jian, Y. Qiao, B. X. Zhao. A comprehensive review of group activity recognition in videos. *International Journal of Automation and Computing*, vol. 18, no. 3, pp. 334–350, 2021. DOI: 10.1007/s11633-020-1258-8.

[26] D. A. Huang, J. J. Lim, L. Fei-Fei, J. C. Niebles. Unsupervised visual-linguistic reference resolution in instructional videos. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 1032–1041, 2017. DOI: 10.1109/CVPR.2017.116.

[27] H. Doughty, D. Damen, W. Mayol-Cuevas. Who′s better? Who′s best? Pairwise deep ranking for skill determination. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 6057–6066, 2018. DOI: 10.1109/CVPR.2018.00634.

[28] B. Singh, T. K. Marks, M. Jones, O. Tuzel, M. Shao. A multi-stream Bi-directional recurrent neural network for fine-grained action detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 1961–1970, 2016. DOI: 10.1109/CVPR.2016.216.

[29] Y. A. Farha, J. Gall. MS-TCN: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 3570–3579, 2019. DOI: 10.1109/CVPR.2019.00369.

[30] Y. Zhao, Y. J. Xiong, L. M. Wang, Z. R. Wu, X. O. Tang, D. H. Lin. Temporal action detection with structured segment networks. In *Proceedings of IEEE International Conference on Computer Vision*, Venice, Italy, pp. 2933–2942, 2017. DOI: 10.1109/ICCV.2017.317.

[31] H. J. Xu, A. Das, K. Saenko. R-C3D: Region convolutional 3D network for temporal activity detection. In *Proceedings of IEEE International Conference on Computer Vision*, Venice, Italy, pp. 5794–5803, 2017. DOI: 10.1109/ICCV.2017.617.

[32] A. Richard, H. Kuehne, J. Gall. Action sets: Weakly supervised action segmentation without ordering constraints. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt

Lake City, USA, pp. 5987–5996, 2018. DOI: 10.1109/CVPR.2018.00627.

[33] H. Doughty, I. Laptev, W. Mayol-Cuevas, D. Damen. Action modifiers: Learning from adverbs in instructional videos. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 865–875, 2020. DOI: 10.1109/CVPR42600.2020.00095.

[34] J. B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, S. Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 4575–4583, 2016. DOI: 10.1109/CVPR.2016.495.

[35] S. N. Aakur, S. Sarkar. A perceptual prediction framework for self supervised event segmentation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 1197–1206, 2019. DOI: 10.1109/CVPR.2019.00129.

[36] A. Kukleva, H. Kuehne, F. Sener, J. Gall. Unsupervised learning of action classes with continuous temporal embedding. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 12058–12066, 2019. DOI: 10.1109/CVPR.2019.01234.

[37] F. Sener, A. Yao. Unsupervised learning and segmentation of complex activities from video. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 8368–8376, 2018. DOI: 10.1109/CVPR.2018.00873.

[38] T. X. Sun, X. Y. Liu, X. P. Qiu, X. J. Huang. Paradigm shift in natural language processing. *Machine Intelligence Research*, vol. 19, no. 3, pp. 169–183, 2022. DOI: 10.1007/s11633-022-1331-6.

[39] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, vol. 1, no. 8, Article number 9, 2019.

[40] T. Schick, H. Schütze. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2339–2352, 2021. DOI: 10.18653/v1/2021.naacl-main.185.

[41] X. L. Li, P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 4582–4597, 2021. DOI: 10.18653/v1/2021.acl-long.353.

[42] P. F. Liu, W. Z. Yuan, J. L. Fu, Z. B. Jiang, H. Hayashi, G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. [Online], Available: https://arxiv.org/abs/2107.13586, 2021.

[43] K. Y. Zhou, J. K. Yang, C. C. Loy, Z. W. Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022. DOI: 10.1007/s11263-022-01653-1.

[44] Y. Yao, A. Zhang, Z. Y. Zhang, Z. Y. Liu, T. S. Chua, M. S. Sun. CPT: Colorful prompt tuning for pre-trained vision-language models. [Online], Available: https://arxiv.org/abs/2109.11797, 2021.

[45] K. Y. Zhou, J. K. Yang, C. C. Loy, Z. W. Liu. Conditional prompt learning for vision-language models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, New Orleans, USA, pp. 16795–16804, 2022. DOI: 10.1109/CVPR52688.2022.01631.

[46] M. M. Wang, J. Z. Xing, Y. Liu. ActionCLIP: A new paradigm for video action recognition. [Online], Available: https://arxiv.org/abs/2109.08472, 2021.

[47] Y. M. Rao, W. L. Zhao, G. Y. Chen, Y. S. Tang, Z. Zhu, G. Huang, J. Zhou, J. W. Lu. DenseCLIP: Language-guided dense prediction with context-aware prompting. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, New Orleans, USA, pp. 18061–18070, 2022. DOI: 10.1109/CVPR52688.2022.01755.

[48] C. Ju, T. D. Han, K. H. Zheng, Y. Zhang, W. D. Xie. Prompting visual-language models for efficient video understanding. In *Proceedings of the 17th European Conference on Computer Vision*, Springer, Tel Aviv Israel, pp. 105–124, 2022. DOI: 10.1007/978-3-031-19833-5_7.

[49] W. L. Taylor. "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, vol. 30, no. 4, pp. 415–433, 1953. DOI: 10.1177/107769905303000401.

[50] J. Devlin, M. W. Chang, K. Lee, K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, USA, pp. 4171–4186, 2019. DOI: 10.18653/v1/N19-1423.

[51] Z. Gan, L. J. Li, C. Y. Li, L. J. Wang, Z. C. Liu, J. F. Gao. Vision-language pre-training: Basics, recent advances, and future trends. [Online], Available: https://arxiv.org/abs/2210.09263, 2022.

[52] F. L. Chen, D. Z. Zhang, M. L. Han, X. Y. Chen, J. Shi, S. Xu, B. Xu. VLP: A survey on vision-language pre-training. [Online], Available: https://arxiv.org/abs/2202.09061, 2022.

[53] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

[54] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, S. Singh. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pp. 4222–4235, 2020. DOI: 10.18653/v1/2020.emnlp-main.346.

[55] T. W. Lin, X. Zhao, H. S. Su, C. J. Wang, M. Yang. BSN: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 3–21, 2018. DOI: 10.1007/978-3-030-01225-0_1.

[56] S. C. Wang, Y. Q. Duan, H. H. Ding, Y. P. Tan, K. H. Yap, J. S. Yuan. Learning transferable human-object interaction detector with natural language supervision. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, New Orleans, USA, pp. 929–938, 2022. DOI: 10.1109/CVPR52688.2022.00101.

[57] S. N. Xie, C. Sun, J. Huang, Z. W. Tu, K. Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Mu-

nich, Germany, pp. 318–335, 2018. DOI: 10.1007/978-3-030-01267-0_19.

[58] T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient estimation of word representations in vector space. [Online], Available: https://arxiv.org/abs/1301.3781, 2013.

[59] L. M. Wang, Y. J. Xiong, Z. Wang, Y. Qiao, D. H. Lin, X. O. Tang, L. van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the 14th European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, pp. 20–36, 2016. DOI: 10.1007/978-3-319-46484-8_2.

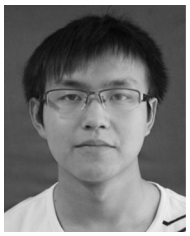
[60] J. Lin, C. Gan, S. Han. TSM: Temporal shift module for efficient video understanding. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Repubic of Korea, pp. 7082–7092, 2019. DOI: 10.1109/ICCV.2019.00718.

[61] B. Y. Jiang, M. M. Wang, W. H. Gan, W. Wu, J. J. Yan. STM: Spatiotemporal and motion encoding for action recognition. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Repubic of Korea, pp. 2000–2009, 2019. DOI: 10.1109/ICCV.2019.00209.

[62] L. M. Wang, Z. Tong, B. Ji, G. S. Wu. TDN: Temporal difference networks for efficient action recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 1895–1904, 2021. DOI: 10.1109/CVPR46437.2021.00193.

[63] D. Zhukov, J. B. Alayrac, I. Laptev, J. Sivic. Learning actionness via long-range temporal order verification. In *Proceedings of the 16th European Conference on Computer Vision*, Springer, Glasgow, UK, pp. 470–487, 2020. DOI: 10.1007/978-3-030-58526-6_28.

[64] D. D. Shan, J. Q. Geng, M. Shu, D. F. Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 9866–9875, 2020. DOI: 10.1109/CVPR42600.2020.00989.



Guyue Hu received the B. Eng. degree in automation from Hefei University of Technology, China in 2016, and the Ph. D. degree in pattern recognition and intelligent systems from National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), China in 2021. He was also a research fellow with School of Computing, National University of Singapore, Singapore from 2021 to 2022. He is currently a research fellow with School of Computer Science and Engineering, Nanyang Technological University, Singapore. He serves as a regular reviewer for a number of international journals and conferences, such as TPAMI, TMM, TCSVT, CVPR, ICCV, ECCV.

His research interests include computer vision, pattern recognition, and computational neuroscience, especially in multi-modal learning, video understanding, and human activity analysis.

E-mail: guyue.hu@ntu.edu.sg (Corresponding author)

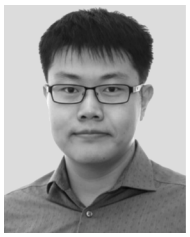
ORCID iD: 0000-0002-6198-8230

Bin He received the B. Eng. degree in automation from Harbin University of Science and Technology, China in 2014, and the Ph. D. degree in mechanical and electronic engineering from Harbin University of Science and Technology, China in 2020. As a joint Ph. D. student, he finished the entire doctoral research at National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), China. He is currently an engineer at North China Computing Technology Institute (alias The 15th Research Institute of China Electronics Technology Group Corporation), China.

His research interests include computer vision, pattern recognition, and intelligent decision-making, especially lie in military intelligence.

E-mail: binhe.cas@foxmail.com

ORCID iD: 0000-0002-3845-7335



Hanwang Zhang received the B. Eng. degree in computer science from Zhejiang University, China in 2009, and the Ph. D. degree in computer science from National University of Singapore, Singapore in 2014. He was a research scientist with Department of Computer Science, Columbia University, USA from 2017 to 2018, and a research fellow with National University of Singapore from 2014 to 2016. He is currently an assistant professor at School of Computer Science and Engineering, Nanyang Technological University, Singapore. He has authored more than 150 scientific papers in these areas in top journals and conferences, including TPAMI, TIP, ICLR, NeurIPS, CVPR, ICCV, ECCV, ACL, EMNLP, etc. He is the recipient of the Best Demo Runner-up Award in ACM MM 2012, the Best Student Paper Award in ACM MM 2013, the Best Paper Honorable Mention in ACM SIGIR 2016, and TOMM Best Paper Award 2018. He is also the Winner of the Best Ph. D. Thesis Award of School of Computing, National University of Singapore, Singapure in 2014.

His research interests include computer vision and multimedia, especially focusing on the fusion of deep learning and reasoning for these fields.

E-mail: hanwangzhang@ntu.edu.sg

ORCID iD: 0000-0001-7374-8739