Eliciting In-Context Learning in Vision-Language Models for Videos Through Curated Data Distributional Properties

Anonymous ACL submission

Abstract

A major reason behind the recent success of large language models (LLMs) is their *in-context learning* capability, which makes it possible to rapidly adapt them to downstream text-based tasks by prompting them with a small number of relevant demonstrations. While large vision-language models (VLMs) have recently been developed for tasks requiring both text and images, they largely lack in-context learning over visual information, especially in understanding and generating text about videos. In this work, we implement **E**mergent **I**n-context **Le**arning on **V**ideos (**EILeV**), a novel training paradigm that induces in-context learning over video and text by capturing key properties of pre-training data found by prior work to be essential for in-context learning in transformers. In our experiments, we show that **EILeV**-trained models outperform other off-the-shelf VLMs in few-shot video narration for novel, rare actions. Furthermore, we demonstrate that these key properties of bursty distributions, skewed marginal distributions, and dynamic meaning each contribute to varying degrees to VLMs' in-context learning capability in narrating procedural videos. Our results, analysis, and **EILeV**-trained models yield numerous insights about the emergence of in-context learning over video and text, creating a foundation for future work to optimize and scale VLMs for open-domain video understanding and reasoning.

1 Introduction

In recent years, the advent of transformer-based (Vaswani et al., 2017) large language models (LLMs) has garnered significant attention in and beyond the AI research community. A central reason for this is their *in-context learning* capability (Brown et al., 2020), which makes it possible to rapidly adapt LLMs to novel tasks by simply prompting them with a few demonstrations. This capability removes the need for the expensive and arduous task-specific fine-tuning required by earlier language modeling approaches.

While in-context learning has been extensively studied and utilized in purely text-based problems in language understanding, reasoning, and generation, there are myriad potential applications for this rapid post-deployment adaptation in processing *video*. For example, in embodied and task-oriented AI, a major challenge is to recognize novel, rare human actions from video that cannot possibly be completely covered in training data (Perrett et al., 2023; Du et al., 2023; Bao et al., 2023). A vision-language model (VLM) capable of in-context learning over video could address this challenge, as it would only require a few related videos of actions as few-shot, in-context examples to recognize and reason about these novel, rare actions. However, while large VLMs for jointly processing text and images have been developed (Li et al., 2022, 2023c; Dai et al., 2023; Zhu et al., 2023a; Peng et al., 2023; Liu et al., 2023), they are typically not optimized for reasoning over multiple images (i.e., frames), crucial for understanding videos. Meanwhile, a handful of open-source VLMs have recently been developed for video understanding (Zellers et al., 2022; Li et al., 2023b; Zhang et al., 2023; Lin et al., 2023), but they lack in-context learning.

In-context learning in text-only, transformer-based LLMs was initially observed to improve with increased model size, along with the size and diversity of training data (Brown et al., 2020). Later, Chan et al. (2022) identified several distributional properties of the training data as causes for this emergent behavior in transformer-based models: (1) bursty distributions with entities that tend to appear in clusters, (2) skewed marginal distributions with a long tail of infrequent items, and (3) dynamic meaning with label multiplicity. However, as their experiments relied on small transformer-based models trained on synthetic image classification data, it remains unclear whether their findings hold true for VLMs trained on video and text at scale.

In this work, we address this question by conducting systematic empirical experiments to investigate whether these training data distributional properties also elicit in-context learning capabilities in VLMs for video. Specifically, we use various text annotations from Ego4D (Grauman et al., 2022), a popular video dataset, to implement **E**mergent **I**n-context **Le**arning on **V**ideos (**EILeV**), a novel VLM training method that satisfies all three properties and successfully elicits in-context learning over video and text. In our experiments, we observe that the **EILeV**-trained models outperform other off-the-shelf VLMs in few-shot video narration on rare actions, and that, through careful ablation studies, each property indeed contributes to this in-context learning capability. Furthermore, our analysis yields a host of new insights around the importance of each property in in-context learning for video.

The contributions of our work are as follows: (1) we propose **EILeV**, a novel training method that can elicit in-context learning capabilities in VLMs for video and text, (2) we validate through systematic ablation experiments that the same data distributional properties that elicit in-context learning in small transformer-based models also apply to VLMs for videos, and (3) we release a set of **EILeV**-trained VLMs with in-context learning capabilities optimized for egocentric videos.

2 Related Work

2.1 In-Context Learning

Brown et al. (2020) discovered in-context learning in LLMs when creating GPT-3. This was a significant departure from fine-tuning which involves parameter updates to adapt LLMs to downstream tasks. Instead, in-context learning enables LLMs to be adapted without parameter updates by prompting them with a few examples of a task as part of the input context for text generation. The size of the model and training data were thought to be key to training a model with in-context learning capabilities.

More recently, there has been more research on the exact causes of in-context learning. Min et al. (2022) proposed MetaICL, a meta-training framework to elicit in-context learning capabilities in text-only language models. MetaICL conditions each example with related in-context examples during training. Chan et al. (2022) investigated the distributional properties of training data for in-context learning. Their findings showed that there are certain properties that encourage in-context learning in transformer-based models, and massive textual data from the web used to train LLMs naturally have those properties. Furthermore, Reddy (2023) found that in-context learning is driven by the abrupt emergence of an induction head. There have also been works with findings about in-context learning in VLMs. Notably, training large generative VLMs with image-text interleaved data has been shown to be an effective technique to improve model performance, especially in tasks involving in-context learning (Alayrac et al., 2022; McKinzie et al., 2024; Wang et al., 2024). Our work combines these insights from prior work around the cause of in-context learning to propose a new VLM training paradigm for video and text, and carefully investigates how they contribute to in-context learning.

2.2 Vision-Language Models (VLMs)

With the recent success of text-only LLMs, there have been various efforts to replicate their success in multimodal settings, especially vision and language. Two different types of approaches in training generative VLMs have been proposed. The first is to train them from scratch using large text and paired image and text datasets (Hao et al., 2022; Huang et al., 2024; Peng et al., 2023; Lu et al., 2023). This approach allows the most controllability and flexibility as the resulting VLM is not dependent on other pre-trained models that may have undesirable behaviors, but it requires a massive amount of compute and data. In order to address these challenges, a number of approaches have been proposed to create VLMs by learning a mapping from a frozen pre-trained vision encoder to the input space of a frozen pre-trained LLM (Alayrac et al., 2022; Li et al., 2023b; Zhao et al., 2023; Li et al., 2022, 2023c; Dai et al., 2023; Liu et al., 2023; Zhang et al., 2023; Lin et al., 2023; Yang et al., 2022; Li et al., 2023d; Zhu et al., 2023a; Laurençon et al., 2023; Maaz et al., 2023; Ye et al., 2023; Gong et al., 2023; Zhang et al., 2024).

Some of these approaches enable the resulting VLMs to process videos by representing them as sequences of still frames; however, only Flamingo (Alayrac et al., 2022), Otter (Li et al., 2023b) and Kosmos-2 (Peng et al., 2023) support in-context learning over video and text as a by-product of their large-scale pre-training. In this work, we conduct thorough investigation of how key properties of training data achieve in-context learning beyond just as a by-product of large-scale training.

3 Three Distributional Properties for In-Context Learning

Since Brown et al. (2020) discovered in-context learning in text-only LLMs, there has been much research into the cause for in-context learning. In particular, Chan et al. (2022) found that three characteristics of the training data are important in eliciting in-context learning in transformer-based models, each of which is abundant in both natural language and video data: *bursty distributions*, *skewed marginal distributions*, and *dynamic meaning*.

Bursty Distributions In-context learning relies on data where entities appear in clusters, or non-uniformly depending on the context. Groups of related entities may be mentioned frequently in some contexts, but much more rarely in other contexts.

Skewed Marginal Distributions In-context learning also relies on data of skewed marginal distributions with a long tail of infrequent items (i.e., a Zipfian distribution). This phenomenon is a long-standing challenge in representing language and images, and has long been observed in text, image, and video datasets collected for research.

Dynamic Meaning Lastly, in-context learning relies on dynamic meaning, where a single entity can have multiple possible interpretations, and multiple entities can map to the same interpretation. In natural language, we observe this property in word senses, homonyms, and synonyms. In the visual world, a particular object may be described in multiple valid ways, e.g., synonyms, physical properties, and hypernyms. Meanwhile, many distinct objects may be grouped based on various descriptors.

4 Problem & Methods

In this section, we first introduce the target problem and dataset for our evaluations of in-context learning. Next, we introduce **EILeV**, our training paradigm which captures all three distributional properties thought to elicit in-context learning, as well as the ablations we use to validate the importance of each property in enabling in-context learning over video and text. We then introduce the model architecture we apply this paradigm to, and lastly discuss how we evaluate the in-context learning capability of VLMs trained on video and text.

4.1 Problem Definition

We target the task of *few-shot video narration* using the Ego4D dataset (Grauman et al., 2022).

Few-Shot Video Narration *Video narration* is a captioning task where given a video, a system must generate a text description of the events occurring in the video. Here, *few-shot video narration* refers to the implementation of this task where a VLM (pre-trained on large-scale video and text data) is conditioned with one or more example videos and narrations before being prompted to generate a narration for a held-out video clip. If conditioning such a VLM on several example videos and narrations improves the quality of narration, this implies that the VLM is indeed capable of in-context learning over video and text.

Ego4D Ego4D is a popular large-scale dataset of egocentric videos that have been densely annotated with human-written English narrations, ideal for our task. Beyond narrations, the dataset includes higher-level class labels for the verbs and nouns associated with each narrated video clip. These annotations enable systematic ablations for all three distributional properties of training data discovered by Chan et al. (2022) to facilitate in-context learning, enabling a systematic study of in-context learning over video and text in VLMs. These ablations are introduced in Section 4.2.

4.2 Training Paradigm & Ablations

Using Ego4D's "Forecasting Hands & Objects Master File", we construct a dataset of interleaved text and video that satisfies these properties, and use it to train and evaluate VLMs. We call this training procedure **E**mergent **I**n-context **Le**arning on **V**ideos (**EILeV**). **EILeV** uses the video and text data provided by Ego4D to implement all three distributional properties necessary for in-context learning: bursty distributions, skewed marginal distributions, and dynamic meaning. To demonstrate the importance of each distributional property captured in **EILeV**, we use Ego4D's detailed annotations to carefully ablate each property as illustrated in Figure 1.

For all experiments, each training data point consists of a *context* with 16 video-narration pairs, and a *query* with a single video-narration pair. We

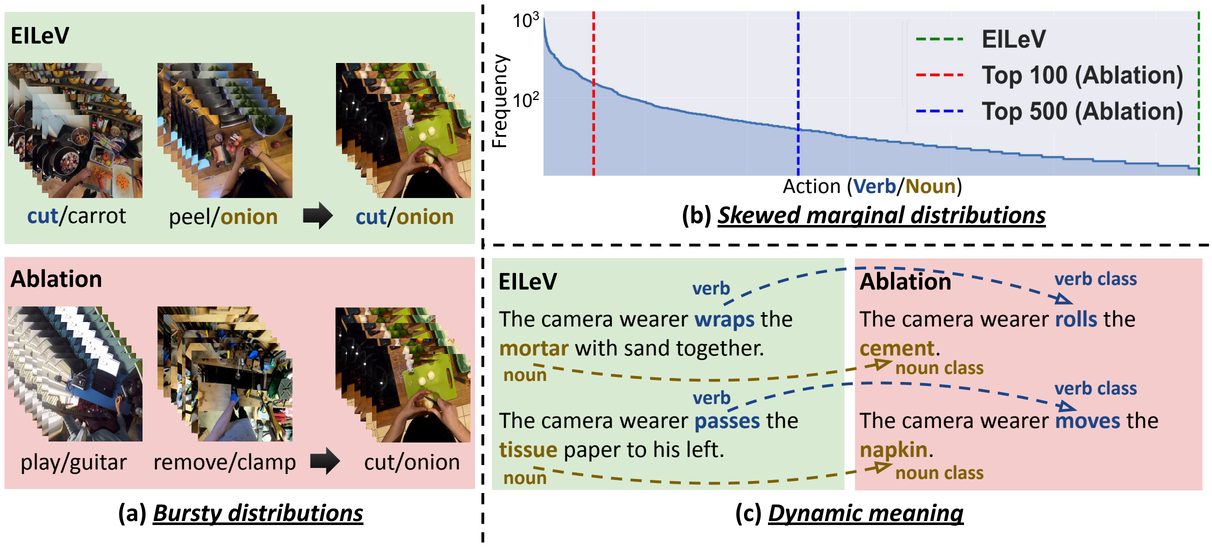

Figure 1: In our proposed training procedure **EILeV**, we ensure that the training data satisfy the following three properties: (a) bursty distributions, (b) skewed marginal distributions, and (c) dynamic meanings. Then, we ablate each property to demonstrate its importance. We ablate property (a) by randomly sampling in-context examples; we ablate property (b) by varying the number of common actions in the training data; we ablate property (c) by canonicalizing verbs and nouns using their corresponding verb and noun classes.

convert the action narrations into question-answer pairs where the narrations are the answers, e.g., e.g., *What is the camera wearer doing? The camera wearer cuts a carrot*. We vary the syntactic form of questions using a set of templates (Appendix C). The training objective is to maximize the likelihood of the sequence of tokens in the ground-truth action narration, conditioned on the context and video clip from the query.

Next, we discuss how each distributional property was incorporated and ablated in **EILeV**.

Bursty Distributions In order to implement bursty distributions in **EILeV**, we take advantage of the annotations in Ego4D, where each video clip is annotated with a verb class and a noun class based on the main action portrayed in the clip. Specifically, we sample video clips and action narrations that share the same verb class as the query for half of the context, and we sample those with the same noun class for the other half. We further ensure that none of the sampled video clips and action narrations match both the verb class and noun class of the query simultaneously. This ensures that the context, while comprising a "burst" of similar concepts, only provides partial information regarding the query. This property can then be ablated by randomly sampling video clips and action narrations without regard to their verb and noun classes. Figure 1 (a) illustrates the two sampling strategies. We can measure the impact of bursty distributions by training VLMs with each type of context and comparing their in-context learning capabilities.

Skewed Marginal Distributions Like most natural datasets, Ego4D's verb and noun class labels have a skewed marginal distribution with a long tail of verb-noun pairs, making it ideal for our study. To study how the skewed marginal distributions of training data affect the in-context learning capability of trained models, we first use the verb and noun class annotations from Ego4D to designate the most frequent 80% verb-noun pairs as *common actions* for training, and the remaining 20% as *rare actions* only for evaluation. It is important to note that while none of the rare actions are part of the common action training data, they may still share either verb or noun classes with common actions. For example, if the training data contain common actions (*put*, *key*) and (*sit*, *bench*), there may exist a rare action (*put*, *bench*) in the evaluation data.

To measure how the skewness of marginal distributions in the training data impacts models' capability to generalize to these novel held-out actions, we then vary the number of common actions in the training data through three experiments. Specifically, we construct a training dataset with only the top 100 common actions (little skewness without a long tail of infrequent actions), one with the top 500 common actions (moderate skewness with a short tail of infrequent actions) and another with all the common actions (highly skewed with a long tail of infrequent items). We uniformly upsample the datasets with top 100 and top 500 common actions to keep all three training datasets to be the same size. Figure 1 (b) shows how these training datasets with different marginal distributions are constructed. Given these curated training datasets, we can measure the impact of the skewness of the marginal distributions of the training data on trained models' in-context learning capability.

Dynamic Meaning For dynamic meaning, we rely on the fact that Ego4D's natural language action narrations contain words of multiple senses, homonyms, and synonyms. To ablate this dynamic meaning property in **EILeV**, we canonicalize verbs and their corresponding objects in the action narrations. Specifically, we prompt an LLM (Llama-2-Chat 7B; Touvron et al., 2023) to replace the verb and its corresponding object of each action narration with their verb and noun class. Figure 1 (c) shows the canonicalization process. We can then measure the impact of dynamic meaning by comparing the in-context learning capability of VLMs trained on data with and without this property.

4.3 Model

To experiment with **EILeV** as discussed above, we adopt a VLM architecture capable of processing sequential data interleaved with both video clips and texts, making it possible to infer patterns and relationships among them and thus support the emergence of in-context learning over them. We initialize our model with BLIP-2 (Li et al., 2023c), a VLM created by learning a transformer-based projection (called a querying transformer or Q-Former) from a frozen pre-trained vision encoder into the input space of a frozen LLM. Since BLIP-2's original implementation is not able to handle data interleaved with video clips and texts, we follow Hao et al. (2022) to perform simple modifications to enable its frozen language model to serve as a universal interface for video clips and texts.[1] Specifically, we first encode all the video clips by independently encoding sampled frames with BLIP-2's frozen Vision Transformer (ViT)-based (Dosovitskiy et al., 2021) vision encoder to produce a sequence of vision tokens for each video clip. The sequence of vision tokens is then compressed by BLIP-2's Q-Former into a fixed-length sequence. The fixed-length sequence is further projected to the word embedding space of the frozen language model of BLIP-2 by a linear layer. It is then interleaved with the text tokens according to the order in which video clips and texts appear in the interleaved data to form the input to the frozen language model. Following the fine-tuning procedure of Li et al. (2023c), we freeze the vision encoder and language model of the BLIP-2 models during training. For all of our experiments, we use BLIP-2 with 2.7 billion parameter OPT (Zhang et al., 2022) as its frozen language model (BLIP-2 OPT-2.7B), and BLIP-2 with XL-size Flan-T5 (Wei et al., 2022) as its frozen language model (BLIP-2 Flan-T5-xl).

[1] While there exist VLMs that already natively support interleaved video and text (Alayrac et al., 2022; Awadalla et al., 2023; Li et al., 2023b), we intentionally chose a VLM that did not to isolate the impact of our **EILeV** training paradigm on VLMs' in-context learning capability.

4.4 Evaluation

To evaluate our various model ablations, we need a means to measure the quality of action narrations generated by models, and the degree to which in-context learning supports this generation.

4.4.1 Action Narration Generation

One major difficulty in evaluating generative models for the action narration generation task is that there is no single correct way to describe the action in a video clip. In an ideal world, we would rely on human annotators to rate how close a generated action narration is to the ground truth, but the cost to do so would be prohibitive. In order to address this challenge, a number of semantic-similarity-based metrics (Zhang et al., 2019; Reimers and Gurevych, 2019) that correlate closely with human judgment have been proposed, and we take advantage of them in our evaluations. Specifically, we report the performance along semantic similarity-based scores produced by Siamese Sentence-BERT Bi-Encoder (STS-BE; Reimers and Gurevych, 2019). For completeness, we also report ROUGE-L (Lin, 2004), a lexical-based text generation metric.

4.4.2 In-Context Learning Capability

To evaluate the in-context learning capability of trained models for action narration, we vary the number of in-context examples in context-query instances (different numbers of "shots") and calculate the above text generation metrics for generated action narrations on the test set. If adding more shots improves narration quality under these metrics, this suggests that the VLM is successfully using in-context learning to adapt to the action narration generation task. Within a single experiment setting, we use the same pre-sampled in-context examples to ensure fair comparison.

5 Experimental Results

In our experiments, we find that the performance of both **EILeV**-trained models strictly increases as more in-context examples (shots) are provided, indicating that **our models successfully acquired in-context learning capabilities during training.** First, in Section 5.1, we establish the in-context learning capability of our models by measuring their performance on rare actions they were not trained on (the key challenge motivating this work), and compare their performance to that of off-the-shelf VLMs. In Sections 5.2, 5.3, and 5.4, we compare their performance to that of models trained on datasets with each key distributional property ablated (as described in Section 4.2) to explore the impact of these training data properties on in-context learning for video and text in VLMs.

5.1 Generalization to Rare Actions

We first compare our **EILeV**-trained models with existing off-the-shelf VLMs in the challenging practical setting that motivated this work: *adaptation to rare actions*. Specifically, we evaluate our models, Kosmos-2 (Peng et al., 2023), and Otter (Li et al., 2023b) on the evaluation set of held-out rare action videos from Ego4D described in Section 4.2.[2] We choose these two models as they are the only open-source large VLMs that support video input and in-context-learning out-of-the-box at the time of writing. Compared to our **EILeV**-trained models, these models have been trained on far more multi-modal interleaved (MMI) data directly related to in-context learning over video (Table 1), as well as other naturalistic multi-modal and text data from the Internet. They also have far more trainable parameters: Kosmos-2 has 1.6 billion and Otter has 1.3 billion, while our models have 188 million (the same number as BLIP-2). Further, unlike our architectural modification that represents each video with a fixed-length sequence, Kosmos-2 and Otter both treat each video as a sequence of images. For an evaluation representative of the practical usage of VLMs, we do not fine-tune models (which requires prohibitive computing power).

[2]Our models were not trained on these rare actions, and Kosmos-2 was not trained on Ego4D. While Otter was trained on Ego4D, the video-text training data was not interleaved as proposed for **EILeV**-trained models, and the low frequency of these actions nevertheless poses a significant challenge.

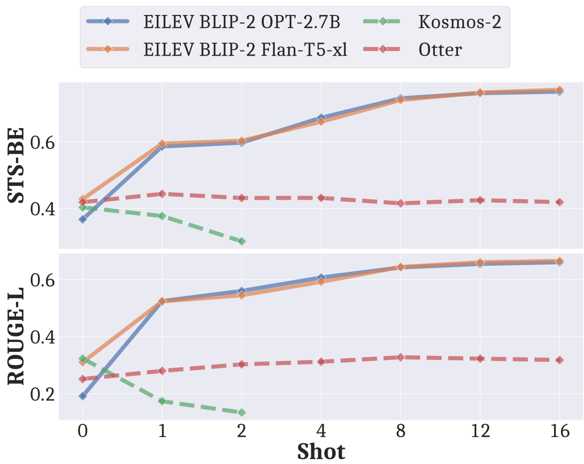

Figure 2: Performance of off-the-shelf VLMs (Kosmos-2 and Otter) on the evaluation set of rare actions for the skewed marginal distributions ablation experiment.

Model	MMI Dataset Size
EILeV BLIP-2 OPT-2.7B & Flan-T5-xl	115K context-query instances
Kosmos-2	71M image-text webpages (Huang et al., 2023)
Otter	101.2M image-text webpages (Zhu et al., 2023b) & 2.8M context-query instances (Li et al., 2023a)

Table 1: Off-the-shelf and **EILeV**-trained VLMs and their multi-modal interleaved (MMI) dataset sizes.

Instead, we rely solely on models' in-context learning capability to adapt to these rare actions.

Figure 2 shows the results of this evaluation.[3] While the zero-shot performance of our **EILeV**-trained models is similar to Kosmos-2 and Otter, **as we provide in-context examples, the performance of our models increases while that of off-the-shelf VLMs does not**. Consequently, our **EILeV-trained VLMs significantly outperform off-the-shelf VLMs**. While Kosmos-2 and Otter have not been fine-tuned on this exact data, they are much larger models trained on an enormous amount of naturalistic data, and their in-context learning capability is a main selling point thought to remove the need for task-specific fine-tuning. Therefore, it is reasonable to expect their performance to improve with more in-context examples

[3]We can only perform evaluations up to 2-shot with Kosmos-2, as it runs out of its context window beyond 2-shot.

or even outperform our models. This observation underscores that *training smaller VLMs with a focused approach like **EILeV** can be advantageous for certain use-cases*, such as generating narrations for novel, rare actions, than training large, generalist VLMs on huge naturalistic datasets.

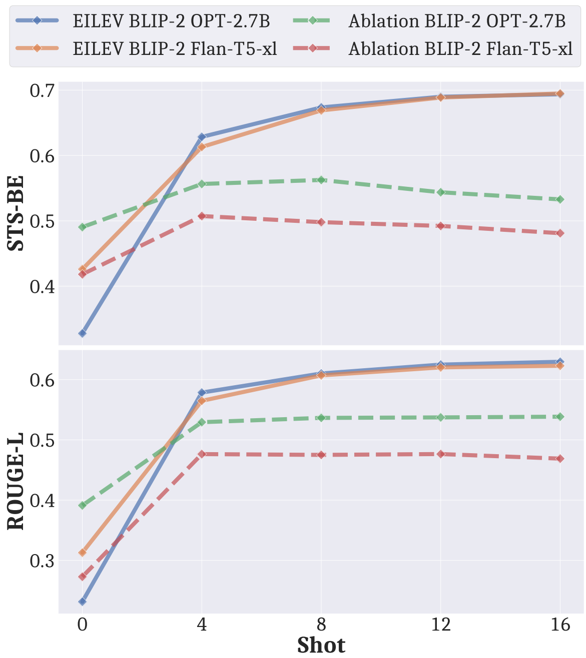

Figure 3: Results for the bursty distributions ablation experiment.

5.2 Bursty Distributions Ablation

Figure 3 shows the results of the bursty distributions ablation experiment. To maintain the same action distributions in both the training and test sets, we use a random train-test split with a ratio of 75/25 for this experiment. Unlike the **EILeV**-trained models, the performance of the models trained on randomly sampled in-context examples (ablation) initially improves from 0-shot to 4-shot, but tapers or even decreases as more examples are provided. This indicates that they failed to acquire in-context learning capabilities during training, suggesting that **bursty distributions are indeed necessary for in-context learning on video and text**. We hypothesize that the initial improvement in performance from 0-shot to 4-shot is mainly due to the fact that ablation models have learned to mimic lexical characteristics from in-context examples. However, as they have failed to learn to exploit the semantic information from in-context examples due to the lack of bursty distributions in training data, they do not benefit from additional in-context examples.

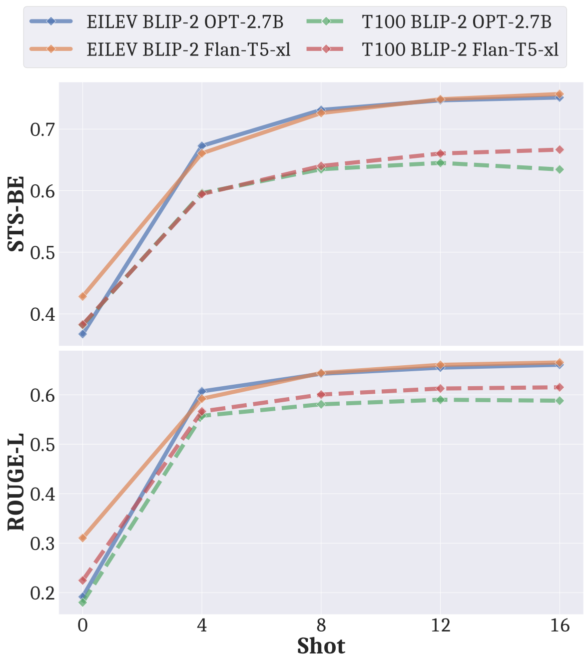

Figure 4: Results for the skewed marginal distributions ablation experiment using a training dataset with top 100 common actions (T100).

5.3 Skewed Marginal Distributions Ablation

Figures 4 and 5 show the results of the skewed marginal distribution ablation experiment. The T100 models trained on data with only the top 100 common actions (little skewness without a long tail of infrequent actions) show a noticeably inferior in-context learning performance to the **EILeV**-trained models that were trained on the training dataset with all the common actions (highly skewed with a long tail of infrequent items). On the other hand, the T500 models trained on data with the top 500 common actions (moderate skewness with a short tail of infrequent actions) show an in-context learning performance that is only slightly worse than the **EILeV**-trained models, indicating that **an increased amount of skewness with a long tail of infrequent items makes in-context learning more likely to appear in VLMs**. Further, we observe that the T500 models outperform their respective **EILeV**-trained models in the 0-shot setting. This is an instance of in-context versus in-weights learning tradeoff (also studied in Chan et al., 2022), a phenomenon where in-context learning capability can reduce pre-trained models' ability to utilize knowledge encoded in their weights during pre-training. Interestingly, we do not observe this pattern with the T100 models, perhaps because the less diverse training data is not representative enough for models to gain sufficient in-weights knowledge.

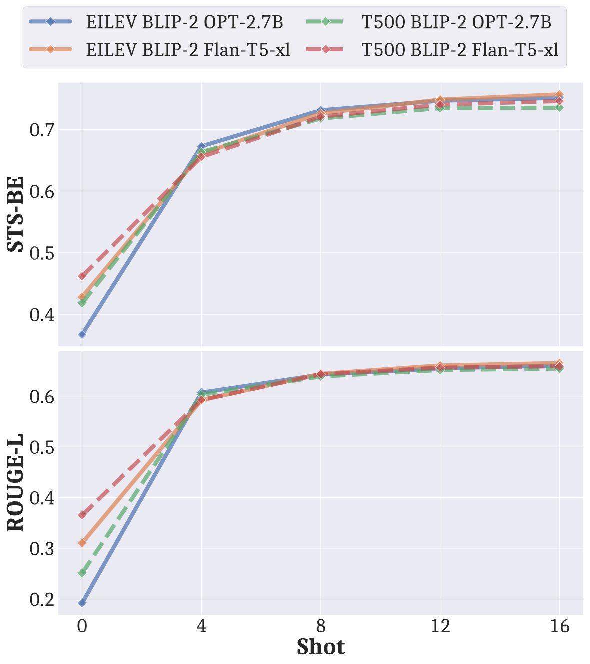

Figure 5: Results for the skewed marginal distributions ablation experiment using a training dataset with top 500 common actions (T500).

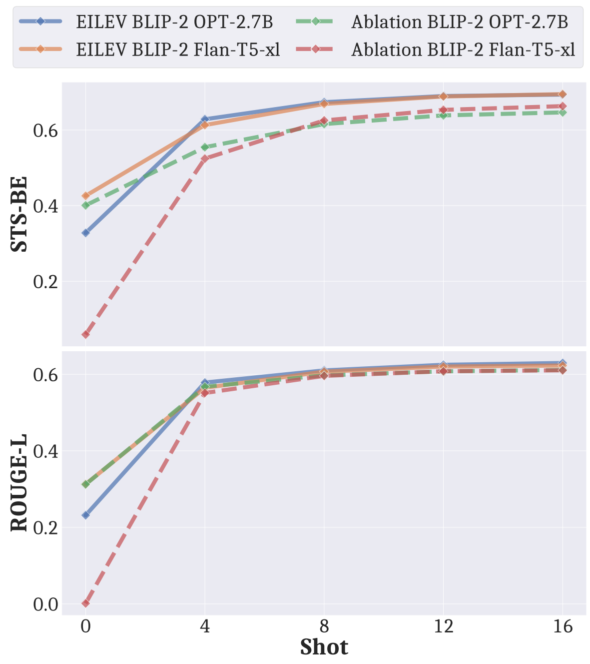

Figure 6: Results for the dynamic meaning ablation experiment.

5.4 Dynamic Meaning Ablation

Figure 6 shows the results of the dynamic meaning ablation experiment. We use a random train-test split with a ratio of 75/25 for this experiment to maintain the same action distributions in both the training and test sets. The ablation models trained on data with verbs and their corresponding objects canonicalized surprisingly acquire some in-context learning capabilities, but the **EILeV**-trained models mostly outperform them. Since the performance gaps under this ablation are smaller than that of the previous ablations, this suggests that **while dynamic meaning plays a role in the in-context capabilities of a VLM, it contributes less than bursty and skewed marginal distributions do**. Interestingly, however, the performance gap is much more pronounced for STS-BE (semantic similarity metric) than ROUGE-L (lexical metric), suggesting that dynamic meaning contributes more to the model's ability to extract semantic information from in-context examples than lexical information.

6 Conclusion

In this work, we conducted a first-of-its-kind systematic investigation of in-context learning in vision-language models (VLMs) trained on videos and text. Specifically, we implemented **E**mergent **I**n-context **L**earning on **V**ideos (**EILeV**), a novel training paradigm capturing three key properties of training data found to induce in-context learning in transformers (Chan et al., 2022): bursty distributions, skewed marginal distributions, and dynamic meaning. In our experiments, we showed that our **EILeV**-trained models exhibit in-context learning capabilities superior to that of off-the-shelf VLMs, as they were significantly more adaptable to novel, rare actions. We demonstrated that all three of these properties are indeed important to optimize the in-context learning capabilities of these models on narrating actions in videos, especially bursty and skewed marginal distributions.

Our work yields new insights about the nature of in-context learning in video and text. For example, we observed that while reducing the skewness of the training data distribution compromised in-context learning capability, it improved in-weights learning in trained models (Chan et al., 2022). We also found that dynamic meaning had a bigger impact on semantic similarity metrics for generated narrations than lexical metrics, suggesting this property is particularly important for acquiring semantic information through in-context learning.

While we focused on action narration in Ego4D (Grauman et al., 2022) as a proof-of-concept, **EILeV** serves as a foundation for the community to build VLMs capable of in-context learning on video and text in broader tasks and domains. We release our **EILeV**-trained models as a resource for future work in egocentric video narration.

7 Limitations

Since our **EILeV**-trained models are optimized and evaluated for action narration generation on egocentric video using in-context learning, their ability to generalize to diverse, real-world scenarios may be limited. However, this focus was by design and necessity. The primary goal of this work was to verify that the three distributional properties identified by Chan et al. (2022) also elicit in-context learning capabilities in VLMs for videos. To that end, we intentionally chose to use Ego4D, a dataset with sufficient annotations to enable our systematic ablation experiments as a proof of concept. Despite this limitation, **EILeV**-trained models may retain some capability to answer other types of questions due to the use of a frozen language model. Furthermore, **EILeV** is a general training method that can be applied to other tasks given the appropriate data.

Additionally, our models may inherit biases from their frozen language models, making it possible that they could generate harmful content. Before deploying such a system for real-world applications, safety measures like guardrails and training data sanitization are crucial to minimize potential negative impact. On the other hand, since we used the diverse and global data from Ego4D to train our models, this may mitigate possible socio-economic bias found in pre-trained visual representations (Nwatu et al., 2023).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Yuwei Bao, Keunwoo Peter Yu, Yichi Zhang, Shane Storks, Itamar Bar-Yossef, Alexander De La Iglesia, Megan Su, Xiao Lin Zheng, and Joyce Chai. 2023. Can foundation models watch, talk and guide you step by step to make a cake? *arXiv preprint arXiv:2311.00738*.

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.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. 2022. Data distributional properties drive emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, 35:18878–18891.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2022. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Yuqing Du, Ksenia Konyushkova, Misha Denil, Akhil Raju, Jessica Landon, Felix Hill, Nando de Freitas, and Serkan Cabi. 2023. Vision-language models as success detectors. In *Proceedings of The 2nd Conference on Lifelong Learning Agents*, volume 232 of *Proceedings of Machine Learning Research*, pages 120–136. PMLR.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian

Fuegen, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 2022. Ego4d: Around the World in 3,000 Hours of Egocentric Video. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*.

Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. 2022. Language models are general-purpose interfaces. *arXiv preprint arXiv:2206.06336*.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. 2023. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. 2024. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36.

Hugo Laurençon, Lucile Saulnier, Leo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023a. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023b. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023c. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023d. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2023. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *arXiv preprint arXiv:2312.17172*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu He, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. 2024. Mm1: Methods, analysis & insights from multimodal llm pre-training. *ArXiv*, abs/2403.09611.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. Metaicl: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809.

Joan Nwatu, Oana Ignat, and Rada Mihalcea. 2023. Bridging the digital divide: Performance variation across socio-economic factors in vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10686–10702, Singapore. Association for Computational Linguistics.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Toby Perrett, Saptarshi Sinha, Tilo Burghardt, Majid Mirhemdi, and Dima Damen. 2023. Use your head: Improving long-tail video recognition. In *Computer Vision and Pattern Recognition*.

Gautam Reddy. 2023. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *The Twelfth International Conference on Learning Representations*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Alex Jinpeng Wang, Linjie Li, Kevin Qinghong Lin, Jianfeng Wang, Kevin Qinghong Lin, Zhengyuan Yang, Lijuan Wang, and Mike Zheng Shou. 2024. Cosmo: Contrastive streamlined multimodal model with interleaved pre-training. *ArXiv*, abs/2401.00849.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. Zero-shot video question answering via frozen bidirectional language models. In *NeurIPS*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. Merlot reserve: Multimodal neural script knowledge through vision and language and sound. In *CVPR*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2024. LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. 2023. Learning video representations from large language models. In *CVPR*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023b. Multimodal C4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*.

A Additional Experiments

A.1 Additional Baselines

We report the performance of three additional baselines on the Ego4D-based dataset used in the main ablation experiments, as well as another dataset constructed from EPIC-KITCHENS-100 (Damen et al., 2022). The first is a naive action classification baseline (“VideoMAE”). Specifically, we fine-tune the “videomae-huge-finetuned-kinetics” variant of VideoMAE (Tong et al., 2022) using the verb and noun class annotations to produce a verb and a noun classifier. The predicted verb and noun classes are then transformed into action narrations using an off-the-shelf LLM (7 billion parameter Llama-2-Chat (Touvron et al., 2023)). Note that this baseline only uses videos as its input, and cannot perform in-context learning. The second are off-the-shelf BLIP-2 models with the architectural modifications from Section 4.3 for interleaved data support ("BLIP-2 OPT-2.7B & Flan-T5-xl"). The third are **EILeV**-trained models with in-context examples ablated, and fine-tune solely on the query ("FT BLIP-2 OPT-2.7B & Flan-T5-xl").

A.1.1 Results on Ego4D

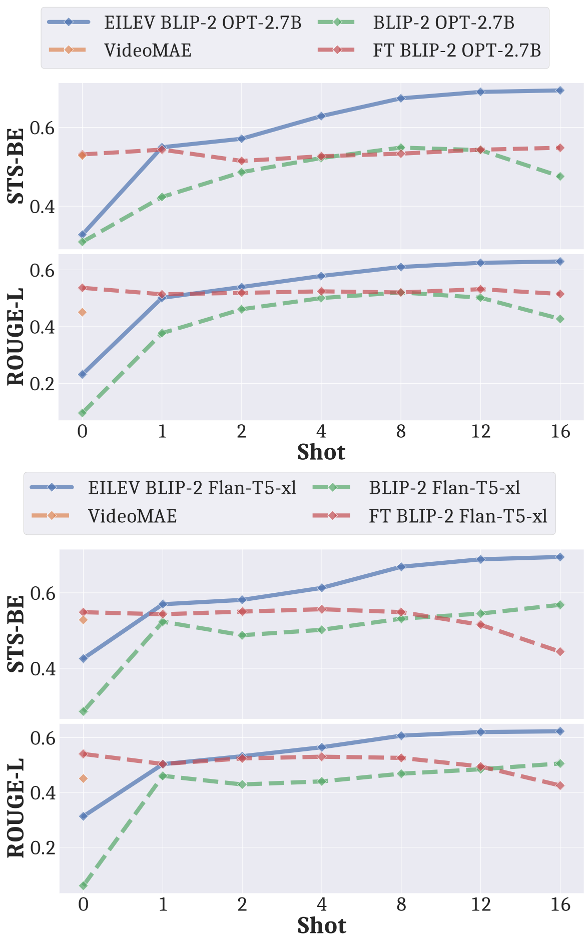

Figure 7: Performance of additional baselines on the Ego4D-based dataset.

Figure 7 reports the performance of the three additional baselines on the Ego4D-based dataset. We use a random train-test split with a ratio of 75/25 for this experiment to maintain the same action distributions in both the training and test sets. The **EILeV**-trained BLIP-2 models demonstrate superior in-context learning capabilities, as their performance improves with an increasing number of shots, ultimately outperforming all baseline models. This is a further indication that **EILeV** has successfully elicited in-context-learning capabilities in them. The VideoMAE and FT BLIP-2 models exhibit the best performance at 0-shot, suggesting they have the most amount of in-weights knowledge due to their fine-tuning. However, VideoMAE cannot process in-context examples, and its 0-shot performance is quickly outperformed by **EILeV**-trained models with only one in-context example. The performance of FT BLIP-2 models stagnates or even declines as the number of shots increases, highlighting their lack of in-context learning capabilities and the importance of the training data design discussed in Section 4.2. These findings about the performance of different models at 0-shot and subsequent shots align with Chan et al. (2022) observations regarding the "tradeoff between in-context learning and in-weights learning," where no models could maintain both in their experiments. In our experiment, the **EILeV**-trained BLIP-2 models are optimized for in-context learning, as evidenced by their subpar performance at 0-shot and superior performance with additional shots, whereas the FT BLIP-2 models show the opposite trend. We leave designing training data to find the right balance for future work.

A.1.2 Results on EPIC-KITCHENS-100

Next, we test if **EILeV**-trained BLIP-2 models trained solely on Ego4D can generalize to out-of-distribution actions via in-context learning. Specifically, we evaluate them on the validation split of a different egocentric video dataset, EPIC-KITCHENS-100, without further fine-tuning. Note that there is a significant distributional shift between Ego4D and EPIC-KITCHENS-100 even though they both contain egocentric videos in the kitchen setting as evidenced by the t-SNE plot in Figure 9. All the experimental setups are same as the experiments on the Ego4D-based dataset except evaluation context-query instances are formed by sampling both the context and the query from the validation set of EPIC-KITCHENS-100. Unlike Ego4D, the action narrations from EPIC-KITCHENS-100 are not full sentences, but simple verb-noun phrases. Therefore, we use an LLM (7 billion parameter Llama-2-Chat (Touvron et al., 2023)) to turn the simple verb-noun phrases into full sentences with "the camera wearer" as the subject.

Figure 8 reports the evaluation results. The performance of the **EILeV**-trained BLIP-2 models improves with an increasing number of in-context examples and ultimately outperforms all the baselines. This indicates that these models can generalize to

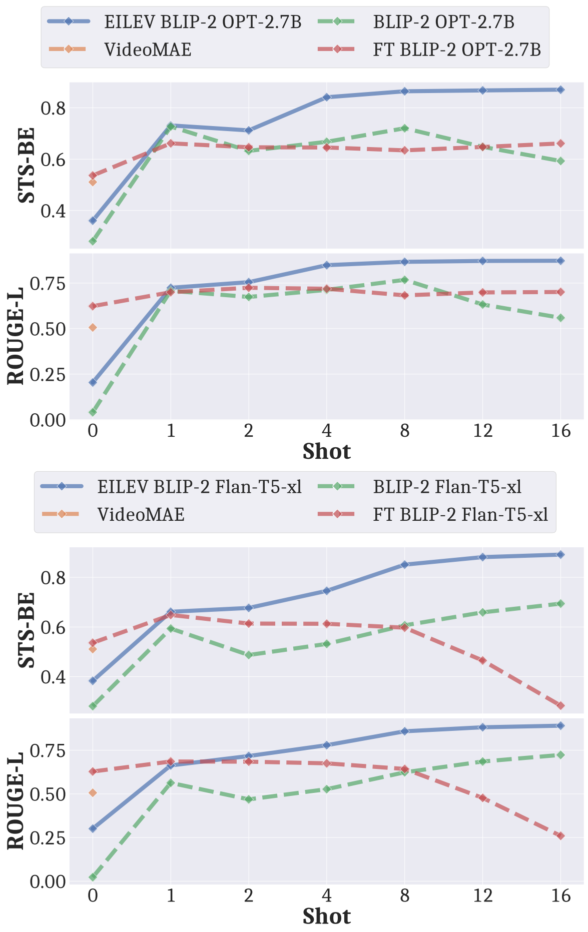

Figure 8: Performance of additional baselines on the EPIC-KITCHENS-100-based dataset

out-of-distribution actions via in-context learning. All the baseline models exhibit similar trends as on the Ego4D-based dataset: they demonstrate the best performance at 0-shot but fail to benefit from the in-context examples.

A.2 In-Context or In-Weights Learning

We now aim to validate that the source of the generalization capabilities demonstrated by the **EILeV**-trained models in Section 5.1 is indeed from in-context learning, not in-weights learning. This is to further reinforce our claim that **EILeV**-trained models can generalize to actions that they have not seen during training, i.e., actions of which they have no direct in-weights knowledge. To that end, we use the frequency of each verb/noun class in the common action training data as the proxy for the knowledge about the verb/noun class encoded into the weights of the model (in-weights learning), and the difference in model performance between 16-shot and 0-shot settings for a particular rare action as the proxy for in-context learning performance. If the model relies on in-weights learning for a particular novel, rare action, the difference in performance for that action between 16-shot and 0-shot settings would be correlated to the frequency of the corresponding verb/noun class in the training data. This outcome is not desired, as we want the model to rely on in-context learning for generating accurate narrations of novel, rare actions unseen during training.

Figure 10 shows the scatter plots between the log verb/noun class frequency in the training data and the difference in STS-BE for the corresponding rare action between 16-shot and 0-shot settings for the **EILeV**-trained models. For example, given a rare action (“put”, “bench”), a point on the scatter plot may refer to the log frequency of “put” in the common action training data in the x-axis and the difference in the STS-BE performance of **EILeV** BLIP-2 OPT-2.7B on (“put”, “bench”) between 16-shot and 0-shot. As the scatter plots and their corresponding R^2 values show, there is a minimal linear correlation between the log verb/noun class frequency in the training data and the difference in STS-BE for the corresponding action from in-context learning. This suggests that the **EILeV**-trained models generate accurate narrations for novel, rare actions via in-context learning rather than in-weights learning, as the linear model does not significantly account for the variance in the observed data.

A.3 Context Modeling and In-Context Learning

In this evaluation, we seek to investigate if the **EILeV**-trained models perform correct context modeling by incorporating the relationships between video clips and narrations. To that end, we evaluate the **EILeV**-trained models and the off-the-shelf BLIP-2 baseline models from Section A.1 on shuffled in-context examples where video clips no longer match the action narrations. We then compare their performance from shuffled in-context examples (the treatment group) to the one from unshuffled in-context examples as the control group. If the performance remains unchanged, it implies that the model does not consider the relationships between in-context video clips and action narrations. On the other hand, if the performance decreases, it implies that the model does take the relationships between video clips and action narrations into account, and the mismatch adversely affects its performance. We do not report the results at 0 and 1-shot since shuffling of the in-context video

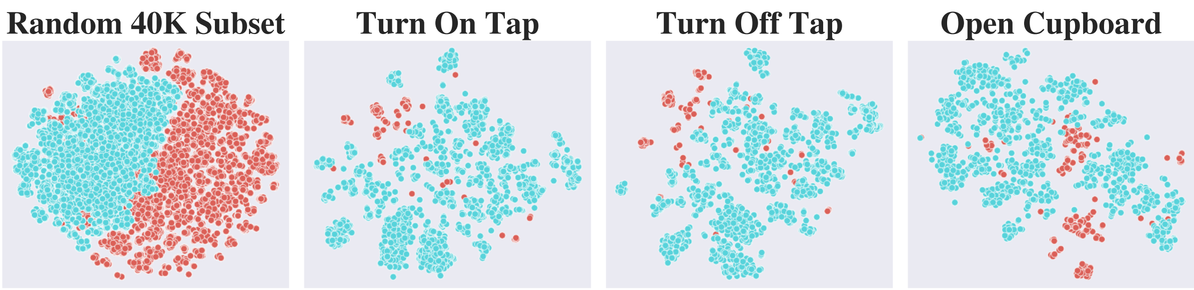

Figure 9: t-SNE plots of the video embeddings from the frozen vision encoder of BLIP-2 OPT-2.7B. Ego4D videos are in red, and EPIC-KITCHENS-100 videos are in blue. Plots for a randomly sampled subset of 40k videos from both and three most common actions from EPIC-KITCHENS-100 are shown. We manually map Ego4D actions to the EPIC-KITCHENS-100 actions.

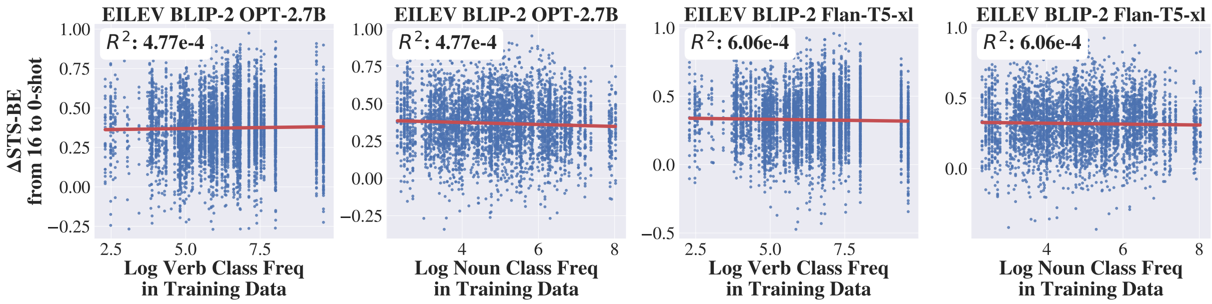

Figure 10: Scatter plots with trend lines and R^2 values between the log verb/noun class frequency in the training data with common actions and the difference in STS-BE (Δ STS-BE) for the corresponding rare action between 16-shot and 0-shot settings for the **EILeV**-trained models.

clips would not have any impact at those settings.

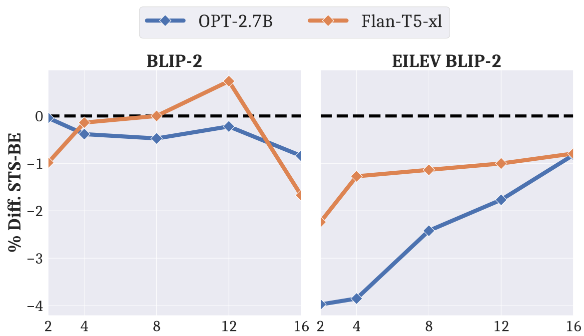

Figure 11: Percentage difference plots between the treatment group with shuffled in-context video clips and the control group. A negative value below the dotted zero line means the STS-BE performance of the treatment group is worse than the control group.

Figure 11 shows the percentage differences in STS-BE from 16-shot to 0-shot between the treatment group and the control group for the **EILeV**-trained models and the off-the-self BLIP-2 models. For the off-the-shelf BLIP-2 models, the percentage differences are small across all shots. This indicates that they rely mostly on the context as a whole rather than the semantic details from the relationships between video clips and action narrations when performing in-context learning. We hypothesize that our proposed architectural modifications (Section 4.3 allow the off-the-shelf BLIP-2 models to tap into the text-only in-context learning capabilities of their frozen language models, which lack the ability to extract semantic details from the relationships between video clips and action narrations. This hypothesis is supported by their subpar in-context learning capabilities from Section A.1, which speaks to the importance of our modifications to the training data. On the other hand, there is a clear drop in performance for the **EILeV**-trained models in terms of the semantic-similarity-based metric STS-BE. This indicates that the **EILeV**-trained models extract detailed semantic information from the correspondence between in-context video clips and action narrations.

B Training Details

In all of our experiments, each video clip is created by taking the four seconds before and after its action narration timestamp, and 8 frames are sampled uniformly from each video clip. The total training batch size is 128 and the optimizer is AdamW (Loshchilov and Hutter, 2018) with the initial learning rate of 1×10^{-5}, weight decay of 0.05 and a linear scheduler. We train for 5 epochs on 8 NVIDIA A40 GPUs using distributed data parallel. We evaluate every 200 steps and select the model with the lowest loss. The training time is about a day and a half.

C Question Templates

Table 2 shows the question-answer pair templates we use in our experiments. They are based on the instruction templates proposed by Dai et al. (2023).

Table 2: List of question-answer pair templates.

What is the camera wearer doing? {narration}
Question: What is the camera wearer doing? {narration}
What is the camera wearer doing? An answer to the question is {narration}
Q: What is the camera wearer doing? A: {narration}
Given the video, answer the following question. What is the camera wearer doing? {narration}
Based on the video, respond to this question: What is the camera wearer doing? Answer: {narration}
Use the provided video to answer the question: What is the camera wearer doing? {narration}
What is the answer to the following question? "What is the camera wearer doing?" {narration}
The question "What is the camera wearer doing?" can be answered using the video. The answer is {narration}