Training-free Deep Concept Injection Enables Language Models for Video Question Answering

Xudong Lin[1], **Manling Li**[2], **Richard Zemel**[1], **Heng Ji**[2], **Shih-Fu Chang**[1]
[1]Columbia University [2]University of Illinois at Urbana-Champaign
xudong.lin@columbia.edu

Abstract

Recently, enabling pretrained language models (PLMs) to perform zero-shot crossmodal tasks such as video question answering has been extensively studied. A popular approach is to learn a projection network that projects visual features into the input text embedding space of a PLM, as well as feed-forward adaptation layers, with the weights of the PLM frozen. However, is it really necessary to learn such additional layers? In this paper, we make the first attempt to demonstrate that the PLM is able to perform zero-shot crossmodal tasks without any crossmodal pretraining, when the observed visual concepts are injected as both additional input text tokens and augmentation in the intermediate features within each feed-forward network for the PLM. Specifically, inputting observed visual concepts as text tokens helps to inject them through the self-attention layers in the PLM; to augment the intermediate features in a way that is compatible with the PLM, we propose to construct adaptation layers based on the intermediate representation of concepts (obtained by solely inputting them to the PLM). These two complementary injection mechanisms form the proposed Deep Concept Injection, which comprehensively enables the PLM to perceive instantly without crossmodal pretraining. Extensive empirical analysis on zero-shot video question answering, as well as visual question answering, shows Deep Concept Injection achieves competitive or even better results in both zero-shot and fine-tuning settings, compared to state-of-the-art methods that require crossmodal pretraining.

1 Introduction

Pretrained language models have been shown to be a powerful base model to deal with tasks beyond natural language processing, such as visual question answering (Lu et al., 2019; Dai et al., 2022) and video question answering (Sun et al., 2019; Li et al., 2020a; Lin et al., 2021; Yang et al., 2021, 2022b). These tasks require reasoning over information from multiple modalities. Thus, the key challenge is to find a common representation so that the information from different modalities can be fused and processed by the PLM. Conventional methods (Lu et al., 2019; Sun et al., 2019) usually rely on a two-stage training process to obtain satisfying results on downstream datasets. Assuming pretrained language models and feature extractors like vision-text contrastive models (e.g., CLIP (Radford et al., 2021)) are available, the first stage aims at crossmodal pretraining on web-collected vision-text dataset with techniques like masked token modeling (Li et al., 2020a; Zellers et al., 2021) or contrastive learning (Xu et al., 2021; Li et al., 2022; Yang et al., 2021) to learn the alignment and fusion of visual and textual inputs. In the second stage, the model is further fine-tuned with human annotation on specific downstream datasets (Antol et al., 2015; Yang et al., 2021; Yu

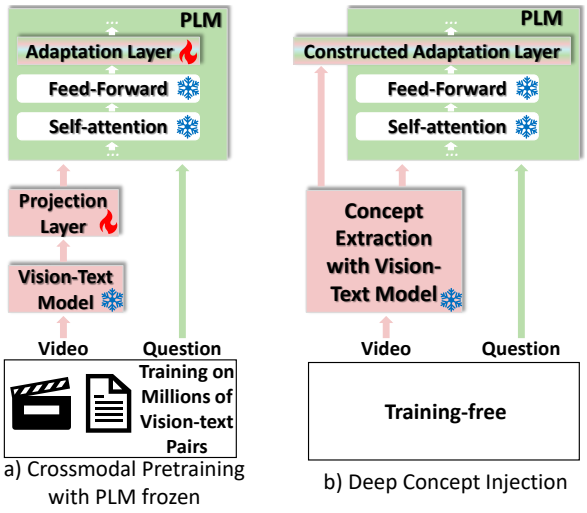


Figure 1: Unlike existing methods of crossmodal pretraining on millions of vision-text pairs, our Deep Concept Injection enables PLMs for zero-shot crossmodal tasks in a training-free manner. The core idea is to leverage concepts as the bridge to inject the visual information in the inference process of PLMs as both input and constructed adaptation layers.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 22399–22416
November 12-16, 2024

et al., 2019; Li et al., 2020a; Xu et al., 2017; Lei et al., 2018; Marino et al., 2019) to obtain better models for specific tasks.

However, such a two-stage training process has been criticized to be lack of efficiency, flexibility and generalization (Lin et al., 2021, 2023; Yang et al., 2022b; Li et al., 2023a). Therefore, researchers (Yang et al., 2022b; Li et al., 2023a) have been actively exploring the possibility of relying solely on the first crossmodal pretraining stage and aims at learning a general vision-language model that can perform well without any additional downstream fine-tuning. Successful representative methods in this line of work like FrozenBiLM (Yang et al., 2022b) freeze the language model and only train a few projection layers and a few adaptation layers during the training process to improve the efficiency. This line of research, while notable for its effectiveness, raises a pertinent question: Is the training of the projection networks truly necessary?

In this paper, we challenge the prevailing methodology and propose a novel method that eliminates the need for training projection networks while enabling the PLMs to perform zero-shot crossmodal tasks. As in Figure 1, our approach, Deep Concept Injection (DCI), injects the observed visual concepts as both additional input text tokens and augmentation in intermediate features within each feed-forwards network to enable PLMs to perceive and reason over multimodal inputs.

Our key insights are two-fold. First, towards zero-shot crossmodal tasks, it is necessary to represent the observed visual information in a way that the PLM directly understands, and our solution is to represent the observation using concepts. Inspired by (Lin et al., 2023) and (Wang et al., 2022), these visual concepts can be extracted through retrieval over a predefined vocabulary given the visual input, with the help of pretrained vision-text contrasting models like CLIP (Radford et al., 2021).

Second and more importantly, in modern PLMs based on Transformers (Vaswani et al., 2017), there are two complementary ways of fusing multimodal information. One commonly used way is to provide visual information as additional elements in the input, where the interaction between visual input and textual input is modeled in the self-attention layers. *However, self-attention layers were trained on natural sentences but not between concept words and a natural sentence. Moreover, the other possibility within feed-forward networks has been ignored.* We propose to leverage the intermediate representations of concept words (when they are solely input to the PLM) to construct adaptation layers and to achieve crossmodal fusion by estimating conditional distribution of the concept given the visual observation and the current word being processed in the PLM.

With the above two key insights, there remains one design choice to complete Deep Concept Injection: how do we choose the set of concepts? One intuitive solution is to leverage existing ontology in computer vision datasets (Krizhevsky et al., 2012; Krishna et al., 2017; Carreira and Zisserman, 2017). However, such generic datasets might not be aligned with the specific downstream tasks we are interested in. To obtain task-relevant prior, we explore two orthogonal solutions. We first exploit the setting where the access to all the possible answer words of the dataset is allowed, which is actually true for open-ended question answering datasets (Xu et al., 2017; Yu et al., 2019; Yang et al., 2021). Second, to further eliminate the assumption over prior information about the task and dataset, we propose to obtain the set of relevant concepts by querying the language model. With extensive empirical analysis on fourteen datasets, the proposed Deep Concept Injection achieves competitive or even better performance than state-of-the-art methods, without any crossmodal pretraining. We believe this paper will stimulate further research and exploration in the field, potentially opening new paths towards more efficient and versatile utilization of PLMs for crossmodal tasks.

The contribution of this paper could be summarized as follows:

- We first challenge the current methodology of zero-shot crossmodal tasks on the necessity of training additional layers and provide a negative answer by injecting observed visual concepts to PLMs to **enable zero-shot crossmodal tasks without any additional training**;
- We propose a novel method, Deep Concept Injection, to introduce visual information to PLMs by both inputting the most probable concepts as additional textual input and **constructing adaptation layers conditioned observed concepts**;
- We provide insightful empirical analysis to facilitate future research, including the necessity of crossmodal pretraining when downstream

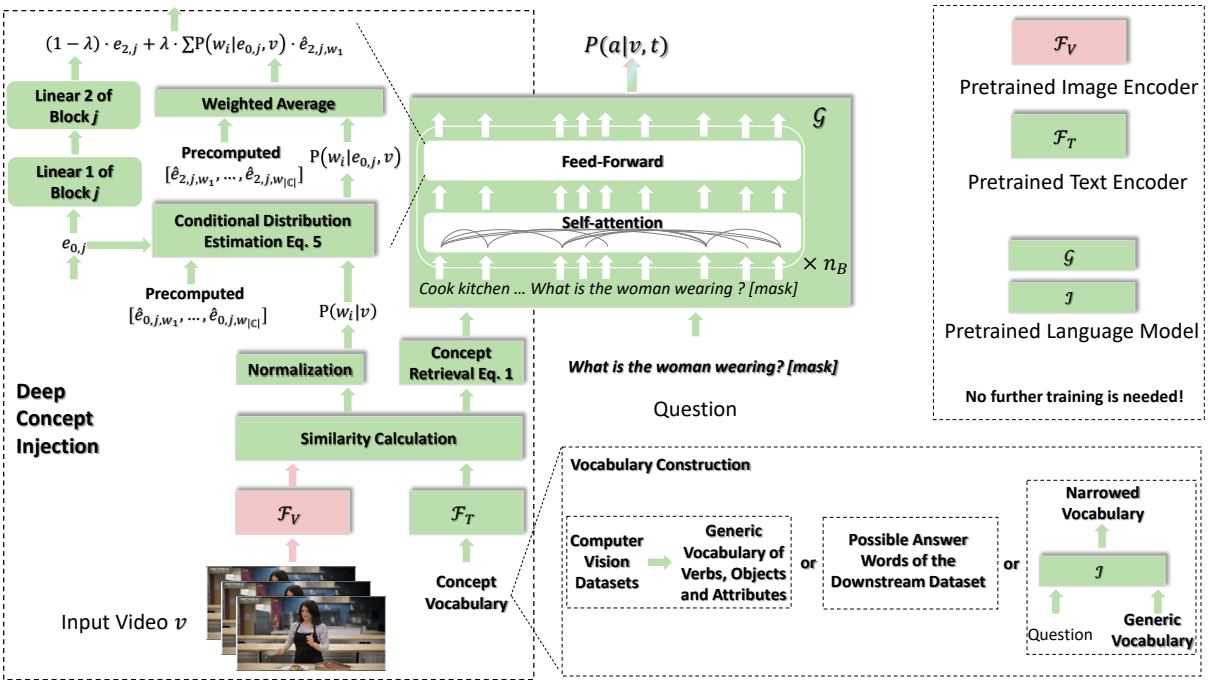


Figure 2: Injecting the observed visual concepts as both additional input text tokens and augmentation in the intermediate features within each feed-forwards network for the PLM enables zero-shot crossmodal tasks without any crossmodal pretraining. The most probable concepts extracted from visual input are additional input text so that visual information will be fused with textual information in the self-attention layers (intuitively, "cook, kitchen, ..." provide context for the question); the concept information is further injected in every feed-forward network via adding intermediate representation of concepts weighted with the conditional distribution given current word being processed and the visual input (intuitively, "cook, kitchen, ..." + "wearing" makes it closer to "apron"). Detailed descriptions of the proposed Deep Concept Injection can be found in Sec 2. This figure is best viewed in color when zoomed in.

fine-tuning is still desired, comparisons with other alternatives that don't require additional training, and DCI's versatile usage

2 Technical Approach

In this section, we first present some preliminaries (more detailed related work is discussed in the supplementary material) and then introduce the Deep Concept Injection in detail. We propose DCI based on two key ideas: speak the "language" that PLMs understand and comprehensively leverage both ways in Transformer block for crossmodal fusion. The first idea motivates us to leverage concepts (e.g., action, objects, attributes and etc.) as the bridge to transform visual information into text representations. The second idea motivates us to also utilize feed-forward networks for crossmodal fusion. Last we discuss possible ways of acquiring prior information for vocabulary construction.

2.1 Preliminaries

Crossmodal tasks. These tasks require the model to fuse information from multiple modalities, e.g., vision and text to return a text response. Specifically, we mainly consider video question answering and image captioning/visual question answering tasks in this paper. In video question answering, given a video v and question t as input, the model is required to predict the correct answer that matches the ground-truth a_l from an answer corpus $\mathbb{A} = \{a_1, .., a_{|\mathbb{A}|}\}$. In image captioning/visual question answering, the problem setting is conceptually identical; the only difference is that the visual input is a single image. In the model descriptions, we will adopt video question answering for illustration.

Pretrained Vision-Text Contrastive Models. We mainly leverage pretrained image-text contrastive models. It consists of a visual encoder $\mathcal{F}_V : \mathbb{R}^{H\times W} \longrightarrow \mathbb{R}^D$ and a text encoder $\mathcal{F}_T : \mathbb{W}^L \longrightarrow \mathbb{R}^D$, where H, W are the height and width, L is the length of the sentence, D is the dimension of the common embedding space and $\mathbb{W}$ is the set of all the words. In this paper, we mainly use it as the concept extractor because of its strong zero-shot recognition abilities (Radford et al., 2021).

Pretrained Language Models. The key is to train a model $\mathcal{G} : \mathbb{W}^L \longrightarrow \mathbb{R}^{|\mathbb{W}|}$ that predicts the

probability of a word given certain context as input. Depending on the actual objective design, the prediction could be for a masked word (Devlin et al., 2018; He et al., 2020) or the next word (Raffel et al., 2019; Chung et al., 2022). The network architecture could be also categorized as encoder-only (Devlin et al., 2018; He et al., 2020), encoder-decoder (Raffel et al., 2019; Chung et al., 2022), or decoder-only (Brown et al., 2020). All the PLMs used in this paper are based on Transformer (Vaswani et al., 2017), which consists of n_B Transformer blocks and each block's main building components are self-attention layers that models the interaction among different words, and feed-forward networks that process each word individually. The feed-forward network essentially consists of two linear layers with one activation layer.

2.2 Deep Concept Injection

In this section, we describe how to inject observed concepts comprehensively and enable crossmodal fusion in both self-attention layers and feed-forward networks.

2.2.1 Injection as Additional Textual Input.

To enable crossmodal fusion through self-attention, we extract visual concepts as additional textual input through the retrieval process as follows. First, we construct the word vectors from a predefined concept vocabulary $\mathbb{C}$; specifically, for each word c_i, we use the text encoder to obtain its word vector $\mathcal{F}_T(w_i)$. For the input video v, we encode it with the pretrained image encoder $\mathcal{F}_V(v)$ frame by frame. Then we compare the similarity between the frame embeddings and each of the words to retrieve k most similar words,

$$w_{1,1}, ..., w_{1,k}, w_{2,1}, ..., w_{F,k} = \arg\max_{i}^{k} \mathcal{F}_T(w_i)^\top \mathcal{F}_V(v), \quad (1)$$

where F is the number of frames in the video v.

Then the retrieved concepts are fed into the pretrained text model with the question t in parallel to obtain final prediction about answer a_l,

$$P(a_l|v,t) = \mathcal{G}(w_{1,1}, ..., w_{1,k}, w_{2,1}, ..., w_{F,k}, t). \quad (2)$$

We follow the temporal order of frames to concatenate retrieved words frame by frame with the question sentence t. Note for simplicity, we use a single variable t to denote the actual sentence of the question and the context text, which contains multiple words. As shown in Figure 2, "cook, kitchen, ..." will interact with question words in the self-attention layer and help to provide information about visual observation, which helps the model to reason over multimodal inputs.

2.2.2 Injection as Augmentation in the Intermediate Features of Feed Forward networks.

Since the concept words are not really natural sentences and thus the interaction is not perfectly modeled in the self-attention layers. The ignored possibility of mutlimodal fusion in PLMs lies in the feed-forward networks. We first describe how the augmentation can be added in a way that the PLM understands and then describe why this process can be considered as constructing adaptation layers.

The key of realizing any training-free augmentation for a pretrained model is to speak in the "language" that the model understands. Therefore, we first extract intermediate representation of each concept when they are input to the PLM individually,

$$\hat{e}_{0,j,w_i} = \mathcal{G}_{0,j}(w_i), \quad (3)$$

where $\hat{e}_{0,j,w_i}$ represents the intermediate representation of a concept w_i, which is input to the feed-forward network in the j-th Transformer block of the PLM. Similarly, we can extract the output representation of the feed-forward network in each Transformer block for each concept word,

$$\hat{e}_{2,j,w_i} = \mathcal{G}_{2,j}(w_i). \quad (4)$$

Note that these extraction processes only need to be done once for all the future crossmodal inference, which makes the amortized complexity to be negligible.

As shown in Figure 2, during inference for crossmodal tasks as in Eq. 2, for simplicity, we denote the input intermediate representation and the output intermediate representation of whichever word is currently being processed as $e_{0,j}$ and $e_{2,j}$, respectively. To fuse crossmodal information, we first compute the conditional distribution with the approximation that $e_{0,j}$ is independent of v,

$$P(w_i|e_{0,j}, v) \approx \frac{P(w_i|e_{0,j})P(w_i|v)}{P(w_i)}. \quad (5)$$

The factorized terms can be obtained as follows,

$$P(w_i|e_{0,j}) = \frac{\exp\left(\hat{e}_{0,j,w_i}^\top e_{0,j}\right)}{\sum_l \exp(\hat{e}_{0,j,w_l}^\top e_{0,j})}, \quad (6)$$

$$P(w_i|v) = \text{Top}_k(\text{Max-pool}(\frac{\exp\left(\mathcal{F}_T(w_i)^\top \mathcal{F}_V(v)\right)}{\sum_l \exp(\mathcal{F}_T(w_l)^\top \mathcal{F}_V(v))})), \quad (7)$$

where the Max-pool is applied along the temporal axis for the video input to handle multiple input frames and Top_k indicates that we only keep the most relevant k concept's probability to be non-zero and then scale the distribution so that the summation of probabilities is 1. This process essentially keeps the most relevant and probable visual concepts of the visual input, which we also find important empirically. We don't assume extra information about $P(w_i)$ and thus we simply apply the uniform distribution. In practice, we simply scale the product of $P(w_i|e_{0,j})$ and $P(w_i|v)$ to ensure the summation to be 1 to obtain the estimation of $P(w_i|e_{0,j}, v)$.

Then we leverage the conditional distribution to augment the output intermediate representation of the feed-forward network by adding the representation of concepts weighted based on the conditional distribution,

$$e_{2,j} = (1-\lambda)\cdot e_{2,j} + \lambda \cdot \sum_i P(w_i|e_{0,j}, v)\cdot \hat{e}_{2,j,w_i}. \quad (8)$$

Both the calculation of the conditional probability and the augmentation of the output intermediate representation can be done in parallel for each word as matrix multiplication, which leads to the equivalence to a feed-forward adaptation network

$$e_{2,j} = (1-\lambda)\cdot e_{2,j} + \lambda \cdot \text{Linear}_2(\text{Act}(\text{Linear}_1(e_{2,j};\theta_1));\theta_2), \quad (9)$$

where θ_2 is the weight matrix of the second linear layer Linear_2 whose row i is the transpose of $\hat{e}_{2,j,w_i}$, θ_1 is the weight matrix of the first linear layer Linear_1 whose column i is $\hat{e}_{0,j,w_i}$ and Act consists of both soft-max and element-wise multiplication with $P(w_i|v)$.

Intuitively, as verified in Figure 3, intermediate representation of "[mask]" could not be close to the answer "hat" but after adding the representation of observed concepts, the model can make correct prediction. Therefore, by further injecting the visual concept in the feed-forward network of each block, the visual information is comprehensively fused with the textual input for the PLM to make better prediction for crossmodal tasks.

2.3 Prior Information Acquisition for Vocabulary Construction

Existing computer vision datasets provide a generic vocabulary of visual concepts $\mathbb{C}$. Inspired by (Wang et al., 2022), we curate a comprehensive visual concept vocabulary of verbs, objects and attributes from Visual-Genome (Krishna et al., 2017; Kuznetsova et al., 2020). We denote the variant using this generic vocabulary as DCI. However, such a vocabulary could be too general for downstream tasks.

We first explore a setting with the access to the answer word vocabulary which either consists of the most frequent answers from the training set provided in the open-ended setting or consists of the answer words from the choices in the multiple-choice setting. This does not leak any information for 8 datasets of open-ended video question answering. We denote this variant as DCI-A.

To generally obtain prior information about the task to narrow down from a generic vocabulary, we propose to prompt a PLM to ask about relevant visual concepts

$$P(w_i|I) = \mathcal{I}(t), \quad (10)$$

where t is the question (and context) and $\mathcal{I}$ is not necessarily the same PLM we use for crossmodal tasks, although in our implementation we use the same model for simplicity of implementation. Then we can narrow down a subset of most n_c probable concept words from the generic vocabulary $\mathbb{C}$. We denote this variant as DCI-LM.

3 Experimental Results

In this section, we will first introduce the implementation and evaluation settings. Then we organize the following subsections by answering a set of important questions. More ablations, further analysis and other details are in the supplementary material.

3.1 Implementation and Evaluation Settings

We mainly compare with state-of-the-art video-language models using frozen PLMs and learned projection layers, FrozenBiLM and provide case studies in contrast to BLIP-2 (Li et al., 2023a) and LLaVA-1.5 (Liu et al., 2023). We follow their settings respectively to implement and evaluate our methods. Based on empirical results, we use $k = 4$, $\lambda = 0.01$, and $n_c = 1500$. More details and comprehensive ablation studies are provided in the supplementary material due to space limit.

Model	MM Samples	GPU hours	iVQA	ANet-QA	TGIF	How2QA	TVQA	LSMDC
Zero-shot Setting								
Random	NA	NA	0.1	0.1	0.1	25.0	20.0	0.1
VQA-T (Yang et al., 2022a)	72M	380	13.3	12.3	-	53.1	-	-
Reserve (Zellers et al., 2022)	1B	196K	-	-	-	-	-	31.0
Flamingo3B (Alayrac et al., 2022)	2.1B	-	32.7	-	-	-	-	-
Flamingo9B (Alayrac et al., 2022)	2.1B	-	35.2	-	-	-	-	-
Flamingo80B (Alayrac et al., 2022)	2.1B	553K	**40.7**	-	-	-	-	-
CLIP (Radford et al., 2021)	NA	NA	9.2	1.2	3.6	47.7	26.1	1.2
DeBERTa-V2 (He et al., 2020)	NA	NA	12.1	23.0	32.3	52.7	55.1	50.0
FrozenBiLM (Yang et al., 2022b)	10M	160	26.8	**25.9**	41.9	58.4	59.7	51.5
DCI (ours)	**0**	**0**	<u>28.0</u>	25.1	**45.2**	**62.8**	**60.7**	**52.4**
DCI-A (ours)	0	0	30.2	25.6	45.6	63.1	60.9	52.8
DCI-LM (ours)	0	0	28.5	25.2	45.3	62.9	60.6	52.6
Fine-tuning Setting								
MERLOT (Zellers et al., 2021)	180M	-	-	41.4	69.5	-	78.7	52.9
SiaSamRea (Yu et al., 2021)	5.6M	-	-	39.8	60.2	84.1	-	-
VQA-T (Yang et al., 2022a)	72M	380	35.4	39.0	-	85.3	-	-
Reserve (Zellers et al., 2022)	1B	196K	-	-	-	-	**86.1**	-
All-in-one (Wang et al., 2023)	138M	11K	-	-	66.3	-	-	-
VindLU (Cheng et al., 2023)	25M	2.0K	-	**44.7**	-	-	79.0	-
FrozenBiLM (Yang et al., 2022b)	10M	160	39.6	43.2	**68.6**	86.7	82.0	**63.5**
FrozenBiLM*	0	0	31.6	41.8	67.4	75.8	70.8	57.1
DCI-A (ours)	**0**	**0**	**42.6**	<u>42.8</u>	<u>68.5</u>	**89.3**	<u>81.7</u>	<u>61.6</u>

Table 1: Comparison with the state-of-the-art methods on manually-labeled video question answering datasets in terms of accuracy (%) and efficiency. Our DCI is built upon CLIP and DeBERTa-V2, as FrozenBiLM. MM Samples indicate the number of video-text samples used in the crossmodal pretraining process. GPU hours denote the additional computation required for it. Bold indicates the best results and underline means relatively better than FrozenBiLM. "-" means unclear from the original paper and "NA" is not applicable. * indicates FrozenBiLM is fine-tuned without loading pretrained projection and adaptation layers from the crossmodal pretraining stage.

FrozenBiLM is evaluated on 8 video question answering datasets: iVQA (Yang et al., 2021), ActivityNet-QA (Yu et al., 2019), TGIF-QA (Jang et al., 2017), How2QA (Li et al., 2020a), TVQA (Lei et al., 2018), LSMDC (Maharaj et al., 2017), which are manually labeled; MSRVTT-QA (Xu et al., 2017) and MSVD-QA (Xu et al., 2017), which are generated automatically from video captions and we report them separately in the supplementary material due to quality concern raised in (Lin et al., 2023). We follow its evaluation setting for each of the datasets to report results. Our models use the same CLIP ViT-L/14 (Radford et al., 2021) model and the same DeBETa-V2-XL (He et al., 2020) model as the FrozenBiLM model. In the fine-tuning setting, to maintain a fair comparison in terms of trainable parameters, we train the same adaptation layers as FrozenBiLM.

For image captioning comparison with BLIP-2 on NoCaps (Agrawal et al., 2019), we use the same Q-Former (after its first Vision-and-Language Representation Learning stage) based on ViT-g (Fang et al., 2022) and the pretrained FlanT5-XL (Chung et al., 2022). After Q-former, the extracted features of an image will have an axis for different learned queries, which can be handled in the same way as the temporal dimension in the video question answering setting illustrated in Section 2.

3.2 DCI's Effectiveness in Training-free Setting

As shown in Table 6, compared to state-of-the-art zero-shot video question answering model FrozenBiLM, without training on 10 million video-text pairs for 160 GPU hours, all the proposed DCI variants generally achieve better or competitive results on all the 6 manually-labeled video question answering datasets. On some of the datasets like iVQA and TGIF-QA, the absolute improvement is up to 3.7% and the relative improvement is up to 12.7%. In spite of the huge difference in terms of the number of parameters in the model (890M v.s. 80B) and the huge number of multimodal samples (2.1B) and cost of training (553K TPU hours), compared to Flamingo80B, our proposed DCI method successfully reduces the gap between FrozenBiLM and such gigantic multimodal large language models. We leave further scaling model size used by DCI as future research.

3.3 Effects of Vocabulary Construction Methods

As shown in Table 6, we observe that generally the DCI-A variant performs the best (such as the

Model	Projection Layer	iVQA	ActivityNet-QA	TGIF-QA	How2QA	TVQA	LSMDC
FrozenBiLM	Learned	26.8	25.9	41.9	58.4	59.7	51.5
FrozenBiLM*	Learned	**27.3**	24.7	**41.0**	53.5	53.4	50.7
CLIP+DeBERTa	Random	7.0	14.2	22.8	46.8	39.4	46.8
CLIP+DeBERTa	Constructed	24.5	24.1	39.5	55.8	57.9	51.0
CLIP+DeBERTa	Concepts	26.5	**25.1**	40.8	**57.6**	**59.4**	**51.4**

Table 2: Comparison between FrozenBiLM and its counterpart without training on the projection from visual input to PLMs. "Projection Layer" indicates how the projection layers are obtained. * denotes no adaptation layers are added for fair comparisons.

Figure 3: Attention visualization of DCI with only injections as inputs and full DCI. With the help of augmentation in the intermediate features, "[mask]" token attends more to "hat", which leads to the correct prediction. Best viewed when zoomed in.

2.2% absolute improvement from the generic vocabulary on iVQA), which is expected as the possible answer words in each dataset provide strong prior information about the task and the dataset. We also find that using the PLM to narrow down from the generic vocabulary helps to improve the performance but not as significant as DCI-A. As the hyper-parameters are tuned with only iVQA, it is still encouraging to observe a rather consistent improvement from DCI-LM. But generally the performance improvement is not as significant as the improvement from pretraining-required FrozenBiLM to our pretraining-free DCI method.

3.4 DCI's Effectiveness in Fine-tuning Setting

Despite this method being proposed in a training-free manner, it is important to understand whether DCI effectively helps to avoid the costly cross-modal pretraining stage. Therefore, we also fine-tune the models with our DCI method. Similar to FrozenBiLM, we freeze the PLM but just update the parameters of the same configured adapter networks from scratch to keep the same number of trainable parameters. As shown in Table 6, compared to directly fine-tuning FrozenBiLM without the crossmodal pretraining stage, our DCI-A equipped model significantly improves the accuracy by up to **13.5% absolute improvement**, which demonstrates the effectiveness of the proposed method for fusing visual information beyond zero-shot setting. When comparing with FrozenBiLM with 10M of more examples for crossmodal pretraining, our DCI-A still outperforms it by up to 3% of absolute gain, which further indicates **it is more important to inject visual information in a way that PLMs easily understand than to simply train them extensively**.

3.5 Alternative Methods without Training

Based on the insights discussed in Eq. 9, we provide a baseline with a constructed projection layer that requires no additional training and also helps us understand methods like FrozenBiLM. The main idea is instead of learning the projection layers, the "projected" visual features in the text embedding space could be obtained by weighted-averaging concept embeddings with the conditional distribution of concepts given the visual input. Formally, $e_t = \sum_i P(w_i|v)_t \cdot e_{w_i}$, where e_t is the "projected" visual feature of the t-th frame and e_{w_i} is the word embedding of word w_i. We further provide another baseline where instead of weighting the word embeddings of concepts, we directly concatenate the most relevant concepts as additional textual input. It is essentially only injecting concepts as inputs, without augmentation in the intermediate features.

As in Table 2, we evaluate these baselines on 6 manually-labeled video question answering datasets, and this baseline performs surprisingly well. The constructed variant significantly outperforms the random initialization and performs slightly lower than the learned FrozenBiLM, which indicates that most of the ability of the learned projection layers and the adaptation layers can be instantly obtained with the simple constructed projection layer. Such constructed projection layers or learned projection layers are inferior to directly appending the most relevant concepts, which implies that a sequence of concept words is a better representation than a single projected visual feature.

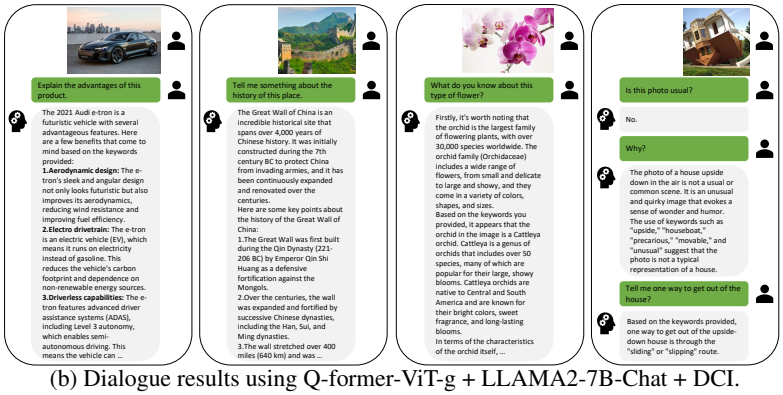
3.6 How do the Constructed Adaptation Layers help inject visual information?

In Figure 3, we visualize the average attention in the last transformer block for results from DCI with

Model	MM Samples	GPU Hours	in-domain C	in-domain S	near-domain C	near-domain S	out-domain C	out-domain S	overall C	overall S
BLIP-2 (Fine-tuned)	129M + 325K	450	123.7	16.3	120.2	15.9	124.8	15.1	121.6	15.8
BLIP-2	129M	380	104.1	14.2	102.0	13.8	**114.1**	14.1	104.8	14.9
DCI (Ours)	0	0	**107.2**	**16.8**	**105.5**	**15.3**	111.1	**14.8**	**107.0**	**15.3**

Table 3: Image captioning results on NoCaps. FLAN-T5-XL is used for fair comparisons. Note that here GPU hours of BLIP-2 are estimated based on the provided configuration in the paper. BLIP-2 fine-tuned on MSCOCO (Chen et al., 2015) is listed for reference.

(a) Dialogue results using Q-former-ViT-g + Flan-T5-XXL + DCI.



(b) Dialogue results using Q-former-ViT-g + LLAMA2-7B-Chat + DCI.

Figure 4: The proposed DCI method generalizes well to multimodal dialogue. Best viewed when zoomed in.

only injection as inputs and full DCI. We observe that the augmentation in the intermediate feature space helps the model attend more to extracted concepts that are relevant to the correct answer. Without the augmentation in the intermediate feature space brought by the Constructed Adaptation Layers, the model predicts a wrong answer even when the correct answer is retrieved as a concept. This verifies that the Constructed Adaptation Layers are complementary to injecting visual concepts as input to the PLM.

3.7 Versatile Usage of DCI

Zero-shot Image Captioning. As shown in Table 3, compared to BLIP-2 relying on 129M of multimodal samples for training the alignment between visual input and large language models, our DCI successfully outperforms in almost every metric setting on the challenging NoCaps image captioning task that stresses on the generalization to novel objects. This encouraging result demonstrates that our DCI method generalizes beyond VideoQA.

Zero-shot Multimodal Dialogue. We show the zero-shot dialogue results in Figure 4. We find the zero-shot multimodal dialogue results to be impressive. With the proposed DCI method, PLMs such as FLAN-T5-XXL and the latest LLAMA2-7B-Chat can instantly be used for multimodal dialogue without any training. For instance, for the Great Wall image, our method retrieves concepts like "china, history, journey, tourism, geography, fortress, travel, dynasty, exploring, castle, fortification..." These concepts highlight how DCI successfully captures the key semantic elements of the image, enabling the model to reason effectively about the question and generate plausible answer about the history of the Great Wall.

4 Related Work

Pre-trained Vision-Text Contrastive Models. Recently, a family of contrastively pre-trained models are introduced, which are learned from large-scale vision-text data (Miech et al., 2020; Radford et al., 2021; Li et al., 2023a). These models typically contain a visual encoder and a text encoder, and learn to map visual and text embeddings into a common

space. They sample positive/ negative pairs from aligned/unaligned image/video and texts, and train the visual and text encoders with a contrastive objective in a self-supervised manner. With access to large-scale multimodal data (e.g., 400 million web image-text pairs), they are shown superior on zero-shot recognition tasks. The resulting visual encoders have also been shown to be great feature extractors for downstream tasks (Li et al., 2020b; Yang et al., 2021, 2022b; Wang et al., 2022; Shen et al., 2021).

Crossmodal Tasks with Pretrained Language Models. Conventional methods (Lu et al., 2019; Sun et al., 2019; Yang et al., 2021) usually rely on a two-stage training process to obtain satisfying results on downstream datasets. Assuming pretrained language models and feature extractors like vision-text contrastive models (e.g., S3D (Miech et al., 2020) and CLIP (Radford et al., 2021)) are available, the first stage aims at training on web-collected vision-text dataset with techniques like masked token modeling (Li et al., 2020a; Zellers et al., 2021) or contrastive learning (Xu et al., 2021; Luo et al., 2021; Li et al., 2022; Yang et al., 2021) to learn to align and fuse visual and textual inputs. In the second stage, the model is further fine-tuned with human annotation on downstream datasets (Yang et al., 2021; Yu et al., 2019; Li et al., 2020a; Xu et al., 2017; Zhou et al., 2018; Wang et al., 2019) for better downstream performance.

Such a two-stage training process has been criticized for a lack of efficiency and flexibility because of the huge cost of the first training stage (Lin et al., 2021, 2023), and they are also not general enough (Yang et al., 2022b; Li et al., 2023a). There are two lines of following research trying to address the limitation of the two-stage training process. One line of work (Lin et al., 2021, 2023) focuses on obtaining competitive models with only the second training stage on downstream datasets and one successful idea is to transform every modality into concept text (Lin et al., 2021, 2023) so that the PLM can immediately understand and leverage the information from other modalities without the expensive first training stage. However, such methods still rely on human annotation and specific training towards each downstream dataset.

The other line of work (Alayrac et al., 2022; Yang et al., 2022b; Li et al., 2023a) relies solely on the first training stage and aims at learning a general vision-language model that can perform well in the zero-shot setting without any additional downstream fine-tuning. During the training process, successful methods in this line of work such as FrozenBiLM (Yang et al., 2022b) freeze the language model and only train a few projection layers and a few feed-forward adaptation layers to project the visual features extracted by a frozen feature extractor like CLIP, to improve the efficiency. The typical training target is, with the video/image as input, generating the associated text. It is noteworthy that, although the pretrained model exhibits the ability to perform zero-shot crossmodal tasks such as video questions answering, to obtain higher performance on downstream tasks, fine-tuning is still crucial (Yang et al., 2022b; Liu et al., 2023) to achieve superior performance. Unlike existing research, we explore a more challenging new problem where there is no additional training or labeled training samples for downstream tasks.

5 Conclusion

In this paper, we present a novel approach to enabling pretrained language models to perform video question answering without any training. The proposed Deep Concept Injection effectively circumvents the necessity of training projection networks, a widely accepted practice in this field, and instead makes insightful use of observed visual concepts as additional input text tokens and as a means for augmenting intermediate features. Extensive results show that they function synergistically to realize strong zero-shot crossmodal capabilities of the PLM and to bypass the costly crossmodal pretraining stage in versatile tasks and settings.

6 Acknowledgement

This research is partially supported by U.S. DARPA ECOLE Program No. #HR00112390060. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. This research alsop obtained support by the funds provided by the National Science Foundation and by DoD OUSD (R&E) under Cooperative Agreement PHY-2229929 (ARNI: The NSF AI Institute for Artificial and Natural Intelligence). We would like to also thank all the other colleagues and anonymous reviewers for their valuable help.

7 Limitations

One limitation in this work is that only crossmodal tasks over vision and text are evaluated. Since we have already covered 15 datasets, we leave further exploiting broader combinations and tasks as future work. However, the proposed approach is rather generic: as long as there is a concept extractor for modality X, preferably a pretrained X-text contrastive model for modality X and text, the proposed DCI can be applied instantly. Another limitation of the proposed method is that it certainly adds additional running time during inference because of the extra computation, but the main complexity still comes from the inference of the large PLM itself. We also want to acknowledge that more complex spatial-temporal relationship is still rather under-explored in this work to be consistent with the main counterpart model such as FrozenBiLM.

We also note that in the current evaluations, the size of the PLM used is still rather limited to a rather small scale. Further scaling up the language model is another interesting future work. We also would like to note that we assume there is no access to good captioning models for all the models evaluated. In practice, further augmenting inputs with captions generated by pretrained captioning models could possibly further improve the performance, which is orthogonal to the setting and approaches explored in this paper.

Due to the nonlinear nature of transformers and multimodal tasks, the community generally lacks effective theoretical tools to analyze such large models, to the best of our knowledge. Therefore, we leave more theoretical analysis as future work.

While we do not anticipate direct negative social consequences stemming from the work, it is important to note our work relies on pre-trained models, which could potentially exhibit certain biases.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. 2023. Vindlu: A recipe for effective video-and-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10739–10750.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Enabling multimodal generation on clip via vision-language knowledge distillation. *arXiv preprint arXiv:2203.06386*.

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

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2022. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,

and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2758–2766.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

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

Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. 2022. Clip-event: Connecting text and images with event structures. *Proc. Conference on Computer Vision and Pattern Recognition (CVPR2022)*.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020b. Cross-media structured common space for multimedia event extraction. In *Proc. The 58th Annual Meeting of the Association for Computational Linguistics (ACL2020)*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. 2021. Vx2text: End-to-end learning of video-based text generation from multimodal inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7005–7015.

Xudong Lin, Simran Tiwari, Shiyuan Huang, Manling Li, Mike Zheng Shou, Heng Ji, and Shih-Fu Chang. 2023. Towards fast adaptation of pretrained contrastive models for multi-channel video-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14846–14855.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2021. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*.

Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. 2017. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6884–6893.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2021. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. *Preprint*, arXiv:1904.01766.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. 2023. All in one: Exploring unified video-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. 2022. Language models with image descriptors are strong few-shot video-language learners. *Proc. The Thirty-Sixth Annual Conference on Neural Information Processing Systems (NeurIPS2022)*.

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

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

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

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022a. Learning to answer visual questions from web videos. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–1.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022b. Zero-shot video question answering via frozen bidirectional language models. *arXiv preprint arXiv:2206.08155*.

Weijiang Yu, Haoteng Zheng, Mengfei Li, Lei Ji, Lijun Wu, Nong Xiao, and Nan Duan. 2021. Learning from inside: Self-driven siamese sampling and reasoning for video question answering. In *Advances in Neural Information Processing Systems*, volume 34, pages 26462–26474. Curran Associates, Inc.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9127–9134.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387.

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

Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

A Can DCI serve as a plug-and-play augmentation for models requiring additional training?

The motivation of DCI is to eliminate additional training and to enable PLMs to perform cross-modal tasks directly. Since there are already trained

Model	Fine-tuned?	iVQA	ANet-QA	TGIF-QA	How2QA	TVQA	LSMDC	MSRVTT	MSVD
FrozenBiLM (Yang et al., 2022b)	No	26.8	25.9	41.9	58.4	59.7	51.5	16.7	33.8
+ DCI-A (ours)	No	**30.6**	**26.1**	**46.3**	**59.5**	**59.8**	**52.4**	**17.3**	**35.0**
FrozenBiLM (Yang et al., 2022b)	Yes	39.6	43.2	68.6	86.7	**82.0**	63.5	47.0	54.8
+ DCI-A (ours)	Yes	**40.4**	**43.3**	**69.5**	**87.1**	81.9	**63.8**	**47.6**	**55.0**

Table 4: Results (%) of plugging DCI-A into FrozenBiLM on iVQA, ActivityNet-QA, TGIF-QA, How2QA, TVQA, LSMDC, MSRVTT-QA and MSVD-QA. "Fine-tuned" indicates whether the FrozenBiLM model is further fine-tuned on each downstream datasets. Bold indicates the better results.

models, it is important and interesting to explore the flexibility of the proposed DCI as a plug-and-play augmentation to these trained models. We take FrozenBiLM for this case study as its trained and fine-tuned checkpoints have all been released. Specifically, for the input sequence, we append the retrieved visual concepts between the projected visual features and the question text; for the augmentation in the intermediate representations, we perform exactly the same augmentation process for every input token.

As shown in Table 4, we extensively evaluate both FrozenBiLM trained with video-text pairs and its variants further fine-tuned on each downstream dataset, with the proposed DCI-A as a plug-and-play augmentation. We observe that even when the projection and adaptation layers are well trained or even fine-tuned towards the specific downstream task, our DCI-A can still help to better fuse the visual information with textual information. This again verifies the necessity of injecting observed concepts and the complementarity with existing approaches.

B Speed Comparison

As shown in Table 5, we measure the inference speed on a V100 GPU with batch size 1 on the validation set of the iVQA dataset. The running time is shown in the following table. The increase in running time of DCI is rather tolerable compared to other models like FrozenBiLM. The time of one ablation experiment of DCI typically takes about 1 GPU minute. Models like FrozenBiLM also need hyper-parameter search, which is much more expensive.

Method	Running time (seconds per iteration)
FrozenBiLM	0.0461 ± 0.0010
DCI (Ours)	0.0495 ± 0.0013

Table 5: Inference speed comparison.

C Comparisons on MSRVTT-QA and MSVD-QA

MSRVTT-QA (Xu et al., 2017) and MSVD-QA (Xu et al., 2017), which are generated automatically from video captions and we report them separately here due to quality concern raised in (Lin et al., 2023). Despite their wide usage in the existing literature, their nature of being automatically generated, which is even shown to be worse than the automatic pretraining data generation pipeline proposed in Just-ask (Yang et al., 2021), determines that they are not suitable for evaluation given all the other six manually annotated video question answering datasets available. Regardless, we observe in Table 6 that the proposed DCI method helps to obtain comparable performance on these two datasets without expensive crossmodal pretraining in both zero-shot and fine-tuning settings.

D Comparison with BLIP-2 on Visual Question Answering

As shown in Table 7, compared to state-of-the-art zero-shot visual question answering model BLIP-2, without training on 129 million video-text pairs for 1 thousand GPU hours, all the proposed DCI variants still generally achieve better or competitive results on all the 3 visual question answering datasets. It is noteworthy that on VQAv2, with a smaller PLM FlanT5-XXL (12B), the proposed DCI even outperforms Flamingo80B by 9.6% of absolute accuracy.

E Instruction-tuning without crossmodal pretraining.

Beyond zero-shot training-free setting, we are also interested in whether the proposed DCI method can also help to bypass the crossmodal pretraining stage when instruction tuning resources are available. As Table 8 shows, compared to training LLaVA-1.5 without crossmodal pretraining, our DCI provides consistent improvement on five evaluation benchmarks (Singh et al., 2019; Lu et al., 2022; An-

Model	MM Samples	GPU hours	MSRVTT-QA	MSVD-QA
Zero-shot Setting				
Random	NA	NA	0.1	0.1
VQA-T (Yang et al., 2022a)	72M	380	5.6	13.5
Reserve (Zellers et al., 2022)	1B	196K	5.8	-
Flamingo3B (Alayrac et al., 2022)	2.1B	-	-	27.5
Flamingo9B (Alayrac et al., 2022)	2.1B	-	-	30.2
Flamingo80B (Alayrac et al., 2022)	2.1B	553K	-	**35.6**
CLIP (Radford et al., 2021)	NA	NA	2.1	7.2
DeBERTa-V2 (He et al., 2020)	NA	NA	6.5	11.7
FrozenBiLM (Yang et al., 2022b)	10M	160	16.7	33.8
DCI (ours)	**0**	**0**	**17.2**	34.5
DCI-A (ours)	0	0	17.6	35.1
DCI-LM (ours)	0	0	17.4	34.4
Fine-tuning Setting				
MERLOT (Zellers et al., 2021)	180M	-	43.1	-
SiaSamRea (Yu et al., 2021)	5.6M	-	41.6	45.5
VQA-T (Yang et al., 2022a)	72M	380	41.8	47.5
All-in-one (Wang et al., 2023)	138M	11K	46.8	48.3
VindLU (Cheng et al., 2023)	25M	2.0K	44.6	-
FrozenBiLM (Yang et al., 2022b)	10M	160	**47.0**	**54.8**
FrozenBiLM*	0	0	46.2	51.9
DCI-A (ours)	**0**	**0**	46.6	54.3

Table 6: Comparison with the state-of-the-art methods on automatically-labeled video question answering datasets in terms of accuracy (%) and efficiency. Our DCI is built upon CLIP and DeBERTa-V2, as FrozenBiLM. MM Samples indicate the number of video-text samples used in the crossmodal pretraining process. GPU hours denote the additional computation required for it. Bold indicates the best results and underline means relatively better than FrozenBiLM. "-" means unclear from the original paper and "NA" is not applicable. * indicates FrozenBiLM is fine-tuned without loading pretrained projection and adaptation layers from the crossmodal pretraining stage.

tol et al., 2015; Hudson and Manning, 2019; Li et al., 2023b). On TextVQA (Singh et al., 2019), and Pope (Li et al., 2023b) and ScienceQA (Lu et al., 2022), our crossmodal-pretraining-free even achieves better results compared to LLaVA-1.5 with crossmodal pretraining. This demonstrates the versatile usage of the proposed DCI method and prompts us to rethink the value of crossmodal pretraining: the performance gap resulting from the absence of crossmodal pretraining is marginal compared to the one resulting from a larger-scale instruction fine-tuning setting, which again challenges the necessity of crossmodal pretraining in the image-language domain.

We observe that the model benefits more on the Language Science subject where the model is required to perform certain reasoning with commonsense based on image context. For example, as shown in Figure 5, the model is asked which word best describes the sound this hammer makes, given an image of some one driving nails on fence. LLaVA answers buzzing, which is incorrect. But with concepts such as "hammer, picket, nail, fence" retrieved, the model successfully answers banging. Such examples indicate that directly using concepts as image representation help reduce possible visual hallucination (which aligns with the improvements on the POPE dataset) or better recalls the commonsense knowledge that the PLM possesses.

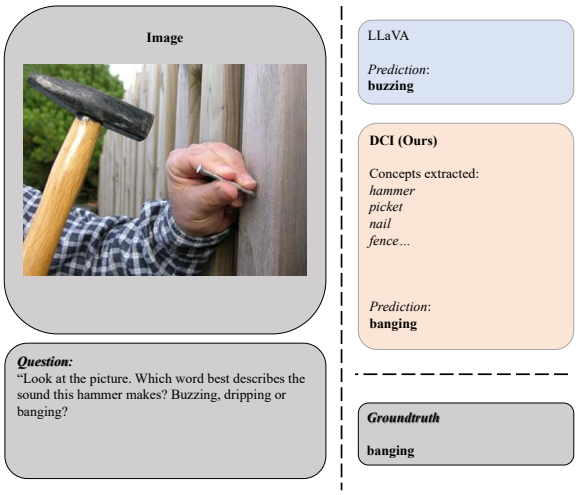


Figure 5: Visualization of results on ScienceQA.

Model	MM Samples	GPU hours	VQAv2 test-dev	OK-VQA test	GQA test-dev
VLKD (Dai et al., 2022)	3.7M	320	44.5	13.3	-
Flamingo3B (Alayrac et al., 2022)	2.1B	-	49.2	41.2	-
Flamingo9B (Alayrac et al., 2022)	2.1B	-	51.8	44.7	-
Flamingo80B (Alayrac et al., 2022)	2.1B	553K	56.3	**50.6**	-
BLIP-2 (Li et al., 2023a)	129M	1K	65.0	45.9	44.7
DCI (ours)	0	0	64.5	46.3	45.2
DCI-A (ours)	0	0	**65.9**	46.8	**45.4**
DCI-LM (ours)	0	0	65.4	46.9	45.2

Table 7: Comparison with the zero-shot state-of-the-art on visual question answering in terms of accuracy (%) and efficiency. Our DCI is built upon the same pretrained models as BLIP-2 ViT-g FlanT5XXL. MM Samples indicate the number of image-text samples used in the crossmodal pretraining process. GPU hours refer to the additional computation required for it. Bold indicates the best results. "-" means unclear from the original paper.

Model	MM Samples	GPU hours	TextVQA	ScienceQA	VQAv2	GQA	POPE
LLaVA-1.5 (Liu et al., 2023) (Full tuning)	558K	320	58.2	66.8	78.5	62.0	85.9
LLaVA-1.5	558K	320	53.7	67.6	**76.5**	**59.1**	86.3
LLaVA-1.5*	0	0	52.0	67.5	74.3	57.2	85.3
DCI (Ours)	**0**	**0**	**54.0**	**69.2**	74.9	57.8	**86.9**

Table 8: Comparison with LLaVA-1.5 in the instruction fine-tuning setting with Vicuna-7B. MM Samples indicate the number of image-text samples used in the crossmodal pretraining process. GPU hours denote the additional computation required for it with V-100 machines. "*" indicates that the pretrained projection layers from the crossmodal pretraining stage are not loaded for fair comparison. Full tuning indicates the setting using larger data (665K) and batch size (128) as in the paper (Liu et al., 2023), and the rest are all obtained using the same smaller training setting (166K, 64).

As Additional Input	As Augmentation in Features	Acc. (%)
✗	✗	12.1
✓	✗	26.5
✗	✓	13.2
✓	✓	28.0

Table 9: Accuracy with different combinations of injection mechanisms on iVQA.

F Ablation Studies

In this section, we report the results of ablation studies on the iVQA dataset.

F.1 Effect of the two Injection Pathways

As shown in Table 9, we observe that injecting observed visual concepts as additional textual context contributes to the main improvement over the language model-only baseline (no injection is used). The Constructed Adaptation Layers help to further improve the performance. This is expected as the direct injecting of additional textual input leverages the well-trained self-attention layers to fuse information between text and vision, and thus, it is easier to provide visual information to the PLM. However, this is not complete or perfect as the PLM may not be able to directly fuse the visual concepts with other textual input well because the visual concepts are not the same as natural sentences. Augmenting the intermediate features helps to further inject visual information explicitly, which complements the previous mechanism by their designs and is verified by the empirical results.

F.2 Constructed Adaptation Layers Inserted in Different Depth

We first ablate on the depth where the constructed adaptation layers are inserted. As shown in Table 10, we generally observe that with fewer layers used to insert the constructed adaptation layers, the resulting models perform worse than the default design where all the blocks are inserted with the constructed adaptation layers, which is expected since without training, it is intuitive to gradually inject visual information block by block.

F.3 Constructed Adaptation Layers with Different Intermediate Embeddings

We then ablate on the different variants of constructing the adaptation layers where different intermediate embeddings are used. As shown in Table 11, we observe that either using all e_0 or all e_2 variants yields lower performance. We suppose this is consistent with the multiple-layer design within the feed-forward networks: the early layer also serves to produce a "distribution" between input and the

Depth	iVQA Accuracy (%)
First Half Feed-forward Networks	27.1
Second Half Feed-forward Networks	27.2
Even Feed-forward Networks	26.9
Odd Feed-forward Networks	26.7
All (Default)	28.0

Table 10: Comparison on the iVQA dataset when different depths of the constructed adaptation layers are inserted at.

Text-conditioned Distribution with	Weighted Average Embeddings with	iVQA Accuracy (%)
e_0	e_0	26.6
e_2	e_2	27.4
e_0 (Default)	e_2 (Default)	28.0

Table 11: Comparison on the iVQA dataset when different intermediate embeddings are used.

Method	iVQA Accuracy (%)
FrozenBiLM	23.8
DCI (Ours)	25.3

Table 12: Comparison with FrozenBiLM on the iVQA dataset when ImageNet pretrained model is used as the feature/concept extractor.

internal knowledge elements and then the "distribution" is used to re-weight internal knowledge elements stored in later linear layers.

F.4 Using ImageNet Classification Model for Concept Extraction

To understand whether our model generalizes beyond vision-text contrastive model for concept extraction, we use the same ViT pretrained on ImageNet21k as FrozenBiLM in its Table 14. As shown in Table 12, The superior results of our DCI achieved again verifies it effectiveness of enabling zero-shot multimodal reasoning without training. The performance is lower than using CLIP for concept extraction as expected, which is also observed by (Alayrac et al., 2022) because "our goal is to use the Vision Encoder as a feature extractor for the Flamingo models in order to capture the whole scene and not just the main object".

F.5 Hyper-parameter Selection

We first vary the three hyper-parameters introduced in the proposed DCI method, the number of concepts retrieved, the injection weight, and the vocabulary size when we use the PLM to narrow down from the generic vocabulary. As shown in Table 13a, we observe that using $k = 4$ produces the best results and changing number of words around 4 does not change the performance too much. As presented in Table 13b, we find that using a relatively small $\lambda = 0.01$ for injection as augmentation in the intermediate feature works better. When λ is significant larger, the performance degrades, which is intuitively understandable as this would change the intermediate representation of the model too much. As shown in Table 13c, we observe that significantly narrowing down the vocabulary by one order of magnitude helps to improve the accuracy but when the vocabulary is too small the performance would also degrade. Overall, we find that within the range we explored, the performance of the method w.r.t. hyper parameters is stable.

F.6 Performance Breakdown on ActivytyNet-QA

We report the detailed performance breakdown based on the manually labeled types of QA in the ActivityNet-QA dataset. We observe that there are certain types of questions that our method achieves significant improvement, such as Color, Number and Yes-No. We believe this is because that these important concepts like colors are directly represented in our method compared to using a projected visual feature vector, which makes it easier for the model to obtain the required information for answering the question. Over all the types, all the methods including our method performs poorly on Temporal-related QA, which indicates a possible future direction for further improvement.

G Additional Details

G.1 Implementation Details

We implement the DCI method using PyTorch and inject our implementation to publicly available

k	Accuracy (%)
2	27.9
4	28.0
6	27.3

(a) The number of retrieved concepts.

λ	Accuracy (%)
0.005	27.8
0.01	28.0
0.015	28.0
0.1	26.5

(b) The injection weight.

n_c	Accuracy (%)
500	27.6
1000	28.1
1500	28.5
2000	28.4
2500	28.2
10738 (Full)	28.0

(c) The vocabulary size.

Table 13: Results for hyper-parameter selection on the iVQA validation set.

Model	Motion	Spatial	Temporal	Yes-No	Color	Object	Location	Number	Other
VQA-T (Yang et al., 2021)	2.3	1.1	0.3	36.3	11.3	4.1	6.5	0.2	4.7
FrozenBiLM (Yang et al., 2022b)	**12.7**	**6.8**	**1.6**	53.2	16.5	17.9	**18.1**	26.2	**25.8**
DCI (ours)	11.0	4.8	0.8	55.2	23.2	**18.6**	10.2	25.7	22.3
DCI-A (ours)	11.3	5.8	1.3	55.3	**24.7**	16.5	11.2	**29.6**	22.0
DCI-LM (ours)	10.8	4.9	1.4	**55.4**	24.6	16.9	11.2	29.2	22.2

Table 14: Results for different types of QA on the ActivityNet-QA test set.

code repositories of the base models, respectively. We use half precision for model parameters to save memory and improve speed during inference. All the experiments on video question answering are done with 4 Nvidia V100-32GB GPUs. Experiments for comparisons with BLIP-2 are done with a Nvidia A100-40GB GPU. Experiments for comparisons with LLaVA are done with 4 Nvidia A100-40GB GPUs.

For comparison with LLaVA-1.5 (Liu et al., 2023) in the instruction-tuning setting, with CLIP-L (Radford et al., 2021) and Vicuna-7B (Chiang et al., 2023) we use a smaller batch size (64), LoRA (Hu et al., 2021) training and only 25% percent of its 665K instruction tuning data due to limited training resources.

For video question answering tasks, we follow the prompt of FrozenBiLM to query the language model with questions and additional input and determine the answer based on the probability obtained for the "[mask]" token. For visual question answering and image captioning, we follow the same setting of BLIP-2 or LLaVA to generate answers and then compare with the ground-truth when comparing with them, respectively.

To construct the vocabulary, we follow VidIL (Wang et al., 2022) to construct vocabulary. There are 2,138 verbs, 6,369 objects and 7,233 attributes curated for the vocabulary. Merging and deduplication results 10,738 unique concept words. We find that directly using all these concept words together as one vocabulary has already helped, so we do not perform further fine-grained processing among different categories of concept words.

When computing the intermediate representations for each concept word, we simply average the representation if there are multiple tokens in the concept word. For fine-tuning experiments, we follow the same hyper-parameters as used in FrozenBiLM. Our code will be made publicly available upon publication.

G.2 Dataset and Evaluation Metric for Ablation Study

iVQA (Yang et al., 2021) contains 10,000 instructional videos. Each video is annotated with one question and five corresponding answers. In the official split, there are 6,000, 2,000, and 2,000 videos for training, validation, and testing, respectively. We use the 2,000 videos in the test set for ablation study in the appendix (when not specified) and follow the test split of all the datasets used in FrozenBiLM and BLIP-2 to report results in the main paper. We follow (Yang et al., 2021) to calculate accuracy with five annotations per question.

H More discussion on zero-shot multimodal dialogue results

One interesting aspect of the results here is that the model was able to recognize some named entities. After checking the reconized concepts, we

find that some of the entities are indeed part of the vocabulary like audi e-tron. For the Great Wall image, the recognized concepts include "china", "fortification", and "tourism". The PLM successfully inferred the most famous Great Wall based on these concepts. Currently, we don't intentionally handle named entities in our vocabulary, but this ability can be further integrated if we can also provide a list of named entities that we want the model to recognize, which will be an interesting future research direction.