Boosting Generic Visual-Linguistic Representation with Dynamic Contexts

Guoqing Ma[1], Yalong Bai[2], Wei Zhang[2], Ting Yao[2], *Senior Member, IEEE*,
Basem Shihada[1], *Senior Member, IEEE*, & Tao Mei[2], *Fellow, IEEE*
[1]King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia
[2]JD Explore Academy, Beijing, China
{guoqing.ma, basem.shihada}@kaust.edu.sa, ylbai@outlook.com
{wzhang.cu, tingyao.ustc}@gmail.com, tmei@live.com

***Abstract*—Pretraining large models on generous multi-modal corpora has accelerated the development of visual-linguistic (VL) representation and achieved great success on various vision-and-language downstream tasks. Learning these models is usually executed by predicting the randomly masked words of captions or patches in images. Such approaches, nevertheless, seldom explore the supervision of causalities behind the caption descriptions or the procedure of generating events beyond still images. In this work, we endow the pretrained models with high-level cognition by delving into dynamic contexts to model the visual and linguistic causalities uniformly. Specifically, we format the *dynamic contexts* of an image as the sentences describing the events *before*, *on*, and *after* image. Unlike traditional caption-wise similarity, we propose a novel dynamic contexts-based similarity (DCS) metric, in which the correlation of potential causes and effects besides immediate visual content are considered to measure the relevance among images. DCS can be further simplified by parameterizing event continuity to relax the requirements on dense contextual event annotations. A new pre-task is designed to minimize the feature distances of dynamically contextual relevant images and incorporate the event causality and commonsense knowledge into the VL representation learning. Models based on our dynamic contexts significantly outperform typical VL models on multiple cross-modal downstream tasks, including the conventional visual commonsense reasoning (VCR), visual question answering (VQA), zero-shot image-text retrieval, and extended image / event ordering tasks.**

***Index Terms*—Vision-language Pretraining, Event Reasoning, Dynamic Contexts**

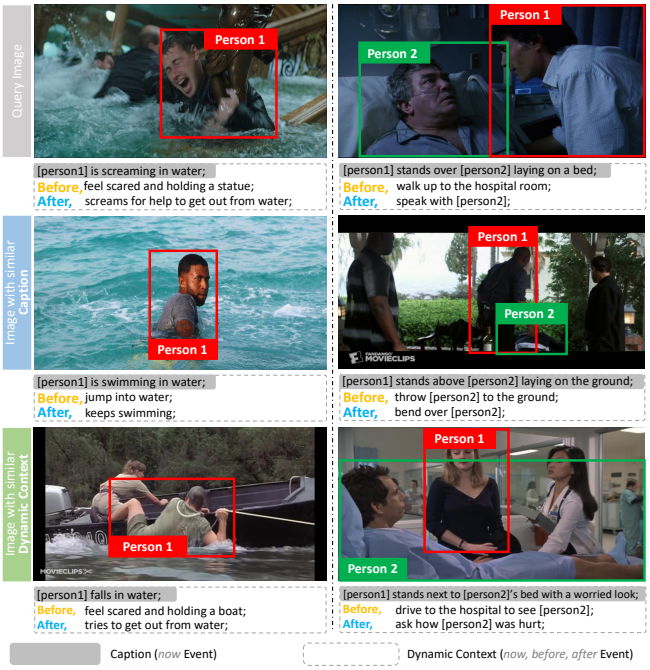

Fig. 1. Images in the first and second rows share similar captions (*now* events) but different *before* and *after* events, while images in the first and third rows have similar dynamic contexts. Dynamic context beyond ⟨image, caption⟩ carries rich causal knowledge of events.

I. Introduction

Motivated by the remarkable performance of large-scale language modeling [1], [2], [3] on target NLP tasks, researchers found that pretrained models on large-scale image-text datasets lead to superior performance over direct training on downstream visual-linguistic (VL) tasks. Existing VL pretrained methods [4], [5], [6], [7], [8], [9] adopt the transformer [10] as the backbone and extend the model to learn the joint representation through visual grounding tasks on large image-text datasets. By mimicking the objective functions in large-scale language modelings such as the mask word modeling and next sentence prediction, the objective functions of these pretrained Visual-Linguistic (VL) networks focus on similar targets, e.g., predicting masked words in captions, reconstructing masked patches in images, and image-caption matching. Nevertheless, still image frames with caption annotations, as shown in Fig. 1, mainly target describing the immediate visual content (*now* events, a.k.a. captions in most images) of the image, and suffer from a short understanding of dynamic situations captured in the image. In contrast, with commonsense knowledge and causality reasoning ability, people can easily reason out rich information about the story underlying the visual scene that goes far beyond the frame of the image and caption descriptions. The causality of event development beyond current visual scenes plays a crucial role in visual commonsense reasoning. It means correlating the dynamic story of *before*, *now*, *after* and the visual content together might bring cognition-level supervision for visual-linguistic representation learning.

Existing VL representation learning models supervised by cross-modal alignment and single-modal cloze testing could learn the correlations between image and caption. Such models

might lead to recognition-level understanding but are still far from cognition-level understanding. Intuitively, the cognitive capability is reflected by whether the learned representations can distinguish images, as human beings, in accordance with the dynamic contexts. From this viewpoint, pretext tasks on low-level text tokens and image patches lack the supervision of dynamic contexts, and are hard to characterize the subtle differences across images. As shown in Fig. 1, similar captions with slightly different wordings might respond to totally different contexts. Meanwhile, the causality of events can be affected by various factors, resulting in non-uniqueness in many cases. As such, directly measuring the dynamic contextual relevance of two given images based on sentence-level similarity would be dogmatic. As depicted in Fig. 2, there can be more than one reasonable assumption and logical inference regarding the cause and effect of the immediate visual content.

To alleviate these issues, we devise a novel learning-based metric concept, namely dynamic contexts-based similarity (DCS), to measure the similarity between any image pairs. To align the data format requirements of existing VL pretraining methods and maintain the consecutiveness of dynamic contexts, we implicitly model the before and after events as a distribution conditioned on the representation of the events as the image presents. Technically, a reasoning model is trained to extract the relations across consecutive events. A brief mathematical equation is then constructed to compute DCS in terms of events happening now while considering multiple potential event developments. Furthermore, we propose a new pretext task for VL pretraining models, aiming to incorporate event causality and commonsense knowledge into the VL representation learning. Specifically, we construct relevant image pairs based on the pre-calculated DCS, and an in-batch contrastive learning framework is applied to the image pairs. Such a pre-task is readily applicable to any VL pretraining methods. Extensive experiments have demonstrated that our proposal of dynamic contexts empowered models leads to a significant performance boost across different visual reasoning downstream tasks. Visualizations on downstream tasks and ablation studies illustrate that enhanced models boost performance by distinguishing the images with different dynamic contexts. We conclude our contributions as follows:

- We propose a novel learning-based criteria (DCS) to measure the similarity of images according to the continuity of contextual events beyond still images.
- Based on the quantification of dynamic contexts relevance, we design a new pretext task to incorporate the event causality and commonsense knowledge to large-scale vision and language pretraining.
- Experimental results on a wide range of visual-linguistic downstream tasks demonstrate the effectiveness of our proposed criteria. Extensive experiments, qualitative analysis, and ablation studies also prove that our newly designed pre-task captures cognitive understanding beyond the still image and can be feasible to be extended to several novel applications, such as image / event ordering tasks.

II. Related Work

A. Multimodal Pretraining

The idea of pretraining transformer-based networks [10] on large raw datasets and transferring on downstream tasks has been spread into the vision-linguistic area. Co-attention [11], [12] is straightforward to align the representation of different modalities to suit the format of vision-linguistic transformer based models. Authors in [12] build a modular network architecture with co-attentions, which consist of two parallel streams for visual and linguistic processing that interact through co-attention transformer layers, and they achieved new state-of-the-arts on the task of VQA. ViLBERT [6] adopts the similar idea and allows its structure for variable depths for each modality and enables sparse interaction through co-attention. They build pretraining tasks on multiple V+L datasets with three loss functions: predicting the masked words given the visual clues, reconstructing the masked image regions given the language clues, and predicting whether the text describes the image content. VL-BERT [5] simplifies the fusion of multi-modality and proposes an effective single-stream architecture, which uses BERT for text feature extraction and Faster-RCNN for image feature extraction. Furthermore, it takes a transformer encoder as its backbone network with the concatenated text and image features as its training input tokens. While adopting the same single-stream architecture, UNITER [8] proposes a new pretraining loss function with optimal transportation, whose target is to optimize the matching loss of word and corresponding image region pair. Recent works [13], [14], [15], [16], [17], [18] on multi-modal pretraining adopt similar single-stream or parallel-streams architectures with transformers as the backbone networks despite the downstream tasks or training schemes.

B. Knowledge Enhanced VL Representations

Besides imitating the network architecture and loss functions of large language modeling, researchers find the performance of the pretrained network benefits a lot from structured knowledge [19], [20]. ERNIE-ViL [7] enhances VL representations by constructing structured knowledge of visual scenes, including the present objects, attributes of objects, and relationships between objects. ERNIE-ViL uses a two-stream cross-modal transformers network to model the joint VL representations and construct object prediction, attribute prediction, and relationship prediction tasks to learn cross-modal detailed semantics alignments. Another direction is to integrate outside knowledge in VL representation learning to enhance the performance on in-domain datasets [21] or deal with the *open knowledge* challenge [22], such as the OK-VQA [23]. SOHO [21] learns to extract image features through a visual dictionary, which is a moving-averaged encoder to group visual pixels with similar visual semantics on the out-of-domain datasets and can be further updated on-the-fly during pretraining. KRISP [22] utilizes a symbolic knowledge graph as well as the implicit knowledge learned from large-scale BERT training to answer the questions focusing on knowledge that is not tied to a specific knowledge base. Since knowledge is organized in a format of graph, graph neural networks [24] are also applied to learn explainable multi-modal models

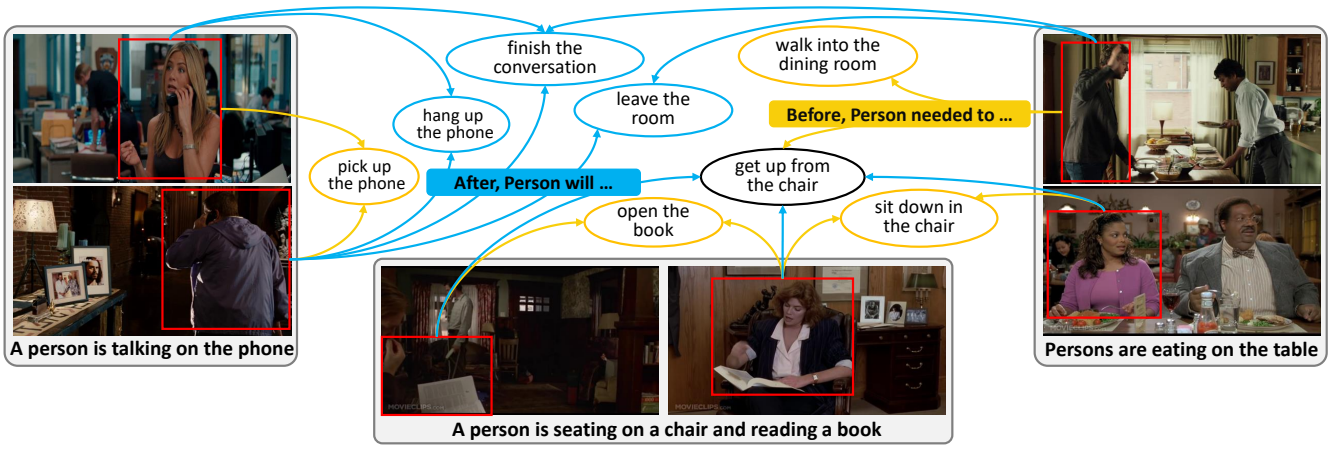

Fig. 2. A sample of dynamic contexts is illustrated with three pairs of images, where each pair shares a similar caption. 1) Each pair of images may have multiple events that happened shortly before (before events, in yellow) or will happen after (after events, in blue). 2) Each pair of images share similar *before* or *after* events. 3) Images with dissimilar captions might also share similar *before* or *after* events.

[25], [26]. They generate path representations by composing the structural, linguistics, and visual information of entities from multi-modal knowledge graphs. Thereafter, the rationale of visual-linguistic interactions can be inferred by leveraging the sequential dependencies within a path. However, existing methods to insert knowledge in VL models during pretraining helps little to understand the rationality, because the sub-task of masked language modeling is powerful enough to understand the token level knowledge [27].

III. Approach

Our approach aims to embed the dynamic context knowledge into existing VL models while aligning its format with the popular datasets utilized in existing methods. We generate dynamic context-based similarities to measure the distances of any image pairs in terms of dynamic contexts with the knowledge of still information. The solution decouples the dynamic contexts with the still information by constructing positive image-caption sample pairs based on the pre-calculated DCS for contrastive learning and maintaining the input format compatibility with existing VL models.

A. Dynamic Contexts based Similarity

Given a still image I_i, its dynamic contexts are defined as a combination of three kinds of event descriptions $[E^i_{before}, E^i_{now}, E^i_{after}]$, where the elements $[e^i_b, e^i_n, e^i_a] \in [E^i_{before}, E^i_{now}, E^i_{after}]$ denotes the sentence description of the event happened "shortly before", "on", and "shortly after" the image taken time, as shown in Fig. 2. With a well-trained sentence embedding model, the representations of these event descriptions can be encoded as feature vectors $[\epsilon^i_b, \epsilon^i_n, \epsilon^i_a]$, respectively. Assume that $\epsilon \in R^n$ is the normalized sentence embedding for each event sentence, and the dot-product function is applied for measuring the similarity of any two sentences embedding [28].

Therefore, for any two images I_i and I_j, given their dynamic context instances of $[e^i_b, e^i_n, e^i_a]$ and $[e^j_b, e^j_n, e^j_a]$, a naive dynamic context-based similarity of these two images can be measured as:

$$\hat{s}_{i,j} = \epsilon_b{}^i \cdot \epsilon_b^{j\,T} + \epsilon^i_n \cdot \epsilon_n^{j\,T} + \epsilon^i_a \cdot \epsilon_a^{j\,T}. \tag{1}$$

Here we ignore the normalization procedure (dividing by 3) for simplification. The representation of $\hat{s}_{ij}$ has the properties of symmetrical, bounded and transitivity, which well represents the similarity of images I_i and I_j. However, there are two main obvious limitations of directly incorporating the supervisions of $\hat{s}_{ij}$ for practical model training:

- *Required dense event annotations.* Almost all advanced VL multimodal models are pretrained on the large-scale corpora [1]. Constructing the dense annotations of dynamic context for these corpora is laborious and expensive and also computationally heavy.
- *Neglected event causality.* Similarity metric $\hat{s}_{ij}$ treats the dynamic contexts independent, hence, it lacks the capture of events causality, which plays an essential role in downstream tasks, such as commonsense reasoning.

To address these two issues, we look back the properties of dynamic contexts for images. Considering a function $f : R^n, R^n \rightarrow R$ as the causality measurement of two vectors, then we have constraints on the function f as follows according to the intuition on causalities:

- *Asymmetry:* $f(w,v) \neq f(v,w)$, $\forall w, v \in R^n$; the causality measurement can distinguish the order of two events, so that the function should be asymmetrical.
- *Continuity:* $w_1 \cdot w_2{}^T \approx 1, v_1 \cdot v_2{}^T \approx 1, f(w_1, v_1) \approx 1 \implies f(w_2, v_2) \approx 1$;

Considering the constraints of causality measurement mentioned above and the sententious structure of equation (1), we apply a linear mapping function as an encoder on normalized events embedding, and use the cosine similarity on the mappings of two consecutive events to *measure* the causalities, represented as follows,

$$\begin{aligned} f_{-1,+1}(\epsilon_{-1}, \epsilon_{+1}) &= \cos(\epsilon_{-1} \cdot W_{-1}; \epsilon_{+1} \cdot W_{+1}) \\ &\doteq \epsilon_{-1} \cdot W \cdot \epsilon^T_{+1}, \end{aligned} \tag{2}$$

where $\cos(x;y) = \frac{x \cdot y^T}{\|x\|_2 \|y\|_2}$ measures the cosine similarity of two vectors x and y; W_{-1} and W_{+1} are parameters of two linear mapping functions on two types of events; (ϵ_{-1} and ϵ_{+1}) denote any two consecutive events embedding, such as (ϵ_b and ϵ_n) or (ϵ_n and ϵ_a); For simplifying the training process, we relax the linear mapping process on two events as a trainable bi-linear function, as long as W is not a symmetric matrix. We can clearly observe the properties of asymmetry and continuity with the definition of $f_{-1,+1}$.

As $f_{-1,+1}$ is a scalar value, and replace $\epsilon_{-1}, \epsilon_{+1}$ with ϵ_b, ϵ_n, we have as follows,

$$\begin{aligned} {\epsilon_b}^i \cdot {\epsilon_b^j}^T &= \frac{1}{f^i_{b,n}} \cdot {f^i_{b,n}}^T \cdot {\epsilon_b}^i \cdot {\epsilon_b^j}^T \cdot f^j_{b,n} \cdot \frac{1}{f^j_{b,n}} \\ &= \frac{1}{f^i_{b,n} f^j_{b,n}} \cdot {\epsilon_n}^i \cdot W^T \cdot {\epsilon_b^i}^T \cdot {\epsilon_b}^i \cdot {\epsilon_b^j}^T \cdot {\epsilon_b}^j \cdot W \cdot {\epsilon_n^j}^T \\ &= \left(f^i_{b,n} \cdot f^j_{b,n}\right)^{-1} \cdot {\epsilon_n}^i \cdot W^T \cdot W \cdot {\epsilon_n^j}^T. \end{aligned} \tag{3}$$

Similarly, by replacing $\epsilon_{-1}, \epsilon_{+1}$ with ϵ_n, ϵ_a, we have:

$$\begin{aligned} {\epsilon_a}^i \cdot {\epsilon_a^j}^T &= \frac{1}{f^i_{n,a}} \cdot f^i_{n,a} \cdot {\epsilon_a}^i \cdot {\epsilon_a^j}^T \cdot {f^j_{n,a}}^T \cdot \frac{1}{f^j_{n,a}} \\ &= \left(f^i_{n,a} \cdot f^j_{n,a}\right)^{-1} \cdot {\epsilon_n}^i \cdot W \cdot W^T \cdot {\epsilon_n^j}^T \end{aligned} \tag{4}$$

Given the pretrained W, $f_{b,n}$ and $f_{n,a}$, by plugging equations (4) and (3) into equation (1), we obtain the same similarity value of $\hat{s}_{i,j}$ on a specific dynamic contexts instance. Therefore, we can neglect any instances of ϵ_b and ϵ_a by ϵ_n with the pretrained parameters.

$$\begin{aligned} \hat{s}_{i,j} =& \hat{\delta} \cdot \cos\left(\epsilon^i_n \cdot W^T; \epsilon^j_n \cdot W^T\right) + \cos\left(\epsilon^i_n; \epsilon^j_n\right) \\ &+ \hat{\delta}_T \cdot \cos\left(\epsilon^i_n \cdot W; \epsilon^j_n \cdot W\right), \end{aligned} \tag{5}$$

where

$$\hat{\delta} = \frac{\left\|\epsilon^i_n \cdot W^T\right\|_2 \left\|\epsilon^j_n \cdot W^T\right\|_2}{f^i_{b,n} \cdot f^j_{b,n}}$$
$$\hat{\delta}_T = \frac{\left\|\epsilon^i_n \cdot W\right\|_2 \left\|\epsilon^j_n \cdot W\right\|_2}{f^i_{n,a} \cdot f^j_{n,a}}.$$

As we presented in equation (5), the dynamic context-based similarity $\hat{s}_{i,j}$ can be calculated with *now* events and the coefficients of $\hat{\delta}$ and $\hat{\delta}_T$. However, as Fig. 2 presents, many images have multiple possible *before* and *after* events, which indicates that coefficients of $\hat{\delta}$ and $\hat{\delta}_T$ are random variables in terms of *before*, *now*, and *after* events. Thereby, instead of considering an instance of $f_{b,n}$ and $f_{n,a}$ from specific a (*before*, *now*, *after*) event combination, we use a statistical value to represent the relationships among event spaces of $\mathcal{E}_b$, $\mathcal{E}_n$ and $\mathcal{E}_a$ to avoid the risk of noise and outlier annotations. Hereby, we obtain our final DCS criteria s_{ij} as follows,

$$\begin{aligned} s_{i,j} =& \delta \cdot \cos\left(\epsilon^i_n \cdot W^T; \epsilon^j_n \cdot W^T\right) + \cos\left(\epsilon^i_n; \epsilon^j_n\right) \\ &+ \delta_T \cdot \cos\left(\epsilon^i_n \cdot W; \epsilon^j_n \cdot W\right) \end{aligned} \tag{6}$$

where

$$\delta = E[\hat{\delta}] = E\left[\frac{\left\|\epsilon^i_n \cdot W^T\right\|_2}{f^i_{b,n}}\right]^2 \approx \left[\frac{1}{|\mathbf{I}|}\sum_{i \in \mathbf{I}} \frac{\left\|\epsilon^i_n \cdot W^T\right\|_2}{f^i_{b,n}}\right]^2,$$
$$\delta_T = E[\hat{\delta}_T] = E\left[\frac{\left\|\epsilon^i_n \cdot W\right\|_2}{f^i_{n,a}}\right]^2 \approx \left[\frac{1}{|\mathbf{I}|}\sum_{i \in \mathbf{I}} \frac{\left\|\epsilon^i_n \cdot W\right\|_2}{f^i_{n,a}}\right]^2,$$

where δ and δ_T are two statistical constants depending on the training dataset and can be calculated with the optimized parameter W with respect to the image samples. The second operation for the above equations satisfies due to the independently selected image samples i and j. In practice, because of $E[\hat{\delta}] = E\left[E[\hat{\delta}|\epsilon^i_n]\right]$, we can calculate the value of $\frac{\left\|\epsilon^i_n \cdot W^T\right\|_2}{f^i_{b,n}}$ for each *before* event given the *now* event ϵ^i_n of image i. We can do the similar calculation for $E[\hat{\delta}_T]$. We apply the average values over the training dataset $\mathbf{I}$ of image samples to estimate the expectations of δ and δ_T. We present the probability density function of the $\frac{\left\|\epsilon^i_n \cdot W^T\right\|_2}{f^i_{b,n}}$ and $\frac{\left\|\epsilon^i_n \cdot W\right\|_2}{f^i_{n,a}}$ with the optimized W in Fig. 3 for a brief overview.

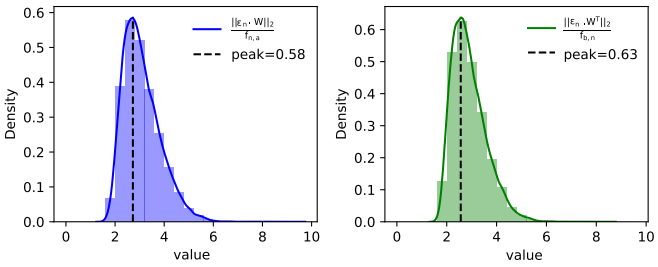

Fig. 3. Probability density function of $\frac{\left\|\epsilon^i_n \cdot W\right\|_2}{f^i_{n,a}}$ and $\frac{\left\|\epsilon^i_n \cdot W^T\right\|_2}{f^i_{b,n}}$ with the optimized W on VisualCOMET dataset.

Eventually, with the event causality measurement function f trained on datasets with human labored dynamic context annotations, such as VisualCOMET [29], the requirement of dense event annotations of *before* and *after* events would be no longer necessary, and the dynamic contexts relevance of image pairs can be smoothly parameterized by W.

B. Learning DCS on VisualCOMET

We extract the event descriptions for each image from the supervised VisualCOMET dataset in the format of $\langle E_{before}, E_{now}\rangle$ pair and $\langle E_{now}, E_{after}\rangle$ pair to be the positive training samples. Every event sentence is embedded into a normalized vector by the pretrained universal sentence encoder (USE) [30] before training the bi-linear model on the event pairs. Any other unpaired events in a training batch are treated as negative samples for the event causality learning. The in-batch contrastive learning loss function is used for optimization by maximizing the positive pair scores and minimizing the negative pair scores.

We split extracted event pairs from the training and validation annotation files of VisualCOMET as training and validation sets, respectively. The model has trained 400 epochs with a batch size of 512 and a learning rate of 1e-5 with a

step decay rate of 0.1 every 100 epochs. With the optimized bi-linear causality model, we can calculate the DCS for any two images to measure their dynamic contexts-based similarity by given image captions.

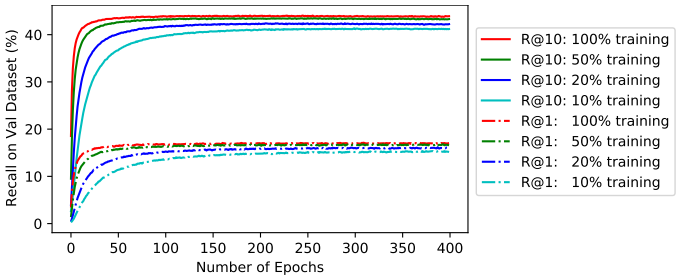

Fig. 4. Recall accuracy on the validation dataset with varied percentages of training dataset used.

Training with Varied Dataset Sizes Pre-text tasks usually need a large dataset for supervision, which requires a heavy labor force to annotate and limits the practical scope. However, a vast dataset for learning DCS might be unnecessary. We apply the pretrained USE model for sentence embedding, making the pre-embedding of sentences out of our consideration. Therefore, the dataset size should only be considerable compared to the applied bi-linear model size to avoid over-fitting during training. Meanwhile, the bi-linear model size depends on the embedding dimension of the USE model. Hence the required dataset size depends on the sentence embedding dimension. To visualize the effect of applying varied dataset sizes, we present the recall accuracy with different ratios of the total training dataset samples in Fig. 4, where the solid and dash-dot lines represent the results of $recall@10$ and $recall@1$.

Analysis The relationship between our defined DCS and the caption-based similarity of events is essential to excavate. To quantify the relations, we find the top-k closest images for each image from the training set of VisualCOMET with both similarity metrics and gather the statistics of overlaps from the top-k selected images. For the clearance of visualization, we normalize the overlapping ratio by dividing the counts of the overlapped image by the dataset size. We present the overlapping ratio statistics for k from 1 to 10 and the summation of overlapped ratio at the bottom in Fig. 5.

As shown in Fig. 5, the summation of the overlapping ratio increases as the k value increases, and when the k value is greater than 5, the ratio is greater than 0.9. In other words, DCS is not counter-intuitive. A correctly and rationally defined DCS metric should measure the dynamic contexts of two images in similar at a high chance if they have similar captions. However, in a limited dataset, the similarity rank based on captions and dynamic contexts might differ. Fig. 6 presents a sample case to explain this phenomenon. The overlapping ratio for k equaling 1 is 0.3, which means the most similar image for any image in the dataset will be the same by the two similarity metrics at the probability of 30%. Therefore, we choose the top-1 similar images to form positive sample pairs to distinguish the effects of DCS with the caption-based similarity significantly.

C. Dynamic Contexts Enhanced Pretraining

The pre-calculated DCS for each paired image does not mean the similarity value of the image and text joint embedding generated by the VL model. Hence it cannot be directly applied as a supervised signal in pre-training, such as minimizing the distance between the similarity values of the joint embeddings and the DCS values in a mean-squared-error (MSE) way. Therefore, based on the pre-calculated DCS for each paired image, we bring a new pretext task named Dynamic Contexts Enhanced (DCE) VL pretraining. The task is designed following the intuition that images with similar dynamic contexts will generate similar representations by the VL models. In order to 1) keep consistency with existing network architectures and 2) insert high-level dynamic contexts knowledge during pretraining, we re-construct a new dataset from the original one with its entry being the format of image and caption $\langle I_i, e_n^i \rangle$. Hereby, during VL pretraining, we hide the high-level dynamic contexts knowledge and keep the still information visible to force models to learn the hidden information with the prior knowledge according to the pre-calculated DCS.

With this ambition, we construct the DCE contrastive learning loss function, allowing the VL models generated joint embeddings of samples with similar dynamic contexts to stay close and keep the embeddings in contrast far away. Contrastive learning requires positive sample pairs and negative sample pairs as supervised information. It requires multiple negative samples to achieve good performance and thus can be computationally expensive (e.g., [31]). In our case, we select the in-batch contrastive learning to balance the performance and computational efficiency. We rank the DCS value among all other images for each query image and select the top-1 image as its positive sample. The in-batch contrastive learning constructs a batch with randomly selected N pairs of similar images iteratively. Optimizing models in a batch iteration treats the N pairs of similar images as positive samples, and all other $N^2 - N$ possible paired combinations are negative samples.

Denote the processed joint representation of the DCE task for sample $\langle I_i, e_n^i \rangle$ as $\mathbf{z}_i$, then we obtain the DCE loss function with in-batch contrastive learning as follows,

$$\mathcal{L}_{\text{DCE}} = -\sum_{i=1}^{N} \log \frac{\exp\left(\cos\left(\mathbf{z}_i; \mathbf{z}_{\arg\max_j(s_{i,j})}\right)/\tau\right)}{\sum_{k=1}^{N} \mathbb{1}_{[k \neq \arg\max_j(s_{i,j})]} \exp\left(\cos\left(\mathbf{z}_i; \mathbf{z}_k\right)/\tau\right)}, \tag{7}$$

where N is the batch size, $\arg\max_j(s_{i,j})$ denotes the index of I_i's top-1 similar image according to the pre-calculated DCS, $\mathbb{1}_{[k \neq \arg\max_j(s_{i,j})]}$ represents the image index in a batch which is not I_i's top-1 similar, and τ is used here to smooth the distribution. The loss function mentioned above achieves the ambition by maximizing the cosine similarity of images similar in dynamic contexts and minimizing the cosine similarity of negative pairs.

IV. Experiments

Typical large-scale VL pretraining [7], [8], [32] apply multiple sub-tasks in parallel to force models learning a fusion of visual and language representations within the backbone

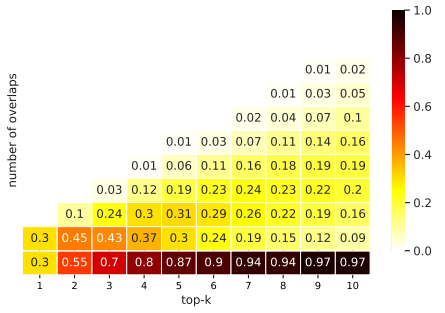

Fig. 5. Overlap ratios of Top-k images on two similarity metrics.

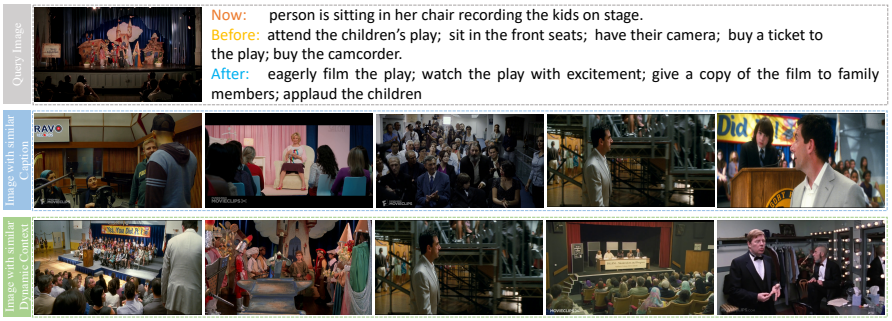

Fig. 6. A query image and its Top-5 similar images from left to right based on captions and dynamic contexts.

network and finetuned on the downstream tasks by transferring the sub-task heads accordingly. In the experiments, we apply the VL pretraining on the well-known benchmark network architectures with our proposed DCE head to force the network to learn high-level causal knowledge beyond the low-level text tokens and image patch representations. In the following subsections, we will go through the pretraining settings, visualize the performance of in-batch contrastive learning of the DCE head, evaluation on downstream tasks, ablation studies to excavate the influence of the DCE head, and visualizations of how the pretrained DCE models capture dynamic context relations on the out-of-domain dataset.

A. Pretraining Settings

We adopt VL-BERT [5] and UNITER [8] as the basic VL pretraining methods for our simulation experiments. Both VL-BERT and UNITER have a similar transformer-based network architecture and combine multiple sub-tasks for pretraining. In detail, VL-BERT states that the loss summation of masked language modeling (MLM), masked visual region construction (MVRC), and image-text matching (ITM) sub-tasks could be regarded as its pretraining objective function. Meanwhile, UNITER introduces an additional word-region alignment (WRA) sub-task, and it randomly shuffles a sub-task to optimize its parameters at each training step. Besides the training schemes, UNITER introduces more pretraining sub-tasks than VL-BERT, while introducing DCE head could be regarded as a further step on UNITER. Therefore, by comparing the basic VL-BERT, UNITER models, and their DCE versions, we can make a comprehensive analysis about the robustness of DCE head in different training schemes and how it influences model performances among other pretraining sub-tasks.

Implementation details During the experiments, we keep both their network backbones and training hyper-parameters as they reported in their works and apply the open-sourced codes [33], [34] to train our enhanced models. Both VL-BERT and UNITER have experimented on small and large size network architectures, we focus on the performance comparisons on large-size models in this paper. To be compatible with the conventional design idea, we construct our newly proposed DCE pre-task as an independent head attached after the backbone network with a two-layer perception module and a GELU

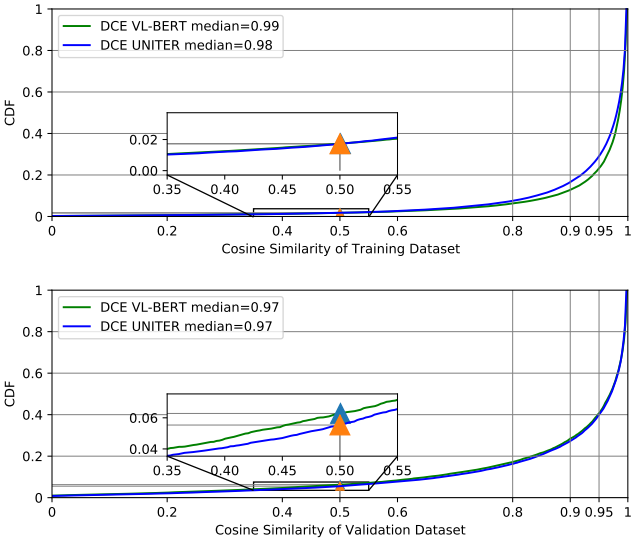

Fig. 7. The CDF of the cosine similarities for dynamic context-based similar image pairs whose representations are generated by the DCE pre-task head on VisualCOMET training and validation splits.

layer [35] as the activation function. We composite image pairs according to the DCS values in a batch and perform DCE head for enhancing the original VL pretraining. Meanwhile, the input entry is still in the format of $\langle image, now\ event \rangle$ to make our DCE head compatible with most existing VL pretraining methods. Noting that the DCE pre-task only takes effect on the data samples of $\langle image, now\ event \rangle$ extracted from the original VisualCOMET dataset during pretraining if without further annotations.

Pretraining datasets We follow the same settings reported in the work of Su *et al.* [5] to train our DCE VL-BERT on the datasets of Conceptual Captions [36], BooksCorpus [37], and VisualCOMET [29], with the sub-tasks of masked language modeling with visual clues, masked RoI classification with linguistic clues, and dynamic contexts enhanced in-batch contrastive learning. Following the same settings reported in [8], DCE UNITER is pretrained on datasets of COCO [38], Visual Genome [39], Conceptual Captions [40], SBU Captions [41], and VisualCOMET [29], with the sub-tasks of masked language modeling, image-text matching, word-region alignment, masked region modeling and dynamic

contexts enhanced in-batch contrastive learning.

B. In-batch Contrastive Learning Performance Visualization

We use cosine similarities of dynamic contexts-based similar image pairs to visualize their feature distances. For each image in the dataset and its pre-calculated DCS-based top-1 similar image pair, we generate the DCE pre-task representations from the pretrained DCE models and calculate the cosine similarity on the paired representations. Fig. 7 presents the cumulative distribution function (CDF) of the cosine similarities for both the training and validation splits of the VisualCOMET dataset. As both the training and validation CDF curves imply, most dynamic context-based similar image pairs have cosine similarity values greater than 0.9, which means the distances of positive image pairs (top-1 similar in terms of dynamic contexts) are minimized as expected. Meanwhile, the CDF values for cosine similarities of both pretrained DCE models lesser than 0.5 is under 2% and 6% for the training and validation splits, meaning almost all positive image pairs are correctly classified as positive in the in-batch contrastive pre-task. The experimental results of pretraining on our DCE models prove that the feature distances of dynamic contexts relevant images are minimized as expected with in-batch contrastive learning, and the performance on the validation split proves the generalization of the new proposed pre-task head on the pretraining models.

C. Evaluation on Downstream Tasks

We verify the effectiveness of incorporating dynamic contexts information to VL representation learning on several typical visual-linguistic downstream tasks, e.g., Visual Commonsense Reasoning (VCR) [42], Visual Question Answering (VQA) [43], and zero-shot image-text retrieval [8].

1) Finetuning on VCR and VQA: We evaluate the performance of VCR and VQA by transferring the pretrained models to each downstream task and finetuning them through end-to-end training on the target datasets. Table I presents the results of DCE models achieving better performance for both network architectures in both downstream tasks than the benchmark models [1]. In conclusion, the DCE VL-BERT has 1.04% and 0.80% accuracy improvements on the VQA test-dev and test-std tasks compared to the baseline VL-BERT. The DCE UNITER has 0.36% and 0.30% accuracy improvements on the VQA test-dev and test-std tasks compared to the baseline UNITER. On the task of Q$\longrightarrow$AR, both DCE models have 0.6% accuracy improvement over the baseline models. Since DCE UNITER's performance improvements are not as significant as that of DCE VL-BERT, to question the statistical significance, we make four trials with different model weights initialization for DCE UNITER pretraining and finetuning the pretrained model on VQA and VCR tasks with the same settings as reported. We report the average values and the maximal absolute deviations of the four trials from the average values.

[1]Both works lack the results of "other", "number" and "yes/no" of VQA, so we report them by training on their public code bases.

Both VCR and VQA require reasoning on the knowledge of image-question. Still, the VCR dataset is more complicated of stories than that of VQA, which in nature requires cognition understanding and commonsense reasoning about the world [42]. Our DCE sub-task constrains the model to learn images-question semantics by forcing dynamic contexts similar images to stay together in the output embedding space. Hence, it performs better on the VCR tasks requiring higher-order cognition than the VQA tasks. Moreover, despite the different training schemes of both benchmark networks, DCE UNTER applies five sub-tasks for pretraining, while Vl-BERT only has three. Therefore, the effectiveness of the DCE sub-task for UNITER during pretraining is diluted by other sub-tasks compared to VL-BERT, which might make performance improvements on VL-BERT more significant than that on UNITER.

Fig. 8 presents a case study of the Q$\longrightarrow$AR task on the VCR validation dataset, where we select the first four cases such that both DCE models make correct choices while base models make wrong choices. None of the selected questions can be directly answered without speculation on the dynamic contexts of images. For instance, the first two questions are equivalent to inference on what will happen next with the images presenting. The last two questions require understanding the background contexts and inference on what has happened before. If training models without speculation, the answers to the third question like "taking out a weapon" or "looking for money" could also be correct. Meanwhile, the answer like "10 has some type of college papers he is shoving in 2's face," which describes the image, could also be a suitable choice. In conclusion, pretraining with dynamic contexts helps downstream tasks better capture higher-order cognition and commonsense reasoning about the world.

2) Zero-Shot Image-Text Retrieval: The evaluations on zero-shot image-text retrieval are more illustrative. Different from the standard image-text matching [45], [46], the influence factors of finetuning are neglected, as no images and captions have been probed during training [8]. Hence it can directly explain the performance of the pretrained models on downstream tasks. We transfer the parameters of the image-text matching sub-task to generate the retrieval score of any image-text pair. Because the reported VL-BERT is not pretrained with the ITM sub-task, we only report the results on the UNITER network.

Table II presents the experimental results of zero-shot image retrieval (ZS IR) and text retrieval (ZS TR) on Flickr test set, following the dataset split and experimental settings in UNITER [47]. It can be found that our DCE UNITER performs better than the base model in terms of different recall and precision for both image and text retrieval. Moreover, the improvements of IR (0.80%) and TR (1.40%) on R@1 are more significant than that on R@5 (IR: 0.20%, TR: 0.30%) and R@10 (IR: 0.04%, TR: 0.30%), proving that DCE UNITER can understand the semantics of image and text more deeply than the base model with dynamic contexts knowledge, leading to better correlation modeling between image and text with more accurate relationship capturing.

Fig. 9 presents several cases of the text and image retrieval tasks. We present the query captions/images and the correct re-

TABLE I

PERFORMANCE COMPARISON TO THE BASELINE PRETRAINING METHODS ON TASKS OF VCR AND VQA. NOTING THAT VQA FOR VL-BERT AND DCE VL-BERT ARE FINETUNED ON THE TRAIN+VAL SPLITS, WHILE UNITER AND DCE UNITER ARE FINETUNED ON TRAIN+VAL+VG SPLITS.

Model	VQA test-dev				VQA test-std				VCR		
	other	number	yes/no	overall	other	number	yes/no	overall	Q→A	QA→R	Q→AR
ViLBERT [6]	-	-	-	70.55	-	-	-	70.92	73.3	74.6	54.8
VisualBERT [4]	-	-	-	70.80	-	-	-	71.00	71.6	73.2	52.4
LXMERT [11]	-	-	-	72.42	63.10	54.20	88.20	72.54	-	-	-
Oscar [13]	-	-	-	73.61	-	-	-	73.82	-	-	-
VL-BERT	61.00	**53.66**	87.03	70.89	61.37	**53.93**	87.46	71.38	75.8	78.4	59.7
DCE VL-BERT	**62.25**	53.41	**88.20**	**71.93**	**62.37**	53.21	**88.45**	**72.18**	**76.94**	**79.48**	**61.30**
UNITER	63.88	56.95	89.91	73.82	64.31	56.34	89.83	74.02	77.3	80.8	62.8
DCE UNITER	**64.31** ± 0.02	**57.39** ± 0.20	**90.18** ± 0.01	**74.18** ± 0.02	**64.47** ± 0.04	**56.65** ± 0.13	**90.28** ± 0.01	**74.32** ± 0.02	**77.57** ± 0.02	**81.35** ± 0.02	**63.48** ± 0.02

Image & Question	Answer Choices	Rationale Choices
 **Do you think 4 will sit down on 9 ?**	No she would walk around it; Yes , if she doesn' t dance , she will sit soon; **No , she won't;** Yes , 4 will put her glove back on , it is on the bench near 1.	4 is a young girl and young girls tend to be shy. she is waiting for a young man like 6 to ask her to dance; She does not have an instrument on stage; 9 is empty and nearby. She will tire of standing in dress shoes soon; **She looks happy and confident. Her dancing posture is on point.**
 **What would happen if 4 bought 6 a drink ?**	6 would be offended; **6 would accept the drink;** 6 would stop talking; 4 would start to talk with 6.	The lady at the top of the stairs is probably the homeowner , so 6 is the housekeeper and the housekeeper would offer guests a drink; **6 is sitting at a bar without a drink, usually means she will be ordering one;** The larger glass of milk is closest to 6, she will continue drinking it after she is done; Based on her mode of dress and position behind the bar, 6 is a bartender and therefore expected to continually offer and serve drinks until asked otherwise.
 **What is 0 grabbing from his pocket ?**	He's about to put it in his pocket; **It is a lottery ticket and he is checking it as the numbers are drawn on tv;** He is taking out a weapon; 0 is looking for the money he wants to deposit with 8.	**He is reaching into inside pocket to take something out;** He is looking off in the distance with a worried look; He has it raised behind his back prepared to strike if the cow ever does come near; 0 has a gun and he has his hand on his arm.
 **Does 10 have proof of 2' s interference in two thousand and sixteen point zero election ?**	No he does not; **No 10 took an oath before testifying that he would not lie when giving his testimony;** No , 2 is fulfilling his job function and did not benefit in the way 10 did; No , 10 has some type of college papers he is shoving in 2's face.	2 sits in a packed room that looks like a conference hall and 10 is in the back with a camcorder recording it; 2 has a textbook and a college sweatshirt on, and many of the girls are holding textbooks , so 10's papers are college related; 2 is at the head of the room and appears to be a star of some sort pontificating on the topic of fame; **2 appears shocked and nervous.**

Fig. 8. A case study of the Q →AR task on the VCR validation dataset. The correct answers and rationals to each question selected by DCE models are emphasized in bold, while the wrong choices selected by UNITER are labeled in gray.

TABLE II

RECALL PERFORMANCE COMPARISON TO THE BENCHMARK PRETRAINING METHODS OF ZERO-SHOT IMAGE AND TEXT RETRIEVAL ON FLICKR1K.

Model	ZS IR			ZS TR		
	R@1	R@5	R@10	R@1	R@5	R@10
ViLBERT [6]	31.86	61.12	72.80	-	-	-
Unicoder-VL [44]	48.40	76.00	85.20	64.30	85.80	92.30
Oscar [13]	57.50	82.80	89.80	73.50	92.20	96.00
UNITER	68.74	89.20	93.86	83.60	95.70	97.70
DCE UNITER	**69.54**	**89.40**	**93.90**	**85.00**	**96.00**	**98.00**

TABLE III

STATISTICS OF DATASETS FOR IMAGE (IQ) AND EVENT ORDERING (EO) TASKS.

	Train set			Validation set		
EO	*before*	*now*	*after*	*before*	*now*	*after*
	238K	467K	469K	59K	29K	59K
IO	# Videos			# Image pairs		
	3,105			24,346		

trieved images/captions by DCE UNITER for each task on the left two columns of each sub-figure. The base UNITER model retrieved images/captions, and the corresponding matched captions/images are presented on the right two columns. As shown in the figure, the wrongly retrieved images and captions by the benchmark UNITER model are somewhat similar to the ground truths. For instance, as the left sub-figure showing, both images on the first row present two women hugging each other, and the matched captions have similar semantics. However, the two images have different dynamic contexts. The right image presents a girl wearing a bike helmet and several people holding bikes in the background, so we can reason that the girl might obtain a good grade in a cycling race, and her friend embraces her for celebration, obviously different from the right image in terms of dynamic contexts. The base UNITER model pretrained with image-caption matching pretext task cannot classify the difference between two images, making an error retrieval on a similar caption semantic. A similar conclusion can be inferred from the left sub-figure. The query image in the first row presents a guy who rides on a motorcycle, illustrated by the retrieved captions. The matched image of the base UNITER retrieved caption presents similarly where a group of men ride motorcycles. However, we can reason that the people on the left might be in the parade as they all wear the same red and black outfits, while the guy in the right image might be a policeman managing traffic.

Fig. 9. A case study of the text and image retrieval tasks on the Flickr1k dataset. The first column of each sub-figure presents the query captions/images, and the second and third columns show the proposed DCE UNITER and UNITER retrieved images/captions. The fourth column gives the matched ground truth of the UNITER retrieved images/captions.

TABLE IV
ABLATION STUDIES ON DOWNSTREAM VCR AND VQA TASKS. WE REPORT THE ACCURACY OF VCR AND TEST-STD ACCURACY OF VQA.

Model	VCR		
	Q$\longrightarrow$A	QA$\longrightarrow$R	Q$\longrightarrow$AR
VL-BERT	75.8	78.4	59.7
VL-BERT$_{abl}$	75.7	78.9	60.1
VL-BERT$_{ood}$	**76.56**	**78.84**	**60.70**
UNITER	77.3	80.8	62.8
UNITER$_{abl}$	77.2	80.9	63.0

Model	VQA test-std			
	other	number	yes/no	overall
VL-BERT	61.4	53.9	87.5	71.4
VL-BERT$_{abl}$	61.8	53.5	87.3	71.4
VL-BERT$_{ood}$	**61.85**	53.83	**87.75**	**71.71**
UNITER	64.3	56.3	89.8	74.0
UNITER$_{abl}$	63.3	56.4	90.0	73.5

The proposed method enhanced with dynamic contexts can accurately distinguish two images with similar captions but different dynamic contexts, increasing the ability to retrieve images and captions more accurately.

D. Ablation Study

The experimental results in the above study prove that the newly proposed sub-task significantly improves the understanding of pretrained models on multiple downstream tasks. However, we introduced extra data samples $\langle image, now\ event \rangle$ extracted from the VisualCOMET dataset into pretraining compared with the baseline methods. Therefore, it is curious whether the additional data samples or the newly proposed sub-task causes increased performance during pretraining. Besides, we learned the DCS model from the VisualCOMET dataset and pretraining the DCE models on the data samples extracted from VisualCOMET. Hence, how the trained DCS models can be generalized on other $\langle image, caption \rangle$ datasets is also worth discovering.

1) Effects of Introducing New Samples in pretraining: We build an ablation study to explore the influence by removing the proposed DCE pre-task head. We pretrain the models of VL-BERT$_{abl}$ and UNITER$_{abl}$ based on the same hyper-parameters and datasets as in the previous section where the DCE models are pretrained except for the DCE pre-task is trimmed. We finetune the pretrained VL-BERT$_{abl}$ and UNITER$_{abl}$ on the downstream tasks of VCR and VQA, and Table IV presents the results. The results in Table IV illustrates that introducing extra data samples in pretraining has slight but not significant improvements for the VCR task on QA$\longrightarrow$R and Q$\longrightarrow$AR. Moreover, it has slightly worsened the VCR task's performance on Q$\longrightarrow$A and the overall result of UNITER on the VQA task. The results are surprising while not strange. The performance of downstream tasks depends on the relevance of pretraining datasets. Larger datasets usually lead to better performance, but not always, as announced by recent studies [27] of decoupling the datasets from VL pretraining. Thereby, discovering deep knowledge from the raw dataset is wiser, while naively increasing the scale of pretraining datasets cannot guarantee performance well everywhere.

2) DCS on Out-of-Domain Dataset: The DCE pre-task in the main experiments pretrained on the extracted data samples from the raw VisualCOMET dataset. At the same time, the DCS is also trained on the dynamic contexts of the same dataset. Hence, it is attractive to discover if the DCE pre-task works on other $\langle image, caption \rangle$ datasets with the DCS model parameters optimized on the VisualCOMET dataset. We apply the optimized bi-linear model W on the Conceptual Captions dataset to generate the DCS of each pair of images and use the DCS to generate positive samples for the DCE head in-batch contrastive learning. We then build an ablation study

Fig. 10. An ordered set of images and the corresponding annotated captions (events happening as showing).

TABLE V
RESULTS OF EVENT AND IMAGE ORDERING ON THE VISUALCOMET DATASET.

Model	EO acc	IO acc	EditDist
VL-BERT	82.27	87.35	0.34
DCE VL-BERT	**83.64**	**89.38**	**0.29**
UNITER	84.28	88.69	0.31
DCE UNITER	**85.15**	**91.41**	**0.22**

by applying the same datasets reported in the baseline VL-BERT model and the DCE pre-task works on the Conceptual Captions dataset for pretraining. The results of downstream tasks of VCR and VQA are presented in Table IV, denoted as VL-BERT$_{ood}$. As the table presented, VL-BERT$_{ood}$ has improvements of 0.8%, 0.4%, and 1.0% on the tasks of VCR and 0.33% on the VQA overall task than the baseline VL-BERT. The performance increments of VL-BERT$_{ood}$ prove that the optimized DCS model and the proposed DCE pre-task can be transferred to other $\langle image, caption \rangle$ datasets to excavate the causalities and commonsense behind still image-caption information.

E. Event and Image Ordering Tasks

To further understand the effects of DCE on visual-linguistic representations learning, and evaluate its performance on event reasoning, we construct two downstream tasks of event ordering (EO) and image ordering (IO) based on the VisualCOMET dataset to check if DCE models can capture the logic beyond the static image and caption. Each task is finetuned from the pre-trained models.

Specifically, for each image, we randomly sample an event from its corresponding dynamic contexts (before events, now events, after events) and compose the data in the format of $\langle image, event \rangle$. We construct a three-label classification task, where EO aims to classify the event into *before*, *now*, or *after*, given the image, event as input. We train on the training split of VisualCOMET and report the corresponding performance on the validation split. The joint representations for each image-event pair generated by the VL backbone networks are fed into a linear mapping layer followed by a cross-entropy loss function for optimization.

IO constructs a more challenging application to re-order the shuffled images sampled from the same video clips in the original VisualCOMET dataset. As images in the VisualCOMET dataset are selected from some video clips, we make a permutation of every two images in the same video clip and label the paired images according to their ground-truth ordering as false (0) or correct (1). We construct the IO training dataset with paired images and their corresponding *now* events. Noting that we select video clips which contain at least three shots to make the ordering task reasonable. Half of the randomly shuffled image pairs are selected to form the training dataset, while the left are used for validation. The VL backbone networks generated representations are fed into a bi-linear mapping layer attached with a binary cross-entropy loss function for optimization. We rebuild the images ordering sequence based on the predicted ordering for all image pairs generated from the same video clip. To verify the re-ordering performance, we use the shortest edit distance (EditDist) dividing by the length of the ground-truth ordering sequence and the classification accuracy on the whole dataset as metrics. We present the statistics of datasets for the EO and IO tasks in Table III. For all experiments, we choose AdamW [48] as the optimizer, and we set learning rates of 1e-5 and 1e-6 for IO and EO, with the batch size to be 40, respectively.

The results of both DCE and base models for event and image ordering tasks are shown in Table V. Fig. 10 presents a sample of the ordered images and the corresponding captions,

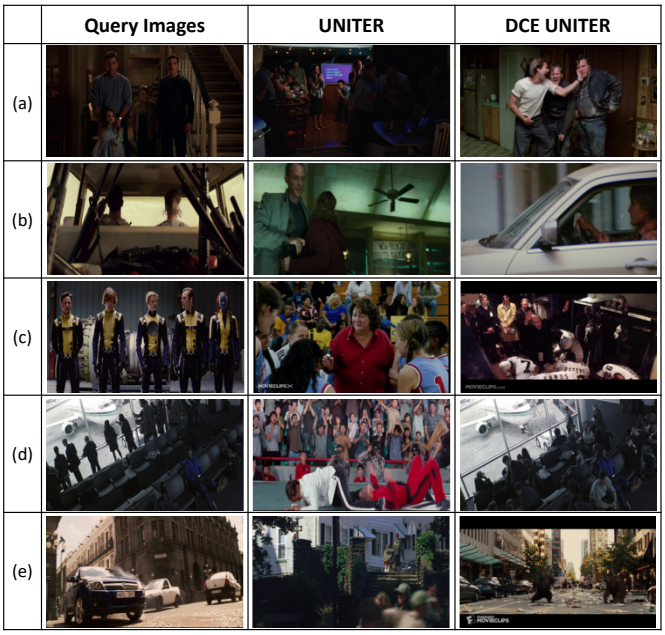

Fig. 11. Visualization of similar images in the VCR test dataset. Figures (a)-(e) present a group of persons, a car, a team, an airport, and a street, respectively.

where baseline models with patch-level information almost unable to distinguish these images. As Table V shows, pretrained DCE models perform better than baseline models in terms of EO and IO, evidencing that the DCE sub-task can better capture the cognition and help models to understand the logic beyond the visual-linguistic information from raw datasets.

F. Visualization of DCS in Out-of-Domains

The essential hypotheses are that we assume images contain wealthy dynamic contexts and similar dynamic contexts generate similar images. The above experimental results implicitly prove the hypotheses with outstanding downstream task performances compared with benchmarks. Further, we want to visualize the image similarities with the pretrained DCE models and show whether the models capture dynamic context relations in out-of-domains. Therefore, we conduct experiments on the VCR test dataset to ensure that no entries have been observed during pretraining and whose domain is similar to the VisualCOMET dataset.

In detail, we generate the representation for each image from the pretrained DCE UNITER by normalizing the output from the DCE head while masking the whole text tokens during experiments. To make a fair comparison, we generate the baseline representation by normalizing the pooled output from the backbone of UNITER. Fig. 11 presents the first five images in the test dataset of VCR and their closest images generated by DCE UNITER and base UNITER.

As shown in Fig. 11, the DCE UNITER can cluster similar images with the query images even if the images are not observed during pretraining. In contrast, the paired images of the base UNITER are not reasonably related to the query images in most cases. The out-of-domain results prove that pretrained models with the DCE sub-task can "understand" the dynamic contexts beyond images even without hints from captions.

V. Conclusion and Future Work

Embedding high-level knowledge from dynamic contexts endows the pretraining of large VL models with causal reasoning capability, and significantly improves the performance of downstream tasks. Particularly, this paper introduces a learning-based metric to measure the contexts similarity of the given image pairs. On this basis, we devise a contrastive learning sub-task to boost VL representation learning with dynamic contexts. Extensive experiments on reasoning related downstream tasks and a new image/event ordering task demonstrate the effectiveness of our mechanisms.

Possible future direction includes extracting the causality of events directly from visual content irrespective of the dependency of dense textual event annotations. Moreover, taking event causalities into account in video understanding task, e.g., video2caption [31], would yield more interesting applications, and further enhance VL representation learning.

References

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/n19-1423

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[3] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018.

[4] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[5] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SygXPaEYvH

[6] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[7] F. Yu, J. Tang, W. Yin, Y. Sun, H. Tian, H. Wu, and H. Wang, "Ernie-vil: Knowledge enhanced vision-language representations through scene graphs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3208–3216.

[8] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *European conference on computer vision*. Springer, 2020, pp. 104–120.

[9] T. Zhu, L. Li, J. Yang, S. Zhao, H. Liu, and J. Qian, "Multimodal sentiment analysis with image-text interaction network," *IEEE Transactions on Multimedia*, 2022.

[10] A. Waswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[11] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 5100–5111.

[12] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6281–6290.

[13] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *European Conference on Computer Vision*. Springer, 2020, pp. 121–137.

[14] H. Zhang, S. Qian, Q. Fang, and C. Xu, "Multi-modal meta multi-task learning for social media rumor detection," *IEEE Transactions on Multimedia*, vol. 24, pp. 1449–1459, 2021.

[15] H. Chen, D. Jiang, and H. Sahli, "Transformer encoder with multi-modal multi-head attention for continuous affect recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 4171–4183, 2020.

[16] L. Huang, Y. Liu, X. Zhou, A. You, M. Li, B. Wang, Y. Zhang, P. Pan, and X. Yinghui, "Once and for all: Self-supervised multi-modal co-training on one-billion videos at alibaba," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1148–1156.

[17] R. Hu and A. Singh, "Unit: Multimodal multitask learning with a unified transformer," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2021, pp. 1419–1429.

[18] Z. Ji, K. Chen, and H. Wang, "Step-wise hierarchical alignment network for image-text matching," in *IJCAI*, 2021, pp. 765–771.

[19] M.-A. Lachaux, B. Roziere, M. Szafraniec, and G. Lample, "Dobf: A deobfuscation pre-training objective for programming languages," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[20] C. Li, B. Bi, M. Yan, W. Wang, S. Huang, F. Huang, and L. Si, "Structurallm: Structural pre-training for form understanding," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 6309–6318.

[21] Z. Huang, Z. Zeng, Y. Huang, B. Liu, D. Fu, and J. Fu, "Seeing out of the box: End-to-end pre-training for vision-language representation

learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 976–12 985.

[22] K. Marino, X. Chen, D. Parikh, A. Gupta, and M. Rohrbach, "Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 111–14 121.

[23] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "Ok-vqa: A visual question answering benchmark requiring external knowledge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3195–3204.

[24] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation learning on graphs: Methods and applications," *arXiv preprint arXiv:1709.05584*, 2017.

[25] A. Holzinger, B. Malle, A. Saranti, and B. Pfeifer, "Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai," *Information Fusion*, vol. 71, pp. 28–37, 2021.

[26] Y. Zhang, S. Qian, Q. Fang, and C. Xu, "Multi-modal knowledge-aware hierarchical attention network for explainable medical question answering," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 1089–1097.

[27] L. A. Hendricks, J. Mellor, R. Schneider, J.-B. Alayrac, and A. Nematzadeh, "Decoupling the role of data, attention, and losses in multimodal transformers," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 570–585, 2021.

[28] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," *Advances in neural information processing systems*, vol. 28, 2015.

[29] J. S. Park, C. Bhagavatula, R. Mottaghi, A. Farhadi, and Y. Choi, "Visualcomet: Reasoning about the dynamic context of a still image," in *European Conference on Computer Vision*. Springer, 2020, pp. 508–524.

[30] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018.

[31] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[32] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5583–5594.

[33] Y. H. Weijie Su, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018. [Online]. Available: https://github.com/jackroos/VL-BERT

[34] L. Chen, "Uniter: Universal image-text representation learning," 2020. [Online]. Available: https://github.com/ChenRocks/UNITER

[35] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[36] E. G. Ng, B. Pang, P. Sharma, and R. Soricut, "Understanding guided image captioning performance across domains," in *Proceedings of the 25th Conference on Computational Natural Language Learning*, 2021, pp. 183–193.

[37] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 19–27.

[38] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[39] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.

[40] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2556–2565.

[41] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," *Advances in neural information processing systems*, vol. 24, 2011.

[42] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[43] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *International Conference on Computer Vision (ICCV)*, 2015.

[44] G. Li, N. Duan, Y. Fang, M. Gong, and D. Jiang, "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 336–11 344.

[45] H. Wang, Y. Zhang, Z. Ji, Y. Pang, and L. Ma, "Consensus-aware visual-semantic embedding for image-text matching," in *European Conference on Computer Vision*. Springer, 2020, pp. 18–34.

[46] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 201–216.

[47] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641–2649.

[48] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

Wei Zhang is now a Senior Researcher at JD.com, Beijing, China. He received his Ph.D degree from Department of Computer Science in City University of Hong Kong. His research interests include computer vision and multimedia, especially visual recognition and generation. He has won the Best Demo Awards in ACM MM 2021 ACM-HK Openday 2013, and served as the Area Chair for ICME, ICASSP, VCIP, and Technical Program Chair for ACM MM Asia 2023. He also served the guest editors for ACM TOMM and Advances in Multimedia.

Basem Shihada (Senior Member, IEEE) is an Associate & Founding Professor of Computer Science and Electrical & Computer Engineering programs in the Computer, Electrical and Mathematical Sciences Engineering (CEMSE) Division at King Abdullah University of Science and Technology (KAUST). His research covers a range of topics in energy and resource allocation in wired and wireless networks, software defined networking, cloud/edge computing, internet of things, data networks, and underwater networks. He obtained his PhD degree in Computer Science from the University of Waterloo. In 2009, he was appointed as visiting faculty in the Department of Computer Science, Stanford University.

Basem's work received several awards and honors, including two best paper awards in AFIN, and IEEE WiMob. He has co-authored more than 170 papers in the field of wireless networks. He serves as a Senior Editor for IEEE Communication Letters, Associate Editor for Frontiers in Communication and Networks, Editor of Photonic Network Journal and served track chair, organizing chair, publication chair, and technical program committee for several conferences such as VTC, InfoCom, Globecom, ICC, Mobiquitous, WCNC, and others. He is a senior member of IEEE and a professional member of ACM. His work on underwater networks, specifically Aqua-Fi, got the world attention and resulted into wide international media coverage.