Ego-VPA: Egocentric Video Understanding with Parameter-efficient Adaptation

Tz-Ying Wu[1,2]   Kyle Min[1]   Subarna Tripathi[1]   Nuno Vasconcelos[2]

[1]Intel Labs   [2]UC San Diego

{tz-ying.wu,kyle.min,subarna.tripathi}@intel.com   nvasconcelos@ucsd.edu

Abstract

Video understanding typically requires fine-tuning the large backbone when adapting to new domains. In this paper, we leverage the egocentric video foundation models (Ego-VFMs) based on video-language pre-training and propose a parameter-efficient adaptation for egocentric video tasks, namely Ego-VPA. It employs a local sparse approximation for each video frame/text feature using the basis prompts, and the selected basis prompts are used to synthesize video/text prompts. Since the basis prompts are shared across frames and modalities, it models context fusion and cross-modal transfer in an efficient fashion. Experiments show that Ego-VPA excels in lightweight adaptation (with only 0.84% learnable parameters), largely improving over baselines and reaching the performance of full fine-tuning.

1. Introduction

Video understanding models have achieved satisfactory performance on various downstream tasks, such as video captioning [33, 38, 47], retrieval [2, 24] and action classification [25, 51]. These models are typically trained on the supervised video datasets of interest [4, 6, 10]. Inspired by visual-language contrastive learning [24, 27], recent research has been shifted to training video foundation models (VFMs) [1, 2, 21, 30, 47, 51] on large datasets, to produce representations that generalize to multiple tasks. Prior work has focused on aligning the video and text representations of the VFM, by developing novel training objectives [1, 21, 45], or leveraging data from other modalities, such as speech recognition [38] or language models [51]. While this research has improved zero-shot performance on unseen domains, there exists a gap between the latter and that of full VFM fine-tuning [21, 41, 51]. This gap ensues from a statistics mismatch between the pretraining data and the application of interest, due to factors such as background variability, video context, etc. It reduces the practical value of VFMs, since fine-tuning usually requires extensive computation, and model parameters can grow exponentially when adapting to multiple tasks. This work focuses on the lightweight
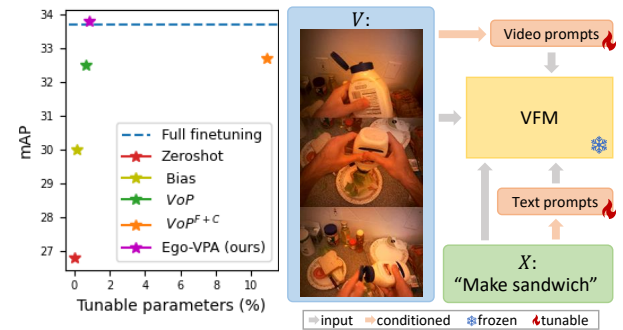

Figure 1. Ego-VPA leverages context-aware prompts to achieve parameter-efficient adaptation for egocentric videos. (Left) Performance vs tunable parameters; (Right) Cross modality prompt-tuning in Ego-VPA where the VFM is frozen.

Table 1. The zero-shot - fine-tuned performance gap (mAP on Charades-Ego) exists in both CLIP-based VFMs [26, 43] and Ego-VFMs [51]. * denotes that only the prompts/adapters are fine-tuned on Charades-Ego.

Method	zero-shot	fine-tuned
CLIP-based VFM		
X-CLIP [26]	24.0	30.0*
Vita-CLIP [43]	25.8	31.3*
Egocentric VFM		
LaViLa [51]	26.8	33.7

adaptation of VFMs for egocentric (first-person view) videos.

The gap between zero-shot and fine-tuning performance also holds when image-based foundation models (IFMs) (i.e. CLIP [31]) are used for image recognition [44]. Several parameter-efficient adaptation techniques have been developed, including the use of adapters [8] and prompt-tuning [14, 52, 53]. This inspired a few recent applications to video understanding [13, 28, 39, 43], based on the use of IFMs to encode individual video frames. We hypothesize that this is insufficient for egocentric video, since IFMs are pretrained on third-person views, not first-person, and lack temporal reasoning capabilities. The use of existing lightweight adaptation techniques is insufficient to overcome these barriers. We validate this hypothesis by demonstrating the inefficiency of two IFM-based video models [26, 43] for egocentric video understanding, as shown in Table 1.

In this work, we investigate efficient prompt tuning based frameworks to address light-weight adaptation of VFMs,

in particular Ego-VFMs [21, 30, 51] which are pretrained on egocentric videos. Since no prior work investigated this problem, we begin with introducing several baselines, including adding learnable prompts only to the text or the video encoder. Single-modality prompts unsurprisingly lead to marginal improvements over zero-shot VFM performance as they fail to capture connections between the text and spatial-temporal context. We then show that improved results can be obtained by prompt-tuning both the video and text encoders, using the recent VoP approach [13], which uses an additional module, a bi-directional LSTM, to connect visual prompts across frames, according to context. While achieving good performance, this introduces a large parameter overhead, compromising the lightweight nature of the adaptation (See trainable parameters in Table 2).

To address this, we propose a parameter-efficient prompt-tuning approach, **Ego-VPA**, for Ego-VFMs. Ego-VPA uses an encoder to project frame feature vectors into a latent prompt space. It then learns an orthogonal basis for this space, the *prompt basis*, which can be seen as a principal component analysis of prompt space. Given the latent space projection of a frame feature vector, Ego-VPA then determines the subspace of k basis-prompts that best approximates it, in the least squares sense. The selected basis prompts are then used to synthesize k video prompts for the video frame, using a linear latent decoder. Since it is trained over the entire dataset, the prompt basis is representative of all frames, allowing efficient cross-frame context modeling. In addition, since the basis prompts capture the semantic content of the video, the method can be naturally extended to cross-modal prompt synthesis, where the basis prompts that best reconstruct the projected video/text feature are used to synthesize video/text prompts. Figure 1 summarizes the overall cross-modal prompt-tuning.

To highlight the effectiveness and efficiency of Ego-VPA, three popular egocentric video datasets (i.e. Charades-Ego [35], EGTEA [20], and EPIC-Kitchens-100 [7]) are evaluated. We show that Ego-VPA outperforms the prompt-tuning baselines and is even superior to the fully fine-tuned Ego-VFM on Charades-Ego and EGTEA. More importantly, Ego-VPA only requires 0.84% additional model parameters, which is much more efficient than other baselines (See Figure 1). Ablations show that Ego-VPA consistently outperforms the baselines when different numbers of frames per video or amounts of training data are used.

Overall, we make three contributions to the efficient adaptation for egocentric video understanding. First, we show that prompt-tuning video encoders based on IFMs are suboptimal for egocentric video tasks due to the inherent domain gap. Second, we propose several baselines that prompt-tune the existing Ego-VFMs and show that these baselines are not effective and efficient. Finally, we propose a novel and efficient prompt-tuning approach, Ego-VPA, that utilizes a local subspace approximation with shared basis

prompts for cross-modal prompt synthesis, enabling context reasoning across frames and modalities.

2. Related Works

Video foundation models (VFMs). VFM learns a generalizable representation for videos, which is applicable to many downstream tasks, such as video captioning [33, 38, 47], retrieval [2, 24], action classification [1, 25, 51]. One approach to achieve this is to expand existing image-language foundation models (IFMs) [31] to the video domain [13, 17, 22, 26, 32, 42, 43, 46]. These works show promising results on some short-term or third-person-view video understanding tasks [15, 16, 36]. However, since most of the existing IFMs [31] are pretrained on static internet images, IFMs require additional temporal reasoning modules to capture scene dynamics and may fail to extract meaningful representations for frames in egocentric videos. Another line of works adopts video-language pre-training (VLP) [1, 2, 18, 21, 23, 30, 33, 41, 45, 47, 50, 51] to learn transferable spatial-temporal representation from large-scale video datasets [2, 9, 25]. For example, UniVL [23], All-in-one [41] and Lavender [18] propose a general pre-training method that can support many tasks and achieve solid zero-shot performance. Recently, LaViLa [51] shows that VLP can benefit from the dense narrations generated by Large Language Models (LLMs). In this work, we focus on LaViLa, which is the SOTA Ego-VFM on egocentric videos [7, 20, 35].

Parameter-efficient adaptation. Adaptation techniques have been widely used in natural language processing (NLP) for efficiently adapt pretrained LLMs for domain-specific tasks. For example, task-specific modules (i.e. adapters) [11, 12, 29, 37] are integrated into transformers for efficient adaptation. An alternative is to prompt-tune [19, 34] the large models, where extra tokens are prepended to the model input and are optimized with domain-specific losses. In both cases, the large model remains fixed. These adaptation techniques in NLP are then adopted in the computer vision field. For example, VL-adapter [39] and ST-adapter [28] propose effective adapter-based methods for the tasks of image-language and video understanding, respectively. VPT [14], CoOp [53] and CoCoOp [52] prompt-tune image transformers and CLIP [31] for image recognition tasks. Recently, VoP [13] extended prompt-tuning techniques to the video-language domain by applying video and text prompts. We adapt VoP and its variants to Ego-VFM as strong baselines and propose an efficient way to model context fusion atop the baselines that allows knowledge sharing across frames and modalities.

3. Egocentric Video Understanding with VFMs

3.1. VFM Preliminaries

Inspired by the success of IFMs on image applications, two types of VFMs have been proposed for video. One extends existing IFMs (e.g. CLIP) to the video domain [13, 26, 43], by first encoding individual frames with an IFM and then

fusing them with a temporal reasoning component. The other directly employs a video encoder to learn spatial-temporal representations [41, 51]. These models use video-language pretraining to learn representations that align two modalities. Similar to IFMs, most VFMs [1, 2, 51] adopt dual-encoder design, where a video encoder ϕ_{vid} and a text encoder ϕ_{txt} extract features from each video V and its corresponding text description X (e.g., "turning on the light"), respectively. ϕ_{vid} and ϕ_{txt} are usually transformers [1, 2, 51]. Given an input sequence $\mathbf{Z}^{(0)} \in \mathbb{R}^{N_t \times d}$ with N_t tokens, the L-layer transformer performs the mapping

$$\mathbf{Z}^{(l+1)} = Att(\mathbf{Z}^{(l)})_{l=0...L-1}, \tag{1}$$

where $Att(\cdot)$ is a transformer block [40]. For $l = 0$, a learnable positional encoding that specifies the relative position between the tokens is added. In the following, we omit all learnable positional encoding for brevity and use the subscript vid and txt to represent variables from the video and text domains, respectively.

Text encoder ϕ_{txt}. Given a textual description X, each tokenized word is mapped into a text embedding $\mathbf{z}_i^{(0)} \in \mathbb{R}^{d_{txt}}$. The transformer of Eq. (1) takes $\mathbf{Z}_{txt}^{(0)} = (\mathbf{z}_{[SOS]}^{(0)}, \mathbf{z}_1^{(0)}, ..., \mathbf{z}_{N_w}^{(0)}, \mathbf{z}_{[EOS]}^{(0)})^T$ as the input, where $\mathbf{z}_{[SOS]}^{(l)}, \mathbf{z}_{[EOS]}^{(l)} \in \mathbb{R}^{d_{txt}}$ are special tokens representing the start and the end of the sequence, which is mapped into a token sequence $\mathbf{Z}_{txt}^{(l)}$ at each transformer layer l, and defines the text feature as $\phi_{txt}(X) = \mathbf{z}_{[EOS]}^{(L)}$, where L is the last transformer layer.

Video encoder ϕ_{vid}. Given a video V with T frames, each frame $\mathbf{V}^f$ is mapped into patch embeddings $\{\mathbf{z}_{(f,p)}^{(0)}\}_{p=1}^{N_p}$, where N_p is the number of patches. A [CLS] token $\mathbf{z}_{[CLS]}^{(0)} \in \mathbb{R}^{d_{vid}}$ is prepended to the input sequence, i.e. $\mathbf{Z}_{vid}^{(0)} = (\mathbf{z}_{[CLS]}^{(0)}, \mathbf{z}_{(1,1)}^{(0)}, ..., \mathbf{z}_{(T,N_p)}^{(0)})^T \in \mathbb{R}^{(N_p T+1) \times d_{vid}}$. The video feature $\phi_{vid}(V) = \mathbf{z}_{[CLS]}^{(L)}$ is then extracted with the video encoder ϕ_{vid} with transformer mapping of Eq. (1). Since the $(N_p T+1)$ attention computations of each patch induce large memory overhead, there is a need to trade-off between space and time. VFMs typically use the TimeSformer [3] architecture, which utilizes two types of attention: a spatial attention block that only attends to features from the same time/frame (i.e. $(\mathbf{z}_{[CLS]}, \mathbf{z}_{(f,1)}, ..., \mathbf{z}_{(f,N_p)})$); and a time attention block that only attends to features from the same location (i.e., $(\mathbf{z}_{[CLS]}, \mathbf{z}_{(1,p)}, ..., \mathbf{z}_{(T,p)})$). The two attention blocks are interleaved to construct each transformer block of Eq. (1), reducing the attention computations per patch to (N_p+T+2). This enables increasing the sampling rate of each video (e.g., from 4 frames in [2] to 16 frames). Note that this is especially important for egocentric videos as camera tends to move faster.

Optimization. Following standard visual-language contrastive learning [31] practice, most VFMs [1, 2, 51] align the

two modalities by optimizing the InfoNCE loss [27]. Given a video dataset $\mathcal{D} = \{V_i, X_i\}_{i=1}^N$, the loss for each batch $\mathcal{B}$ is

$$\mathcal{L}_{cl} = -\frac{1}{|\mathcal{B}|} \left[\sum_{i \in \mathcal{B}} \log \frac{e^{(\mathbf{v}_i^T \mathbf{t}_i / \tau)}}{\sum_{j \in \mathcal{B}} e^{(\mathbf{v}_i^T \mathbf{t}_j / \tau)}} + \log \frac{e^{(\mathbf{t}_i^T \mathbf{v}_i / \tau)}}{\sum_{j \in \mathcal{B}} e^{(\mathbf{t}_i^T \mathbf{v}_j / \tau)}} \right], \tag{2}$$

where τ is the temperature, $\mathbf{v} = \phi_{vid}(V)$ and $\mathbf{t} = \phi_{txt}(X)$.

3.2. Generalization Ability of VFMs for Egocentric Video

Standard VFMs are trained from large datasets, typically collected on the web, which consist mostly of exocentric (third-person-view) videos. We refer to them as Exo-VFMs. Egocentric videos are captured from a first-person view by wearable devices, which have many hand-object interactions and motion blurs caused by head and body movements. This renders Exo-VFMs sub-optimal for egocentric video understanding [21]. The introduction of the large-scale egocentric video dataset Ego4D [9] sparked the development of egocentric VFMs (Ego-VFMs) [1,21,30,51]. However, while Ego4D is a very large dataset by egocentric video standards, it is unclear whether models learned from it can generalize to a broad set of egocentric video domains, namely to egocentric datasets other than Ego4D. To test this premise, we start by conducting some preliminary experiments on the generalization ability of Ego-VFMs to the popular Charades-Ego [35] dataset.

Table 1 compares the performance of Ego-VFM LaV-iLa [51] to that of two CLIP-based Exo-VFMs, X-CLIP [26] and Vita-CLIP [43], under the zero-shot setting. While the Ego-VFM has improved performance, likely because it mitigates the change of perspective, the results are not drastically superior. In fact, there is a considerable gap between the zero-shot performance of the Ego-VFM and that after it is fine-tuned to Charades-Ego. The advantage of the Exo-VFMs is that they basically expand CLIP into a VFM using prompting. Since the IFM is fixed, this is a lightweight operation. For these models, the adaptation to new egovideo datsets is not difficult. As shown in the right column of Table 1, when their prompts and additional modules (beyond CLIP) are fine-tuned on Charades-Ego, their performance outperforms the zero-shot application of the Ego-VFM. Hence, from a practical standpoint, the adaptation of the Exo-VFMs to a new egocentric setting is superior to the zero-shot application of the Ego-VFM. However, they still underperform the fine-tuning performance of the latter.

These observations suggest the following conclusions. First, Exo-VFMs like X-CLIP [26] and Vita-CLIP [43] are quite versatile and can cover large domain gaps with lightweight adaptation. However, they cannot match the performance of Ego-VFMs fine-tuned on the egovideo dataset of interest. Second, current Ego-VFMs appear to overfit to Ego4D dataset, requiring fine-tuning for effective performance on alternative egovideo datasets, like

Charades-Ego. However, this requires a large computation and parameter overhead, and reduces the practical value of VFMs, especially if fine-tuning is required for multiple tasks.

This raises the question of how to design lightweight adaptation modules for the Ego-VFM models that, similarly to the Exo-VFMs of Table 1, can mitigate the performance gap across egovideo domains with reduced memory and training. We focus on the LaViLa [51] model, since it is the current state-of-the-art Ego-VFM, where TimeSformer [3] is adopted as the video encoder to trade off the space-time resolutions. In view of the established efficacy of prompt-tuning for other large foundation models [14, 53], we explore prompt-tuning as a parameter-efficient way to adapt Ego-VFMs to downstream egocentric video applications.

4. Ego-VFM Prompt-tuning Baselines

Prompt-tuning introduces a set of learnable prompts, which are the only parameters optimized during the adaptation, leaving the rest of the VFM frozen. Specifically, the input sequence $\mathbf{Z}^{(0)}$ of the transformer of Eq. (1) is augmented with learnable prompts $\mathbf{P} = (\mathbf{p}_1, ..., \mathbf{p}_M) \in \mathbb{R}^{d \times M}$. While Exo-VFMs like X-CLIP [26] and Vita-CLIP [43] rely on prompt-tuning of IFMs, it is non-trivial to prompt-tune a TimeSformer-based VFM (such as LaViLa) as it utilizes divided space-time attention. We next introduce several baseline solutions, based on prompt-tuning methods in the literature. As shown in Figure 2, these insert prompts in the text encoder (TPT), video encoder (VPT), or both (VoP).

Text prompt-tuning (TPT). Given text embedding $\mathbf{Z}_{txt}^{(0)} \in \mathbb{R}^{(N_w+2) \times d_{txt}}$, TPT prepends text prompts $\mathbf{P}_t = (\mathbf{p}_{t,1}, ..., \mathbf{p}_{t,M_t}) \in \mathbb{R}^{d_{txt} \times M_t}$ to the input text emebeddings, i.e. $\tilde{\mathbf{Z}}_{txt}^{(0)} = (\mathbf{z}_{[SOS]}^{(0)}, \mathbf{P}_t, \mathbf{z}_1^{(0)}, ..., \mathbf{z}_{N_w}^{(0)}, \mathbf{z}_{[EOS]}^{(0)})^T$. The output feature $\phi_{txt}'(T) = \tilde{\mathbf{z}}_{[EOS]}^{(L)}$ is then extracted with the transformer in Eq. (1).

Video prompt-tuning (VPT). Given a video V, the video embedding of $\mathbf{Z}_{vid}^{(0)} \in \mathbb{R}^{(N_pT+1) \times d_{vid}}$ is extracted as described in section 3.1. To prompt-tune the video encoder, the visual prompts $\mathbf{P}_v = (\mathbf{p}_{v,1}, ..., \mathbf{p}_{v,M_v}) \in \mathbb{R}^{d_{vid} \times M_v}$ are prepended to the input sequence, i.e. $\tilde{\mathbf{Z}}_{vid}^{(0)} = (\mathbf{z}_{[CLS]}^{(0)}, \mathbf{P}_v, \mathbf{z}_{(1,1)}^{(0)}, ..., \mathbf{z}_{(T,N_p)}^{(0)})^T$. Note that we apply prompt tuning only to the spatial attention blocks of the TimeSformer [3] as we found that prompt-tuning both blocks has no additional gain (See Appendix). Inspired by [14], beyond introducing visual prompts at the input layer, we further prepend layer-specific prompts $\mathbf{P}_v^{(l)}$ to every transformer layer. In the following, we omit the layer l superscript of these deep prompts, for brevity. The output of VPT is then written as $\phi_{vid}'(V) = \tilde{\mathbf{z}}_{[CLS]}^{(L)}$.

Text-video prompt-tuning (VoP). Recently, VoP [13] expanded CLIP into a video model by prompt-tuning of both modalities. Several prompt-tuning variants are proposed in [13]. Vanilla VoP adopts both TPT and VPT,

and the visual prompts are shared across all frames. To propagate contextual information across frames, VoP^C uses a context modeling module (CMM) to generate frame-specific prompts $\{\mathbf{P}_v^1, ..., \mathbf{P}_v^T\}$ conditioned on the context information from other frames. To adapt it to LaViLa, while [13] adopts the [CLS] token from each frame-specific CLIP, we use $\mathbf{z}_f = AvgPool(\mathbf{z}_{(f,1)}, ..., \mathbf{z}_{(f,N_p)})$ as the contextual feature of frame f. In addition, we devise two types of prompting for the space attention block of the TimeSformer. For intra-frame attention, each patch token $\mathbf{z}_{(f,p)}$ can only attend to the prompts associated with frame f (i.e. $(\mathbf{z}_{[CLS]}, \mathbf{P}_v^f, \mathbf{z}_{(f,1)}, ..., \mathbf{z}_{(f,N_p)})$). For inter-frame attention, each patch token $\mathbf{z}_{(f,p)}$ can attend to all the prompts across frames (i.e. $(\mathbf{z}_{[CLS]}, \mathbf{P}_v^1, ..., \mathbf{P}_v^T, \mathbf{z}_{(f,1)}, ..., \mathbf{z}_{(f,N_p)})$). Following [13], we adopt *intra-frame/inter-frame* attention in the first K/last $L-K$ transformer layers and integrate this strategy with VoP^C as the VoP^{F+C} variant.

5. Ego-VPA

Our experiments (see Section 6.2) show that VoP^{F+C} prompting significantly improves the performance of the Ego-VFM on new egocentric datasets. However, this method still falls short in two aspects. First, the CMM module, which is itself a bi-directional LSTM network, still requires a substantial number of model parameters. Second, there is no connection between text and visual prompts, limiting knowledge transfer across modalities. We next propose a novel prompt-tuning technique for Ego-VFMs such as LaViLa, denoted as Ego-VPA, which achieves context fusion across frames (like CMM) and modalities in an extremely lightweight fashion. This is implemented by the proposed prompt synthesis scheme, using a shared prompt basis. Note that we keep the inter-frame attention layers (last $L-K$ layers) identical to VoP^{F+C} for fair comparison.

5.1. Video Prompt Synthesis

The main challenge for video prompting is how to design prompts that capture contextual connections across frames. Prompt design is almost trivial for stand-alone prompts, which are vectors with few parameters learned by back-propagation. However, once a prompt depends on other prompts, there is the need to learn the *functional dependence* between them. This can be done by learning a prompt synthesis network, e.g., a transformer that simultaneously generates prompts for image and text [49], or a recursive network, such as the LSTM of [13] to synthesize prompts recursively, which is more sensible for video. While small, these networks have many more parameters than the prompts themselves, and can sacrifice the lightweight nature of the adaptation. For example, in Table 2, both VoP^C and VoP^{F+C} require about 10% of the VFM parameters.

Frame-specific prompt synthesis. In this work, we seek a more efficient solution, inspired by the compression literature and illustrated in Figure 3. We assume that the prompt
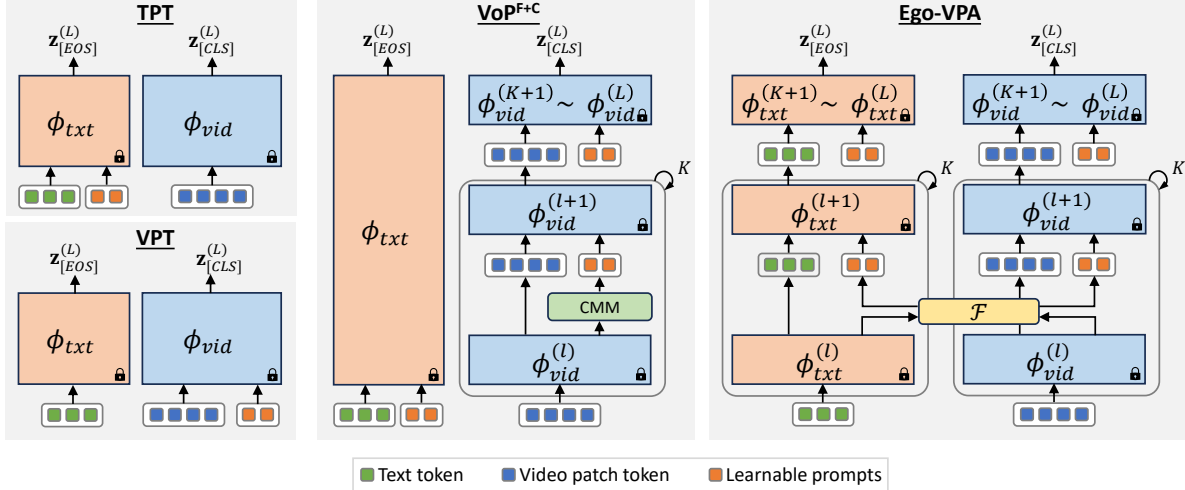
Figure 2. **Models.** We adapt SOTA prompt-tuning methods to Ego-VFMs (See section 4), i.e. TPT, VPT, and VoP^{F+C}, where CMM is a context modeling module. The proposed Ego-VPA leverages a set of basis prompts $\mathcal{F}$ for cross-modal prompt synthesis, enabling context modeling across frames and modalities in a highly efficient way (See section 5).

information lies on a lower dimensional latent space $\mathcal{H}$, onto which frame features $\mathbf{z}_f$ are mapped by an encoder $h_{vid}(\cdot)$: $\mathbb{R}^{d_{vid}} \to \mathcal{H} \subset \mathbb{R}^{d_f}$, where $d_f \leq d_{vid}$. Prompts then inhabit a B-dimensional subspace $\mathcal{P}$ of $\mathcal{H}$ ($B < d_f$) of orthogonal basis $\mathcal{F} = \{\mathbf{f}_1, ..., \mathbf{f}_B\}$. These vectors, denoted as *basis prompts*, can be thought of as the principal components of prompt space. Given a frame feature vector $\mathbf{z}_f$, we seek a small number (k) of basis prompts that best approximate $h_{vid}(\mathbf{z}_f)$. For this, we identify the k-dimensional subspace of $\mathcal{P}$ that has minimum least squares reconstruction error, i.e., the solution of

$$\alpha_f^* = \arg\min_{\alpha_f, ||\alpha_f||_0 = k} ||h_{vid}(\mathbf{z}_f) - \mathbf{F}\alpha_f||_2, \tag{3}$$

where $\mathbf{F} \in \mathbb{R}^{d_f \times B}$ contains the B basis prompts in $\mathcal{F}$ as columns and $||\alpha||_0$ is the 0-norm (number of non-zero elements) of α. We refer to this as the local reconstruction subspace for $h_{vid}(\mathbf{z}_f)$ and denote the set of vectors

$$\mathcal{S}_v^f = \{\mathbf{f}_i | \alpha_{f,i}^* \neq 0\} \tag{4}$$

as the *best local reconstruction basis* for $h_{vid}(\mathbf{z}_f)$.

The vector α_f^* has a closed form solution due to the orthogonality of the basis $\mathcal{F}$. For any matrix $\mathbf{G}$ containing k columns of $\mathbf{F}$, the minimizer of $||\mathbf{h} - \mathbf{G}\alpha||_2$ is

$$\alpha = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{h} = \mathbf{G}^T h_{vid}(\mathbf{z}_f), \tag{5}$$

since $\mathbf{G}$ is orthogonal, leading to the subspace distance

$$||\mathbf{h} - \mathbf{G}\alpha||_2 = (\mathbf{h} - \mathbf{G}\alpha)^T(\mathbf{h} - \mathbf{G}\alpha)$$
$$= ||\mathbf{h}||_2 - 2\mathbf{h}^T\mathbf{G}\alpha + ||\alpha||_2 = ||\mathbf{h}||_2 - ||\alpha||_2. \tag{6}$$

It follows that the solution of Eq. (3) is the subspace that maximizes the magnitude of the α vector. Since α has the form of Eq. (5), this occurs when α includes the largest k dot-products $h_{vid}(\mathbf{z}_f)^T\mathbf{f}_i$, i.e

$$\alpha_{f,i}^* = h_{vid}(\mathbf{z}_f)^T\mathbf{f}_i \times \mathbb{1}_{i \in s(\mathbf{z}_f, \mathcal{F}; k, h_{vid})} \tag{7}$$
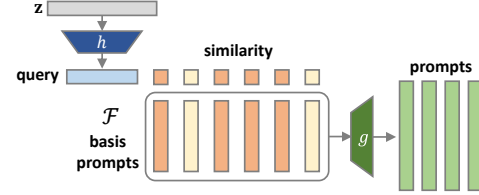

Figure 3. **Prompt Synthesis.** Token $\mathbf{z}$ is projected into the subspace by $h(\cdot)$, and sparsely approximated by the top-k similar prompts in the prompt basis $\mathcal{F}$, which are finally mapped into k prompts by the mapping $g(\cdot)$. (h, g) can be (h_{vid}, g_{vid}) or (h_{txt}, g_{txt}) for visual or text prompt generation, respectively. $k = 4$ in this illustration.

where $\mathbb{1}_{(\cdot)}$ is the indicator function,

$$s(\mathbf{z}, \mathcal{F}; k, h) = \text{top-}k(\{h(\mathbf{z})^T\mathbf{f}_1, ..., h(\mathbf{z})^T\mathbf{f}_B\}) \tag{8}$$

and top-k returns the indices of the largest k elements of its argument. Given α_f^*, the k basis-prompts in the best reconstruction basis $\mathcal{S}_v^f$ of Eq. (4) are mapped to feature space by a latent space decoder $g_{vid}(\cdot) : \mathbb{R}^{d_f} \to \mathbb{R}^{d_{vid}}$. This produces a set of prompts

$$\mathbf{P}_v^f = g_{vid}(\mathbf{F}\mathbf{A}_v^f) \tag{9}$$

where $\mathbf{A}_v^f \in \{0, 1\}^{B \times k}$ is a matrix whose i^{th} column is the one-hot code for the ith index in $s(\mathbf{z}_f, \mathcal{F}; k, h_{vid})$.

Prompt learning. The learning goal is to derive the basis $\mathcal{F}$ of the prompt space over the entire dataset. This is the set of vectors $\mathbf{f}_i$ that minimize the reconstruction error of Eq. (3) under the orthogonality constraint $\mathbf{f}_i^T\mathbf{f}_j = \mathbb{1}_{i=j}$. We ensure that all vectors have unit norm by introducing a normalization layer after the encoder $h_{vid}(\cdot)$ and learn the basis by minimizing the Lagrangian

$$\sum_{f=1}^{T} ||\sum_{\mathbf{f}_i \in \mathcal{S}_v^f} \alpha_{f,i}^*\mathbf{f}_i - h_{vid}(\mathbf{z}_f)||_2 + \sum_{\mathbf{f}_i, \mathbf{f}_j \in \mathcal{F}} \xi_i \mathbf{f}_i^T\mathbf{f}_{j \neq i}, \tag{10}$$

where ξ_i are Lagrange multipliers set to $\xi_i = 1, \forall i$ in all experiments. Since the basis is optimized on the entire dataset, the subspace spanned by $\mathcal{F}$ is a global low dimensional approximation to the space of prompts, akin to a principal component analysis (PCA) of components $\mathbf{f}_i$. Under this view, the approach can be seen as the computation of a localized PCA, which selects the k principal components that best approximate $h_{vid}(\mathbf{z}_f)$ to synthesize the prompts of Eq. (9).

Cross-frame context modeling. Since the frame-specific visual prompts $\mathbf{P}_v^f$ are conditioned on the frame feature $\mathbf{z}_f$, they encapsulate the context of frame f. Hence, cross-attention between $\mathbf{z}_{[CLS]}$ and the prompts $\mathbf{P}_v^1,...,\mathbf{P}_v^T$ can summarize knowledge across frames, without requiring additional modules like the bi-directional LSTM of CMM [13]. The hyper-parameters d_f, B, and k trade-off the number of parameters required to store the prompt basis, with the reconstruction error of the local least squares approximation, and the desired number of prompts to be added per transformer stage. In any case, because the basis $\mathcal{F}$ and the encoder/decoder pair h_{vid}, g_{vid}, are the only learned parameters, the approach is very efficient. In all our experiments, both h_{vid} and g_{vid} are implemented with a single linear layer of $d_{vid} \times d_f$ parameters. Hence, the proposed approach only requires $d_f(B + 2d_{vid})$ parameters, which is usually much smaller than the $(16 + 2M_vT)d_{vid}^2$ additional parameters of CMM. Please refer to the appendix for a detailed comparison. In Table 2, we show that good adaptation performance can be obtained with only 0.84% of the VFM parameters.

5.2. Cross-modal Prompt Synthesis

While the text description and image frames originate from two modalities, the underlying semantic context should be shared, since both refer to the same video event (e.g. "cut a tomato"). Intuitively, visual content in the video can benefit text domain features and vice versa. To enable knowledge transfer between the two modalities, we propose cross-modal prompt synthesis, where the prompt basis $\mathcal{F}$ of section 5.1 is shared across modalities, as shown in Figure 4.

Similar to video prompt synthesis, a text feature $\mathbf{z}_t = \mathbf{z}_{[EOS]}$ is mapped into prompt space by a text encoder $h_{txt}(\cdot) : \mathbb{R}^{d_{txt}} \to \mathbb{R}^{d_f}$, $h_{txt}(\mathbf{z}_t)$ is used to query the prompt basis $\mathcal{F}$ using $s(\mathbf{z}_t, \mathcal{F}; k, h_{txt})$ in Eq. (8), and a prompt set produced with $\mathbf{P}_t = g_{txt}(\mathbf{F}\mathbf{A}_t)$, where $\mathbf{A}_t \in \{0,1\}^{B \times k}$ is a matrix whose i^{th} column is the one-hot code for the ith index in $s(\mathbf{z}_t, \mathcal{F}; k, h_{txt})$. $g_{txt}(\cdot) : \mathbb{R}^{d_f} \to \mathbb{R}^{d_{txt}}$ is an additional prompt generator that maps basis vectors into text prompts. The projection functions h_{vid}, h_{txt} and joint basis $\mathcal{F}$ are jointly optimized with a loss

$$\mathcal{L}_{syn} = \frac{1}{|\mathcal{B}|}\sum_{i \in \mathcal{B}} l_s(V_i, X_i) \tag{11}$$

where

$$l_s(V,X) = \sum_{f=1}^{T}\|\sum_{\mathbf{f}_i \in \mathcal{S}_v^f} \alpha_{f,i}^* \mathbf{f}_i - h_{vid}(\mathbf{z}_f)\|_2$$
$$+\|\sum_{\mathbf{f}_i \in \mathcal{S}_t} \alpha_{t,i}^* \mathbf{f}_i - h_{txt}(\mathbf{z}_t)\|_2 + \sum_{\mathbf{f}_i, \mathbf{f}_j \in \mathcal{F}} \mathbf{f}_i^T \mathbf{f}_{j \neq i}. \tag{12}$$
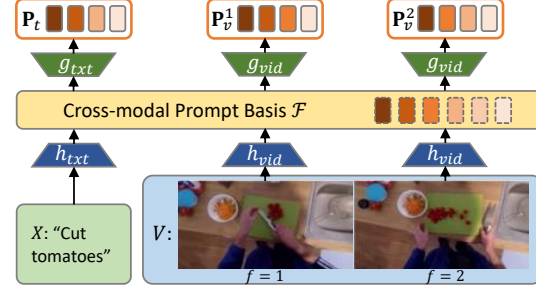

Figure 4. **Cross-modal Prompt Synthesis.** The basis prompts $\mathcal{F}$ are shared across frames and modalities, but different mapping functions h, g are adopted per modality to synthesize the prompts.

with $\alpha_{t,i}^* = h_{txt}(\mathbf{z}_t)^T \mathbf{f}_i \times \mathbb{1}_{i \in s(\mathbf{z}_t, \mathcal{F}; k, h_{txt})}$, similar to Eq. (10).

5.3. Training

The prompt basis $\mathcal{F}$ and projections h_{vis}, h_{txt}, g_{vis} and g_{txt}, are jointly optimized using a combination

$$\mathcal{L} = \mathcal{L}_{cl} + \lambda \mathcal{L}_{syn}, \tag{13}$$

of the contrastive loss of Eq. (2) and the cross-modal prompt synthesis loss of Eq. (11), where λ is a hyperparameter. However, since the prompt basis is randomly initialized, the basis prompts do not contain semantic information in the early stages of training. This problem is compounded by the use of top-k basis prompt selection, where only the basis prompts will be updated, and these updated basis prompts will then be selected again in the next iteration. To prevent this, instead of using the top-k selection rule during training, we sample k basis prompts from a multinomial distribution π_m, which is a mixture

$$\pi_m = \gamma \pi_{sim} + (1-\gamma)\pi_{invf}, \tag{14}$$

of the distribution π_{sim} of similarities between query and basis prompts, and the inverse of the basis prompt selection frequency π_{invf}. The mixture coefficient γ is set to 0 at the beginning of training and gradually increased to 1 in later epochs. This increases the possibility that basis prompts rarely seen in the dataset are learned. Note that the top-k basis prompts are always selected during inference.

6. Experiments

In this section, we validate the efficiency and effectiveness of Ego-VPA by comparing it with SOTA methods and ablating on different components. More results in the appendix.

6.1. Experimental Setup

Datasets. We adopt LaViLa [51] as the Ego-VFM, which is pretrained on Ego4D [9], containing 4M video-text pairs for egocentric videos. Both the proposed method and baselines are evaluated on Charades-Ego [35], EGTEA [20], and EPIC-Kitchens-100 [7], which are widely used in egocentric video research. Charades-Ego [35] is a fine-grained action

Method	Tunable Params (%)	Charades-Ego mAP	EGTEA Mean Acc	EGTEA Top-1 Acc
Zero-shot	0%	26.8	28.90	35.51
Full fine-tuning	100%	32.9/33.7*	67.77	71.37
Bias [5,48]	0.12%	30.0	55.29	61.52
TPT [53]	0.002%	29.7	51.93	58.36
VPT [14]	0.66%	31.7	63.11	68.35
VoP [13]	0.67%	32.5	66.36	70.72
VoP^C [13]	10.64%	32.4	67.55	71.91
VoP^{F+C} [13]	10.86%	32.7	68.70	73.24
Ego-VPA (Ours)	0.84%	**33.8**	**69.17**	**73.39**

Table 2. Results on Charades-Ego and EGTEA compared to state-of-the-art prompt-tuning methods introduced in section 4. * denotes the number in [51], using $4\times$ batch size compared to ours.

	Prompt Generation	Cross Modality	Orthogonality Constraint	Prompt Query	mAP
(m1)	CMM		N/A	N/A	32.7
(m2)	PS			$\sim\pi_m$	32.8
(m3)	PS		✓	$\sim\pi_m$	33.0
(m4)	PS	✓		$\sim\pi_m$	33.3
(m5)	PS	✓	✓	$\sim\pi_m$	**33.8**
(m6)	PS	✓		top-k	32.8
(m7)	PS	✓	✓	top-k	33.5

Table 3. Ablations on prompt generation, the orthogonality constraint imposed by the 2nd term of Eq. (10), and the prompt query (PS: prompt synthesis; $\sim\pi_m$: sampling Eq. (14)). (m1) is VoP^{F+C} [13]; (m2-7) are Ego-VPA variants, and (m5) is full Ego-VPA.

classification dataset containing 33, 114 trimmed action segments for training, spanning across 157 classes. While both egocentric and exocentric views are provided, we only train and evaluate egocentric videos using the official splits, as in [21, 51]. EGTEA [20] is an egocentric cooking video dataset, containing 10,321 action instances from 106 fine-grained classes. We only use the video data and follow the same protocol for training and evaluation as in [51]. EPIC-Kitchens-100 [7] is an egocentric cooking video dataset of 100 hours, containing 67,217/9,668 clips for training/validation. We evaluate the multi-instance retrieval (MIR) task to test the generalization of Ego-VPA, which contains a text-to-video (T->V) and video-to-text (V->T) retrieval task.

Metrics. At inference, video and text features are extracted with ϕ_{vid} and ϕ_{txt} respectively. For action classification, we compute the cosine similarity per video between the video feature and the text feature of each action class as the classification score. We report commonly used metrics for these datasets. Since each testing video in Charades-Ego is multi-label, we report the mean average precision (mAP) over the 157 classes. For EGTEA, we report top-1 accuracy (Top-1 Acc) and average accuracy over all classes (Mean Acc). For the EPIC-Kitchens-100 MIR task, mean Average Precision (mAP) and Normalized Discounted Cumulative Gain (nDCG) for T->V and V->T retrieval are reported.

Model architecture. The Ego-VFM architecture is inherited from LaViLa [51], where the text encoder ϕ_{txt} is a 12-layer ($L=12$) Transformers with $d_{txt}=512$ and the video encoder ϕ_{vid} a 12-layer ($L=12$) TimeSformer with $d_{vid}=768$. We prompt the model with $M_v = M_t = 8$, $K=8$, setting $d_f=512, B=10$ for the prompt basis. All models are trained with 16 frames per video ($T=16$) unless explicitly noted. More details on implementation are reported in the appendix.

Computing. As in section 4, we freeze the Ego-VFM model parameters and only learn a small portion of prompts on the downstream datasets. Ego-VPA does not require large computing resources compared to prior work. We train all the experiments with a batch size of 4 per GPU using 8 NVIDIA Titan Xp GPUs. This is $4\times$ smaller than the batch size in [51]. The memory required for Ego-VPA training is around 2/3 of that required to fine-tune the full model.

6.2. Comparisons to SOTA Prompt-tuning Methods

Table 2 presents the result on Charades-Ego and EGTEA, as a function of the tunable model parameters. The poor zero-shot performance shows that current Ego-VFMs, like LaViLa, are somewhat overfitted to the Ego4D dataset. Full fine-tuning significantly improves performance, but requires optimization of the entire VFM, which is inefficient. Prompt-tuning methods require less parameter optimization. Bias [5,48] only fine-tunes the bias terms in the model but not the weights, and thus has limited adaptation capacity. TPT [53] and VPT [14] extend the input sequence with learnable prompts for the text and video encoders, respectively. These methods are weaker than VoP [13], where prompt-tuning is performed for both encoders. VoP^{F+C}, a variant of VoP using a CMM module and frame-aware attention layers, is the best-performing baseline. However, the introduction of the CMM module significantly increases the parameter counts, requiring around 10% of the model size for the adaptation. The proposed Ego-VPA consistently outperforms all other methods on both datasets, even beating full fine-tuning, with only 0.84% trainable parameters. While the adaptation to EGTEA produces much stronger results, indicating that the domain gap to Ego4D is smaller, the zero-shot performance of LaViLa is still quite weak, reinforcing the importance of efficient adaptation methods.

6.3. Ablation Studies

This section ablates different designs of Ego-VPA with the Charades-Ego dataset, unless explicitly noted.

Context modeling module. We first compare the proposed prompt synthesis (PS) with the CMM module of VoP^{F+C} [13]. As shown in Table 3, despite using much fewer trainable parameters, video-only prompt synthesis (m3) outperforms the CMM in VoP^{F+C} (m1). When cross-modal knowledge transfer is enabled with cross-modal prompt synthesis (m5), the performance further improves largely. This validates the effectiveness of context modeling across frames and modalities.

Orthogonality constraint. The closed-from solution of Eq. (5) only holds if the basis-prompts in $\mathcal{F}$ are orthogonal. The orthogonality constraint in the 2nd term of Eq. (10) is important to guarantee this property. This is validated by (m2-3), (m4-5), and (m6-7) in Table 3, where a gain is
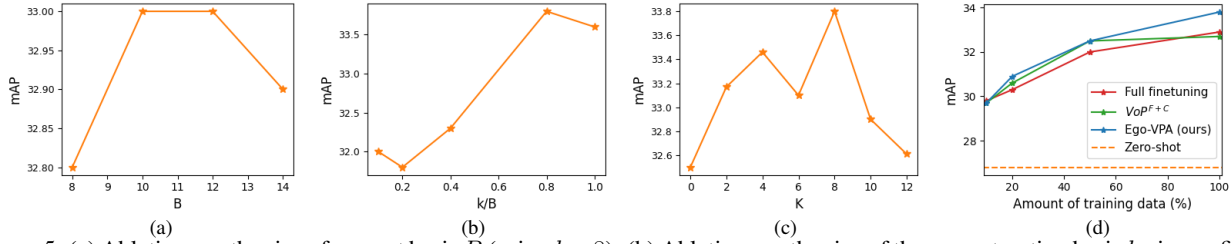
Figure 5. (a) Ablations on the size of prompt basis B (using $k=8$). (b) Ablations on the size of the reconstruction basis k given a fixed B. (c) Ablations on *intra/inter-frame* attention boundary K. (d) Ablations on different amounts of training data.

consistently observed when the orthogonality constraint is imposed, especially in the cross-modal setting.

Prompt query strategies. Table 3 compares the pure top-k selection rule with the sampling strategy of Eq. (14) (m4-7). The latter encourages every basis prompt in $\mathcal{F}$ to be updated in the early training stages. In the later stages, as γ approaches 1, it reduces to top-k selection. (m4-7) shows that the strategy of Eq. (14) allows more distributed feature updates and improves performance.

Size of prompt basis and reconstruction basis. In Figure 5a, we ablate the size of prompt basis B when $k=8$ for video prompt synthesis. Note that we use 8 visual prompts per-frame (i.e. $M_v = 8$), so $B = 8$ means that all basis prompts are used to generate prompts. Increasing B allows a slight improvement since some features may not co-exist across frames. However, the performance saturates quickly, suggesting that the prompt space has a very low dimension. Conversely, we ablate the size of reconstruction basis k given a fixed B for cross-modal prompt synthesis in Figure 5b. In general, increasing k enhances the expressiveness of the projected features, since more basis prompts are used for feature recovery. However, there is a benefit to the local subspace approximation, as best results are usually obtained for $k/B < 1$. In Figure 5b, the optimal ratio is 0.8.

Intra/inter-frame attention boundary. As discussed in section 5.1, we adopt *intra-frame* attention in the first K layers of the video encoder and *inter-frame* attention in the remaining, allowing shallow layers to adapt lower-level features, and deeper layers to fuse high-level semantics. We ablate such attention boundary by training the model with different values of K, as shown in Figure 5c. Note that both visual and text encoders have 12 layers (i.e. $L = 12$). Results show that setting $K = 8$ leads to the best performance, coherent to the observation of [13].

Number of frames. We validate the robustness of Ego-VPA by training with different numbers of frames on both Charades-Ego and EGTEA. As shown in Table 4, the performance of all methods increases with the number of frames, indicating that temporal resolution is an important factor for video understanding. Ego-VPA consistently improves over VoP^{F+C} and the full fine-tuning performances, showing that it is robust and effective across time resolutions.

Amount of training data. To evaluate the proposed

Dataset	Charades-Ego (mAP)			EGTEA (mAcc)		
# of frames	4	8	16	4	8	16
Zero-shot	24.4	26.0	26.8	27.0	28.5	28.9
Full fine-tuning	28.3	31.4	32.9	56.1	62.9	67.8
VoP^{F+C} [13]	28.6	30.9	32.7	59.1	63.6	68.7
Ego-VPA (ours)	**29.3**	**31.5**	**33.8**	**60.5**	**64.4**	**69.2**

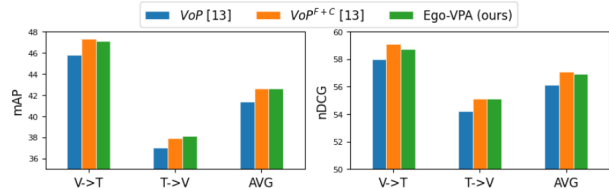Table 4. Ablations on using different numbers of frames per video.

Figure 6. Ego-VPA can generalize to EPIC-Kitchens-100 multi-instance retrieval task.

Ego-VPA on the low-data regime, we adapt the models with 10%, 20%, 50%, and 100% data respectively. As shown in Figure 5d, both Ego-VPA and the SOTA VoP^{F+C} are effective for various amounts of training data, reaching comparable or even superior performance than full fine-tuning and improving largely over the zero-shot results. However, the SOTA model still underperforms Ego-VPA in general.

6.4. Generalization to Retrieval Tasks

We further evaluate multi-instance retrieval tasks [7] on Epic-Kitchens-100. Figure 6 shows that Ego-VPA performs on par with VoP^{F+C}, which has 10% more trainable parameters. Compared to vanilla VoP, which has a similar number of parameters, Ego-VPA has clearly better performance. This shows that Ego-VPA is more parameter-efficient and effective across different egocentric video tasks.

7. Conclusions

We propose Ego-VPA, a novel parameter-efficient adaptation method for Ego-VFMs. Atop a frozen Ego-VFM, we sparsely approximate projected video frame/text features with a shared prompt basis and synthesize video/text prompts accordingly. This is shown to enhance context fusion across frames and cross-modal transfer, achieving improved visual-language alignments. Through extensive experiments, we show that Ego-VPA is both efficient and effective, outperforming SOTA methods with much fewer learnable parameters.

References

[1] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23066–23078, 2023. 1, 2, 3

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 1, 2, 3

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021. 3, 4

[4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 1

[5] Han Cai, Chuang Gan, Ligeng Zhu, and Song Han. Tinytl: Reduce memory, not parameters for efficient on-device learning. *Advances in Neural Information Processing Systems*, 33:11285–11297, 2020. 7

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022. 2, 6, 7, 8

[8] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Jiao Qiao. Clip-adapter: Better vision-language models with feature adapters. *ArXiv*, abs/2110.04544, 2021. 1

[9] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2, 3, 6

[10] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056, 2018. 1

[11] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR, 09–15 Jun 2019. 2

[12] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 2

[13] Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, and Donglin Wang. Vop: Text-video co-operative prompt tuning for cross-modal retrieval. In *CVPR*, 2023. 1, 2, 4, 6, 7, 8

[14] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 1, 2, 4, 7

[15] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2

[16] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 2

[17] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7331–7341, 2021. 2

[18] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23119–23129, 2023. 2

[19] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 2

[20] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, pages 619–635, 2018. 2, 6, 7

[21] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *arXiv preprint arXiv:2206.01670*, 2022. 1, 2, 3, 7

[22] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *European Conference on Computer Vision*, pages 388–404. Springer, 2022. 2

[23] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 2

[24] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, pages 9879–9889, 2020. 1, 2

[25] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019. 1, 2

[26] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022. 1, 2, 3, 4

[27] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 3

[28] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477, 2022. 1, 2

[29] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. Adapterhub: A framework for adapting transformers. *arXiv preprint arXiv:2007.07779*, 2020. 2

[30] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023. 1, 2, 3

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 1, 2, 3

[32] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6545–6554, 2023. 2

[33] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968, 2022. 1, 2

[34] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020. 2

[35] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. 2, 3, 6

[36] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[37] Asa Cooper Stickland and Iain Murray. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5986–5995. PMLR, 09–15 Jun 2019. 2

[38] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019. 1, 2

[39] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237, 2022. 1, 2

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[41] Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. All in one: Exploring unified video-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608, 2023. 1, 2, 3

[42] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 2

[43] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Vita-clip: Video and text adaptive clip via multimodal prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23034–23044, 2023. 1, 2, 3, 4

[44] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022. 1

[45] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 1, 2

[46] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022. 2

[47] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10714–10726, 2023. 1, 2

[48] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based

masked language-models. *arXiv preprint arXiv:2106.10199*, 2021. 7

[49] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022. 4

[50] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021. 2

[51] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023. 1, 2, 3, 4, 6, 7

[52] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 1, 2

[53] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022. 1, 2, 4, 7