Tell, Don't Show!: Language Guidance Eases Transfer Across Domains in Images and Videos

Tarun Kalluri [1] **Bodhisattwa Prasad Majumder** [2] **Manmohan Chandraker** [1]
https://tarun005.github.io/lagtran/

Abstract

We introduce LaGTran, a novel framework that utilizes text supervision to guide robust transfer of discriminative knowledge from labeled source to unlabeled target data with domain gaps. While unsupervised adaptation methods have been established to address this problem, they show limitations in handling challenging domain shifts due to their exclusive operation within the pixel-space. Motivated by our observation that semantically richer text modality has more favorable transfer properties, we devise a transfer mechanism to use a source-trained text-classifier to generate predictions on the target text descriptions, and utilize these predictions as supervision for the corresponding images. Our approach driven by language guidance is surprisingly easy and simple, yet significantly outperforms all prior approaches on challenging datasets like GeoNet and DomainNet, validating its extreme effectiveness. To further extend the scope of our study beyond images, we introduce a new benchmark called Ego2Exo to study ego-exo transfer in videos and find that our language-aided approach LaGTran yields significant gains in this highly challenging and non-trivial transfer setting. Code, models and proposed datasets are publicly available at https://tarun005.github.io/lagtran/.

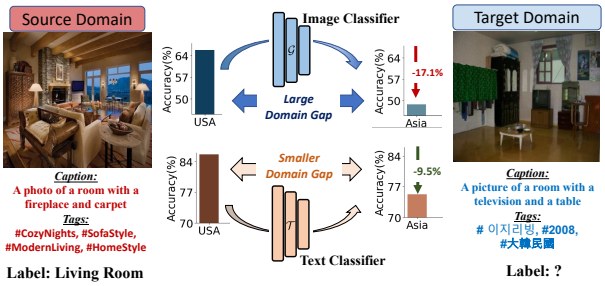


Figure 1: **A summary of our insights for LaGTran**: In a domain transfer setting with labeled source and unlabeled target domain data, we observe significantly more drop incurred while transferring an image-classifier trained on source images to target (17.1%), compared to a text-classifier trained on corresponding text descriptions of source images (9.5%). We use this insight to build a simple framework called LaGTran that leverages these text descriptions easily available in both domains to improve transfer in images and videos.

1. Introduction

Despite great strides in the performance in several applications of computer vision recent years, achieving robustness to distribution shifts at test-time still remains a challenge. In particular, a fundamental need to improve generalization to domains without manual supervision arises due to the cost and scarcity of acquiring labeled images. A dominant paradigm to address this limitation has been unsupervised domain adaptation (UDA), which uses labels from a related source domain along with distribution alignment techniques to bridge the domain gap (Ganin & Lempitsky, 2015; Long et al., 2018; Saito et al., 2018; Xu et al., 2019; Sharma et al., 2021; Wei et al., 2021; Chen et al., 2022; Zhu et al., 2023). Despite their noted success, their limitations in addressing challenging transfer beyond regular domain shifts (Saenko et al., 2010; Venkateswara et al., 2017; Peng et al., 2017) is recently highlighted (Prabhu et al., 2022; Kalluri et al., 2023). We posit that a part of this limitation potentially stems from their dependence on pixel-level data alone to bridge domain gaps, as accurately characterizing shifts and devising bridging strategies solely based on images becomes challenging beyond standard domain shift scenarios.

In contrast, we propose an alternative approach to ease transfer across such challenging shifts by instead leveraging ubiquitously available language guidance during training. Our framework, called **LaGTran** for **La**nguage **G**uided **Tr**ansfer **A**cross Domai**n**s, is surprisingly simple to implement, yet shows extreme effectiveness and competence in handling transfer across challenging domain shifts in images and videos compared to any image-based adaptation method.

*Equal contribution [1]UC San Diego [2]Allen Institute for AI. Correspondence to: TK <sskallur@ucsd.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Our key insight lies in observing that text guidance, which is readily available in the form of metadata for internet-sourced datasets or easily generated with emerging image captioning models, requires no human annotation while offering a more suitable avenue in transferring discriminative knowledge even across challenging domain shifts.

We further illustrate this property in Fig. 1, where we examine the transferability of image and text classifiers trained using image or text supervision respectively between USA and Asia domains from the GeoNet dataset (Kalluri et al., 2023). We observe significantly less drop (9.5%) when applying a text classifier trained on the source text to target text, compared to 17.1% drop incurred when transferring an image classifier to target images. As text operates in a significantly lower-dimensional space, language modality naturally has lesser domain gaps as opposed to images or videos. Furthermore, text descriptions often contain valuable attributes and identifiers that enhance the ability to accurately recognize images in a standard classification setting, suggesting more favorable domain robustness and discriminative properties of language descriptions compared to images.

In the current work, we incorporate these observations to improve transfer in a scenario where the source domain has text descriptions accessible along with the labels, but the target domain only has text descriptions corresponding to the images. Accordingly, we first train a text classifier using the source domain language descriptions and labels and transfer this classifier to assign pseudo-labels to the target text descriptions, which, from Fig. 1, would yield more robust pseudo-labels compared to the common image-based transfer (Liu et al., 2021a; Sun et al., 2022; Kumar et al., 2020). We, therefore, directly use these pseudo-labels as supervision for the unlabeled target images to train an image classifier jointly with source labels. This simple technique, free of any complicated adaptation mechanisms, shows remarkably strong performance surpassing competitive baselines and prior UDA methods.

To further demonstrate the broad usefulness of `LaGTran` beyond images, we introduce and study a novel benchmark for transfer learning in videos called Ego2Exo, which focuses on the previously under-explored challenge of transferring action recognition between ego (first-person) and exo (third-person) perspectives in videos (Li et al., 2021b; Ohkawa et al., 2023; Quattrocchi et al., 2023; Xue & Grauman, 2023) from a transfer learning standpoint. We curate Ego2Exo benchmark using cooking videos from the Ego-Exo4D dataset utilizing key step annotations to assign action labels and atomic action descriptions for textual guidance. Our language-aided transfer shows remarkable utility in this challenging setting, significantly outperforming prior Video-UDA methods (Chen et al., 2019b; Wei et al., 2023).

In summary, our contributions are three-fold.

- A novel framework `LaGTran` highlighting the feasibility of incorporating various forms of readily available text supervision in enhancing transfer across domain shifts (Sec. 3.2).
- A new dataset Ego2Exo to study the problem of cross-view transfer in videos with fine-grained labels covering a diverse pool of actions and free-form text descriptions providing language guidance (Sec. 4.3).
- Demonstration of the competence of `LaGTran` across a variety of domain shifts, with non-trivial gains over UDA methods on challenging datasets like GeoNet (+10%), DomainNet (+3%) and the proposed Ego2Exo (+4%) datasets (Sec. 4).

2. Related Work

Domain robustness in computer vision. A suite of methods have been proposed to improve accuracy on an unlabeled target domain by leveraging labels from a different source domain using unsupervised adaptation (Ben-David et al., 2006; 2010; Ganin & Lempitsky, 2015; Long et al., 2018), where prior works propose various domain alignment strategies including MMD-based (Tan et al., 2020; Long et al., 2017; Sun & Saenko, 2016; Kang et al., 2019), adversarial (Bousmalis et al., 2016; Tzeng et al., 2017; Saito et al., 2017; Chen et al., 2019a; Tzeng et al., 2015; Wei et al., 2021), class-specific adaptation (Pei et al., 2018; Saito et al., 2018; Luo et al., 2019; Xie et al., 2018; Kumar et al., 2018; Gu et al., 2020), clustering (Deng et al., 2019; Park et al., 2020; Li et al., 2021a; Kalluri & Chandraker, 2022) instance-specific adaptation (Sharma et al., 2021; Kalluri et al., 2022; Wang et al., 2022), self-training (French et al., 2017; Liu et al., 2021a; Sun et al., 2022; Prabhu et al., 2022) and more recently transformers (Xu et al., 2021) or patch-based mechanisms (Zhu et al., 2023). Similar ideas have also been explored in video domain adaptation (Choi et al., 2020), with extensions to incorporate temporal alignment (Chen et al., 2019a; Wei et al., 2023; Sahoo et al., 2021; Dasgupta et al., 2022). However, all these uda methods predominantly operate in the pixel-space, and often fall short in bridging more complicated forms of domain shift in challenging transfer settings (Kalluri et al., 2023; Prabhu et al., 2022). While some contemporary efforts utilize pre-trained CLIP models (Lai et al., 2023; 2024; Zhang et al., 2023) or LLMs (Chen et al., 2024) for domain alignment, we show that per sample natural language guidance can be equally effective. Based on this, our work introduces a new paradigm to study robustness, where we build a simple framework using rich textual descriptions to overcome large domain gaps in image and video recognition tasks.

Language supervision in computer vision. The recent proliferation of internet-sourced datasets highlights the

ready availability of natural language supervision without the need for any labeling or annotation efforts in images (Thomee et al., 2016; Changpinyo et al., 2021; Schuhmann et al., 2022; Mahajan et al., 2018; Desai et al., 2021) and videos (Miech et al., 2019; Bain et al., 2021; Grauman et al., 2022; 2023). This availability of language supervision has been effectively utilized to learn scalable weakly supervised models (Mahajan et al., 2018; Singh et al., 2022), robust vision-language representations (Radford et al., 2021; Jia et al., 2021; Pham et al., 2023; Desai & Johnson, 2021; Sariyildiz et al., 2020; Lin et al., 2022; Zhao et al., 2023; Goyal et al., 2022), text-conditioned generative models (Rombach et al., 2022; Ramesh et al., 2021; Saharia et al., 2022) and improving sampling techniques for self-supervised learning (El Banani et al., 2023). Even in the absence of associated language supervision, recent innovations showed the potential of generating correlated descriptions for images using image-to-text or image captioning models (Li et al., 2023; Liu et al., 2023; Achiam et al., 2023). Despite this ubiquity and proven effectiveness of language supervision for vision tasks, little attention has been directed at leveraging their utility in improving transfer learning across domains. In this work, we use language guidance to develop a straightforward mechanism to improve image and video classification on domains without manual supervision.

Domain robustness using language supervision. Recent emergence of large-scale pre-trained vision-language foundational models such as CLIP (Radford et al., 2021) enabled strong zero-shot generalization across diverse domains and tasks (Devillers et al., 2021). However, the zeroshot inference using frozen pre-trained models still fall short of supervised fine-tuning (Radford et al., 2021; Pham et al., 2023; Andreassen et al., 2021), which in-turn suffers from poor generalization to distributions outside the fine-tuning data (Kumar et al., 2022; Wortsman et al., 2022). Prior works explored robust fine-tuning of zero-shot models, but do not leverage target domain data (Udandarao et al., 2023) or language supervision (Wortsman et al., 2022) during fine-tuning. While recent works incorporate language guidance for domain generalization (Dunlap et al., 2023; Wang et al., 2024; Liu & Wang, 2023; Huang et al., 2023; Min et al., 2022; Gokhale et al., 2021), they mostly rely on domain or class descriptors and do not leverage semantically richer free form text supervision from target images during transfer. One work which is closest to ours is (Goyal et al., 2023), which uses label names as text descriptions while we use free-form captions for images or atomic annotations for actions. In video recognition literature, prior works seek to align ego and exo views using language pre-training (Xu et al., 2024; Huang et al., 2024). In contrast to these efforts, we show that incorporating language aided transfer through diverse supervision yields a remarkably effective framework for improving domain robustness in both images and videos.

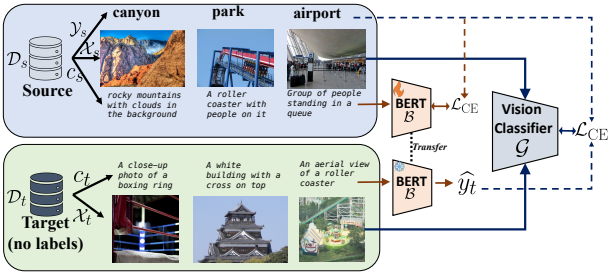


Figure 2: **An overview of training using LaGTran:** We operate in a setting where the labeled source domain and unlabeled target domain data possess cheaply available or easily generated language descriptions for each image. LaGTran proceeds by first training a BERT-classifier $\mathcal{B}$ using source captions and labels (Eq. (1)), and using the trained model to generate pseudo-labels $\hat{y}_t$ for the target captions and corresponding images (Eq. (2)). We then use this generated supervision along with source domain data in jointly training a **Vision classifier** $\mathcal{G}$ for image or video classification (Eq. (3)).

3. Method Details

3.1. Problem Description and Background

We consider the setting of unsupervised cross-domain transfer, with access to labeled data from a source domain $\mathcal{D}_s : \{X_s^i, y_s^i\}_{i=1}^{N_s}$ along with unlabeled data from a target domain $\mathcal{D}_t : \{X_t^i\}_{i=1}^{N_t}$, where $X_s \sim P_s$, $X_t \sim P_t$, N_s and N_t are the number of samples in source and target domains, and the covariate shift assumption means marginal distributions $P_s \neq P_t$ (Ben-David et al., 2006; 2010). However, different from prior works, we additionally assume access to natural language descriptions, denoted by c_i, corresponding to each image or video input in both source and target domains during training. Consequently, we denote the labeled source domain with $\mathcal{D}_s : \{X_s^i, y_s^i, c_s^i\}_{i=1}^{N_s}$ and the unlabeled target domain with $\mathcal{D}_t : \{X_t^i, c_t^i\}_{i=1}^{N_t}$. These text descriptions are readily available through associated metadata in web-collected images (Mahajan et al., 2018), or can be effortlessly generated with state-of-the-art image-to-text models (Li et al., 2023). In Sec. 4, we show robust performance using text descriptions derived from a variety of sources, including: image metadata (e.g., alt-text, hashtags) for web-sourced images, state-of-the-art image captioners for manually curated datasets, as well as action descriptions or narrations in videos. Note that our setting requires language descriptions c_i only during training and not during inference or deployment, and therefore incurs no speed or memory overhead at test-time when compared with prior approaches.

3.2. LaGTran for Cross-Domain Transfer

Overview. The training pipeline used in LaGTran for cross-domain transfer is summarized in Fig. 2, where we first train a BERT sentence classifier using the (text, label) pairs from the source domain dataset, and utilize this trained classifier to infer predictions on all the descriptions from the target domain. We then use these predictions as pseudo-labels for the target images, and train a joint vision classifier along with the labeled source domain images.

Training the text classifier. We use the supervised text-label pairs from the source domain (c_i^s, y_i^s) and train a BERT (Devlin et al., 2019) sentence classifier $\mathcal{B}$ to predict the category label from an input text description, using the training objective

$$\phi^* = \arg\min_{\phi} \mathbb{E}_{(c_i,y_i)\sim\mathcal{D}_s} \mathcal{L}_{\text{CE}}(\mathcal{B}(c_i;\phi), y_i), \quad (1)$$

where ϕ denotes the parameters of the BERT classifier and $\mathcal{L}_{\text{CE}}$ is the supervised cross-entropy loss. We adopt a pre-trained Distill-BERT (Sanh et al., 2019) model from HuggingFace as the sentence classifier $\mathcal{B}(;\phi)$, and fine-tune it on the source domain data. We observed sub-optimal performance using other pre-trained backbones such as T5 (Raffel et al., 2020) or GPT-2 (Radford et al., 2019) (Tab. 5). Across all datasets and experiment settings used in this paper, we feed the raw text descriptions directly into the sentence classifier network without any preprocessing or manual curation. We observed remarkable robustness of the trained classifier in handling several challenges posed by unfiltered text, including their variable lengths across images, language barriers prevalent in geographically diverse data, unrelated tags and descriptions commonly found in web-sourced images or potentially imperfect captions from state-of-the-art captioning models.

To further illustrate our motivation to use text classifier for label transfer, we show the tSNE visualizations of the feature embeddings derived from a source-trained sentence classifier, and compare them to the features derived from a source-trained image classifier in Fig. 3. Evidently, the features computed using the text classifier (Fig. 3c and 3d) are well-separated (more intra-class separation) and well-aligned (less inter-domain separation) compared to image classifier (Fig. 3a and 3b) further validating our hypothesis that the text descriptions of same-class images from both within and across domains lie close to each other.

Cross-modal supervision transfer. We distill the powerful discriminative knowledge learned from text into images through cross-modal (text to image) supervision transfer in the target domain. Specifically, we first freeze the weights of the source-trained BERT classifier $\mathcal{B}$ and compute pseudo-labels on all the target images using their corresponding text descriptions. For an image x_i^t with caption c_i^t,

$$\hat{y}_i^t = \arg\max_{\mathcal{C}} \mathcal{B}(c_i^t;\phi^*), \quad (2)$$

where $\mathcal{C}$ is the set of categories in the classification task. Using these predictions, we construct a pseudo-labeled target dataset, given by $\widehat{\mathcal{D}_t} = \{x_i^t, \hat{y}_i^t\}_{i=1}^{N_t}$. Finally, we combine this pseudo-labeled target images along with manually labeled source domain images to train an image classifier backbone $\mathcal{G}$ by sampling an equal number of images from both source and target in each mini-batch to eliminate effects caused by different dataset sizes.

$$\arg\min_{\theta} \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_s} \mathcal{L}_{\text{CE}}(\mathcal{G}(x_i;\theta), y_i) + \mathbb{E}_{(x_i,\hat{y}_i)\sim\widehat{\mathcal{D}}_t} \mathcal{L}_{\text{CE}}(\mathcal{G}(x_i;\theta), y_i) \quad (3)$$

Note that the inference is performed exclusively using the trained image-based classifier $\mathcal{G}(;\theta^*)$ on image inputs, and neither the text inputs nor the sentence-classifier $\mathcal{B}$ is needed or used at test-time.

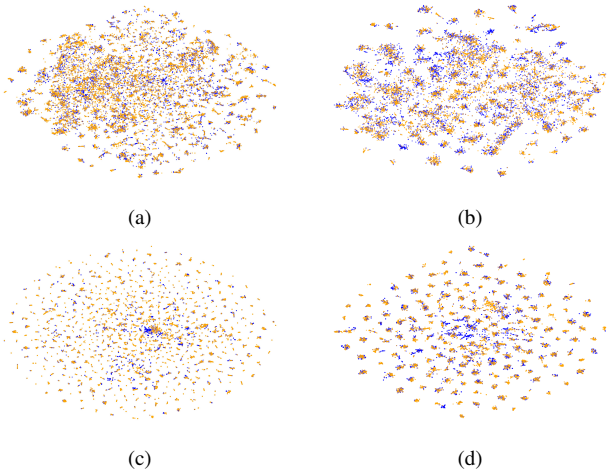


Figure 3: **tSNE visualization of cross-domain features on GeoNet.** We show improved domain-alignment with better class-separation in source and target when extracting features from a text-classifier (Fig. 3c-3d) compared to features from image-classifier (Fig. 3a-3b) highlighting better transferability through text modality. (Source in orange and target in blue).

3.3. Extending LaGTran to Handle Outliers

Owing to the simplicity in the design, LaGTran can easily be extended to the case where the target domain potentially contains outlier samples from outside the category set, also called open-world or universal adaptation (UniDA) (You et al., 2019; Saito et al., 2020). While classical transfer necessitates complete matching between source and target category spaces, open-world transfer relaxes this requirement, allowing the possibility of encountering images from previously unseen and outlier categories during test-time in

the target domain (You et al., 2019; Saito et al., 2020). The task is then to accurately classify a test-image into one of $\mathcal{C}_s$ categories shared between source and target domains while simultaenously detecting outlier images from target private classes. To suit LaGTran for UniDA, we modify Eq. (2) to additionally label predictions made by the text-classifier network $\mathcal{B}$ with an *outlier* class using maximum softmax probability thresholding (Hendrycks & Gimpel, 2016) after training.

$$\hat{y}_i^t = \begin{cases} \arg\max_{\mathcal{C}_s} \mathcal{B}(c_i^t; \phi^*) & \text{if } \max_{\mathcal{C}_s} \mathcal{B}(c_i^t; \phi^*) > \tau \\ |\mathcal{C}_s| + 1 & \text{otherwise,} \end{cases} \tag{4}$$

where τ is a threshold used to detect outlier samples during inference. We then proceed to train a downstream classifier on $|\mathcal{C}|_s+1$ classes using data from supervised source and psuedo-labeled target data from shared classes as well as the outlier class. During inference, we assign a test-image to one of the $\mathcal{C}_s$ classes or the special outlier class based on the prediction. We heuristically choose $\tau = 0.75$ and do not ablate on this. We show in Sec. 4.2 that this simple extension yields highest accuracy on the challenging GeoUniDA dataset (Kalluri et al., 2023), highlighting the versatility of LaGTran to handle diverse styles of domain transfer.

4. Experiments

4.1. LaGTran for Image Classification

Datasets. We adopt GeoNet (Kalluri et al., 2023) and DomainNet (Peng et al., 2019) datasets which together cover a range of domain shifts across varying difficulty levels. GeoNet is the largest dataset for domain adaptation with more than 750k images, proposed to study a practical real-world problem of geographic disparities in images for two tasks - GeoImnet for image classification from 600 classes and GeoPlaces for scene recognition from 205 classes. DomainNet is a challenging dataset proposed for adaptation with 400,000 images from 345 classes. Following prior work (Wei et al., 2021; Kalluri et al., 2022), we show our results on all 12 transfer settings from the 4 most studied domains *real, clipart, sketch* and *painting*. We use a ViT-base (Dosovitskiy et al., 2020) backbone as the image encoder on the GeoNet, and follow prior work (Zhu et al., 2023) and use Swin-base backbone (Liu et al., 2021b) for experiments on DomainNet. Complete training details are included in the supplementary (Supp. C).

Source of text supervision. For GeoNet, we use text supervision from the metadata publicly released along with the dataset, and concatenate the tags, alt-text and free-form captions provided for each image to create the text descriptions. For the DomainNet dataset, since no associated text descriptions are provided, we use a BLIP-2 (Li et al., 2023) model to generate short captions for each image from all the domains. Note that our method only requires text during training, and inference is done solely based on images.

Baselines. A possible argument for the effectiveness of text supervision might be the direct presence of label information in the text description, eliminating the need for any manual supervision at all. To study this in greater detail, we devise two strong baselines to derive psuedo-labels directly using the text descriptions in the target without using any source domain data as follows. We first use a pre-trained Sentence-BERT (Reimers & Gurevych, 2019) encoder, and compute the label embeddings of all the category names as $\mathbf{L} \in \mathbb{R}^{|\mathcal{C}| \times d}$, where d is the embedding dimension of the sentence encoder, followed by zero-shot inference using: (i) **TextMatch,** where we compute the embedding of each text description $e_i^t \in \mathbb{R}^{1 \times d}$ from the target domain, and assign psuedo-label to the label with the highest similarity score with the text embeddings: $\hat{y}_i = \arg\max_{|\mathcal{C}|}(e_i^t \cdot \mathbf{L}^T)$, and (ii) **nGramMatch,** where we additionally compute the set of all n-grams $\{w\}$ for each text description c_i for $n = \{1, 2, 3, 4\}$ and find the embeddings for each of these ngrams separately: $\mathbf{W} \in \mathbb{R}^{|w| \times d}$. The pseudo-label is then assigned to the label with the highest similarity score with the best matching ngram: $\hat{y}_i = \arg\max_{|\mathcal{C}|} \max_w(\mathbf{W} \cdot \mathbf{L}^T)$. Once the psuedo-labels are generated, we proceed with training a joint model using Eq. (3) as before.

In addition to these, we also compare the zero-shot classification obtained using **CLIP** (Radford et al., 2021) with ViT-base backbone. We adopt domain-aware prompting following prior work (Dunlap et al., 2023; Liu & Wang, 2023), where we incorporate the domain information into the prompt-text (eg: `A sketch of a <class>` instead of `A photo of a <class>` to classify sketch images).

LaGTran Outperforms UDA on GeoNet As noted in (Kalluri et al., 2023), previous UDA methods often fall short of bridging geographic disparities, highlighting the challenge of geographical transfer with image data alone. From Tab. 1, LaGTran achieves 60.24% average Top-1 average accuracy on GeoNet, beating all previous UDA methods and strong baselines by significant margins, providing solid validation to our transfer approach using language guidance. Specifically, LaGTran outperforms the source-only baseline by $\sim$14% and best adaptation approach PMTrans (Zhu et al., 2023) by $\sim$10% on the average accuracy, highlighting the natural benefit conferred by training while leveraging text supervision in source and target domains. LaGTran even surpasses zeroshot accuracy using domain-aware prompting on CLIP (Radford et al., 2021) by $\sim$10%, while being trained on order of magnitude fewer data compared to CLIP's hundreds of millions of image-text pairs. Remarkably, we also

	GeoImnet		GeoPlaces		Average
	U→A	A→U	U→A	A→U	
Unsupervised Adaptation					
Source Only	52.46	51.91	44.90	36.85	46.53
CDAN (Long et al., 2018)	54.48	53.87	42.88	36.21	46.86
MemSAC (Kalluri et al., 2022)	53.02	54.37	42.05	38.33	46.94
ToAlign (Wei et al., 2021)	55.67	55.92	42.32	38.40	48.08
MDD (Zhang et al., 2019)	51.57	50.73	42.54	39.23	46.02
DALN (Chen et al., 2022)	55.36	55.77	41.06	40.41	48.15
PMTrans (Zhu et al., 2023)	56.76	57.60	46.18	40.33	50.22
Zeroshot Classification					
CLIP† (Radford et al., 2021)	49.84	53.83	43.41	54.34	50.36
TextMatch	49.68	54.82	53.06	50.11	51.92
nGramMatch	49.53	51.02	51.70	49.87	50.93
LaGTran	**63.67**	**64.16**	**56.14**	**57.02**	**60.24**

Table 1: LaGTran outperforms all prior methods by $>10\%$ on average with the challenging GeoImnet benchmark with 600 classes and GeoPlaces with 205 classes designed for geographical transfer. All methods use a ViT-B backbone. †denotes domain aware-prompting. Best values in **bold**, second best underlined. U:USA, A:Asia.

outperform the strongest baseline *TextMatch* by $\sim 8\%$, underlining the fact that in cases when the text descriptions might not always have embedded label information directly, using labels from a source is still advantageous.

LaGTran achieves new SOTA on DomainNet We summarize the results on DomainNet in Tab. 2, where LaGTran yields large gains over several prior UDA methods and all the competitive baselines, setting new state-of-the-art. Notably, many prior methods return lesser numbers than directly training on a source model (Saito et al., 2018; Zhang et al., 2019; Du et al., 2021; Li et al., 2021a), indicating their poor scalability to natural domain shifts in large-scale data. While more recent innovations in UDA such as self-training (Sun et al., 2022) and patch-based mixing (Zhu et al., 2023) offer improved performance, LaGTran still outperforms these methods on most tasks. Finally, our superior accuracy compared to both baselines *TextMatch* and *nGramMatch*, that employ target-only pseudo-labeling, underscores the significance of having access to supervised text data and labels from a source domain for enhanced target accuracy. Further, a potential explanation for limited benefits being observed in LaGTran, as well as all previous UDA methods, when compared to CLIP under non-real to real transfer could be the precise match between images from standard categories in the real-target domain and the multi-million-scale training data utilized in CLIP. This alignment potentially eliminates any significant domain shift between the train and test settings, unlike in LaGTran where we train on non-real images yet achieve accuracies that are competitive to CLIP. Notably, LaGTran still outperforms CLIP on all domains in GeoNet (+10%) and most domains in DomainNet (upto +7%) while being trained on multiple-orders of magnitude lesser image-text pairs than CLIP.

4.2. LaGTran Improves Transfer with Outliers

We show our results on open-world transfer setting using the GeoUniDA dataset (Kalluri et al., 2023), which examines unsupervised transfer across geographies in the presence of geographically unique classes in both source and target along with common classes. Specifically, GeoUniDA contains 62 shared classes between source and target, along with 138 private categories in each domain. We follow OVANet (Saito & Saenko, 2021) to adopt the H-score evaluation metric, which gives equal importance to closed-set and open-set accuracies by measuring the harmonic mean of both. In addition to standard works that address outlier detection through universal adaptation (You et al., 2019; Saito & Saenko, 2021; Saito et al., 2020), we also train a baseline model using only the source domain data performing test-time outlier detection using MSP thresholding (Hendrycks & Gimpel, 2016). As shown in Tab. 3, LaGTran achieves a H-score of 61.16%, significantly surpassing the baseline source-only accuracy as well as all prior universal adaptation approaches by $>10\%$, indicating that language guidance naturally provides a strong signal to detect target samples while handling outliers in open-set target domain data.

4.3. LaGTran for Video Domain Adaptation

Ego2Exo dataset. Despite rapid advances in methods (Chen et al., 2019b; Munro & Damen, 2020; Choi et al., 2020; Wei et al., 2023) and benchmarks (Munro & Damen, 2020; Plizzari et al., 2023) for video domain adaptation, little insight is available into their ability to address challenging settings such as transfer between ego (first-person) and exo (third-person) perspectives in videos. While prior efforts studying ego-exo transfer require paired videos from both views (Quattrocchi et al., 2023; Sigurdsson et al., 2018; Huang et al., 2024) or do not leverage target unlabeled data (Li et al., 2021b; Ohkawa et al., 2023), limited works study unsupervised domain transfer from ego to exo views due to the lack of a suitable benchmark.

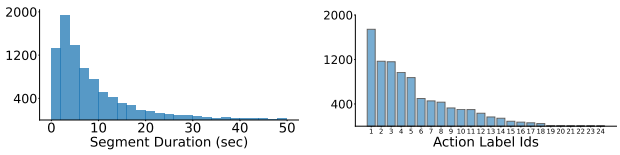
Therefore, we introduce a new benchmark called Ego2Exo to study transfer between the ego and exo views in videos. We curate our dataset using the recently proposed Ego-Exo4D (Grauman et al., 2023), utilizing their keystep annotations for action labels, and atomic descriptions as text supervision. We manually remap the labels to a coarser hierarchy to ease the difficult task of predicting very fine-grained action classes from short clips (eg: `add coffee beans` vs. `add coffee grounds`). Complete details about our dataset creation process are included in Supp. B. Our proposed Ego2Exo consists of video segments labeled with actions from one of the 24 keysteps from ego and exo views of the corresponding actions. We obtain 4100 ego-videos and 4986 exo-videos capturing variety of actions and scenes.

Source	Real→			Clipart→			Sketch→			Painting→			
Target	C	S	P	R	S	P	R	C	P	R	C	S	Avg.
Unsupervised Adaptation													
Source Only	63.02	49.47	60.48	70.52	56.09	52.53	70.42	65.91	54.47	73.34	60.09	48.25	60.38
MCD (2018)	39.40	25.20	41.20	44.60	31.20	25.50	34.50	37.30	27.20	48.10	31.10	22.80	34.01
MDD (Zhang et al., 2019)	52.80	41.20	47.80	52.50	42.10	40.70	54.20	54.30	43.10	51.20	43.70	41.70	47.11
CGDM (Du et al., 2021)	49.40	38.20	47.20	53.50	36.90	35.30	55.60	50.10	43.70	59.40	37.70	33.50	45.04
SCDA (Li et al., 2021a)	54.00	42.50	51.90	55.00	44.10	39.30	53.20	55.60	44.70	56.20	44.10	42.00	48.55
SSRT-B (Sun et al., 2022)	69.90	58.90	66.00	75.80	59.80	60.20	73.20	70.60	62.20	71.40	61.70	55.20	65.41
MemSAC (Kalluri et al., 2022)	63.49	42.14	60.32	72.33	54.92	46.14	73.46	68.04	52.75	74.42	57.79	43.57	59.11
CDTrans (Xu et al., 2021)	66.20	52.90	61.50	72.60	58.10	57.20	72.50	69.00	59.00	72.10	62.90	53.90	63.16
PMTrans (Zhu et al., 2023)	74.10	61.10	**70.00**	79.30	63.70	62.70	77.50	73.80	62.60	79.80	69.70	61.20	69.63
Zero-shot Classification													
CLIP† (Radford et al., 2021)	72.39	60.90	66.81	**81.37**	60.90	66.81	**81.37**	72.39	66.81	**81.37**	72.39	60.90	70.38
TextMatch	71.36	64.30	65.32	81.25	65.65	64.85	81.09	72.65	63.94	81.08	70.84	**64.17**	70.14
nGramMatch	68.92	59.82	63.15	76.35	61.72	62.87	76.35	69.28	62.51	76.04	68.52	60.52	67.17
LaGTran	**77.30**	**68.25**	67.35	81.31	**67.03**	**66.81**	80.78	**75.62**	**68.08**	79.23	**73.80**	63.44	**72.41**

Table 2: LaGTran sets new state-of-the-art on DomainNet-345 dataset, outperforming prior methods and baselines in most tasks. All models use Swin-B backbone, and UDA numbers are taken from (Zhu et al., 2023). †denotes domain aware-prompting. Best values in **bold**, second best underlined. R:Real, C:Clipart, S:Sketch, P:Painting.

Method	Closed Set Acc.	Open Set Acc.	H-score
Source Only w/MSP	38.00	73.90	50.20
UniDA (You et al., 2019)	27.64	43.93	33.93
DANCE (Saito et al., 2020)	38.54	78.73	51.75
OVANet (Saito & Saenko, 2021)	36.54	66.89	47.26
LaGTran	52.98	72.35	**61.16**

Table 3: **Results on open-world transfer on GeoUniDA** shows strong performance of LaGTran even with target outlier classes, achieving the highest H-score. Baseline numbers takes from (Kalluri et al., 2023).



(a) Segment Duration (b) Label Distribution

Figure 4: **Dataset Statistics for Ego2Exo:** Fig. 4a Shows the distribution of segment durations of action videos from Ego2Exo which range from 0.4sec-1min. Fig. 4b shows the long-tail of category distribution in Ego2Exo indicating the challenge in robust classification and transfer.

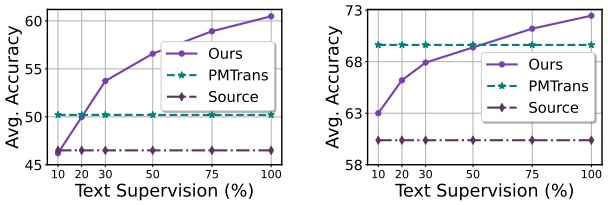
The atomic action descriptions from all the timestamps within each segment form the text supervision for that segment. The same procedure is applied to the validation videos yielding 3147 segments with both ego and exo views. The distribution of the duration of segments in the benchmark, along with the label distribution for ego and exo domains is presented in Fig. 4. We provide more details about the construction of the dataset in the supplementary material. The videos, labels along with the text descriptions are publicly available on our project page.

	Ego→Exo	Exo→Ego	Avg.
Unsupervised Adaptation			
Source Only	8.39	15.66	12.03
TA3N (Chen et al., 2019b)	6.92	27.95	17.44
TransVAE (Wei et al., 2023)	12.06	23.34	17.70
Zero-shot Video Recognition			
EgoVLP (Lin et al., 2022)	5.89	19.35	12.62
LaVILA (Zhao et al., 2023)	5.86	23.16	14.51
TextMatch	10.36	13.57	11.97
nGramMatch	11.50	15.46	13.98
LaGTran	**12.34**	**30.76**	**21.55**
Target Sup.	17.91	33.19	25.55

Table 4: **Results on Ego2Exo benchmark** LaGTran achieves the highest accuracy compared to prior video UDA methods as well as zeroshot video-text pre-trained models. Best values in **bold**, second best underlined. All methods use pre-extracted omnivore-base features, EgoVLP and LaVILA use Timesformer-base backbone.

Training details. We use the pre-computed Omnivore-base (Girdhar et al., 2022) features provided along with the EgoExo4D dataset for training and evaluation, and follow the same strategy for training all the other baselines as well as prior adaptation methods for fair comparison. We use the top-1 accuracy on the validation set for evaluation. More details on the training procedure are provided in the supplementary, Supp. C. We compare LaGTran for video with prior UDA approaches (Chen et al., 2019a; Wei et al., 2023) as well as Video-CLIP based methods with domain-aware prompting (Lin et al., 2022; Zhao et al., 2023).

LaGTran efficiently handles cross-view transfer. Firstly, we highlight the importance of studying robustness across ego and exo views in Tab. 4 by examining the ego-test



(a) Accuracy on GeoNet (b) Accuracy on DomainNet

Figure 5: **Impact of the amount of text supervision on the target accuracy.** LaGTran outperforms strong UDA methods while requiring text supervision from only 20% of samples in GeoNet and 50% in DomainNet, with potential for further enhancement with increased text data.

accuracy of a model trained directly on ego videos, which achieves 33.19%, compared to a model transferred from exo-videos, which only achieves 15.66%. Similarly, a model trained on ego videos achieve only 8.4% for recognition in exo view, compared to a potential 17.91% achievable by training directly on exo videos, indicating a significant domain shift. Current state-of-the-art video adaptation methods (Wei et al., 2023) yield limited gains to bridge these gaps, highlighting the need for novel approaches to address this challenge. Moreover, zeroshot video classification accuracy using EgoVLP (Lin et al., 2022) and LaVILA (Zhao et al., 2023) also show limited gains. Notably, LaGTran which efficiently leverages action descriptions available alongside the videos, achieves an accuracy of 21.55% on average significantly outperforming the source-only baseline by 9% and prior adaptation methods by >4%. LaGTran also outperforms pseudo-labeling using *nGramMatch* or *TextMatch*, as the text descriptions, independently developed from keystep labels, often lack utility for deciphering the action category labels on their own. We also note the substantial scope for further improvement in future, both in terms of the low within-domain accuracy as well as the remaining gap to supervised target accuracy.

4.4. Analysis and Ablations

How much text supervision is needed for LaGTran? Since natural language supervision is fundamental to LaGTran, we analyze the impact of the amount of supervision available on the eventual target accuracy. We retrain LaGTran by assuming text supervision from only $\mu\%$ of images in both source and target domains, where $\mu = \{10, 20, 30, 50, 75, 100\}\%$, and simply discard the target images that do not have corresponding textual supervision. As shown in Fig. 5, LaGTran outperforms image-only method PMTrans (Zhu et al., 2023) even with just 20% text supervision in GeoNet(Fig. 5a) and 50% in DomainNet(Fig. 5b), indicating its high data efficiency. Notably, the graph remains unsaturated, suggesting the potential for further improvement through the collection of more cheaply available text supervision in the target domain.

Model	params (M)	GeoImnet	GeoPlaces	DomainNet
T5-small (Raffel et al., 2020)	60.87	73.93	63.61	68.57
CLIP-T (Radford et al., 2021)	63.16	79.87	66.45	71.15
GPT-2 (Radford et al., 2019)	124	77.88	66.65	69.60
DistilBERT (Devlin et al., 2019)	67.1	**83.53**	**69.31**	**71.43**

Table 5: **Comparison of text-classifier backbones** using text-classification accuracy on GeoNet and DomainNet datasets. BERT backbone outperforms other text-pretrained backbones and vision-language pre-trained CLIP-T.

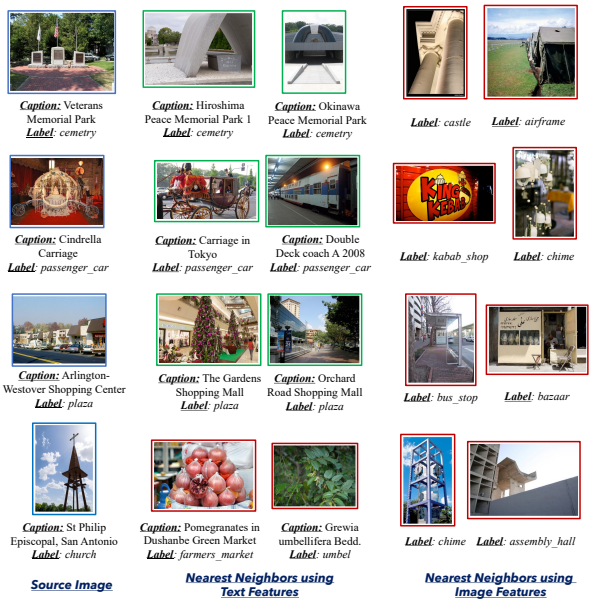


Figure 6: **Visualization of nearest neighbors** of the leftmost source image, using text-trained and image-trained features, along with *ground truth* labels for each image from GeoNet. We observe better "same-class" retrievals using text-captions due to reduced domain gap, as opposed to images.

Effect of text classifier backbone. We compare different choices of text classifiers such as DistilBERT (Sanh et al., 2019), T5-Small (Raffel et al., 2020), GPT2 (Radford et al., 2019) as well as text branch of CLIP (Radford et al., 2021) (CLIP-T) using text-classification accuracy on our datasets. We refer readers to the respective papers for details on their architectures and pre-training datasets. From Tab. 5, DistilBERT yields best text-classification accuracy on all our three benchmarks, outperforming text-only models like T5 and GPT2. Despite large-scale vision-language pre-training, CLIP-T did not yield substantial benefits.

Importance of source domain images. While the majority of our accuracy gains stem from the text guidance, the source domain images providing noise-free supervision are also important in learning strong models on the target

domain. We observed that joint training using source and pseudo-labeled target yields improvements of 1.57% for DomainNet and 0.8% on Ego2Exo benchmarks compared to target-only training. More importantly, training jointly on source and target allows deploying a single joint model across both domains as opposed to domain specific models, greatly optimizing inference costs.

Nearest neighbors using image and text features. We show the top-2 nearest neighbor retrievals using text-features computed from source-trained text-classifier as opposed to image-features in Fig. 6. We observe more robust retrievals based on text-features corresponding to the captions of the images, rather than the images directly signifying the reduced domain gap in the text space. We also note a failure case in the last row of Fig. 6, where neither the text features nor the image features could retrieve the image from the correct class *church*.

5. Conclusion and Future Directions

We introduce a novel framework called `LaGTran` to use readily available text supervision and enhance target performance in unsupervised domain transfer scenarios. We first start with the observation that traditional domain alignment approaches yield limited benefits beyond well-understood domain shifts, followed by insights that language provides a semantically richer medium of transfer with reduced domain gaps. This leads to a language-guided transfer mechanism where we train a text classifier on language descriptions from a source domain and then use its predictions on descriptions from a different target domain as supervision for the corresponding images. Despite being conceptually simple and straightforward, we show the remarkable ability of our method to outperform competitive prior approaches on challenging benchmarks like GeoNet and DomainNet for images and proposed Ego2Exo for videos. Through an emphasis on cost-effective or easily producible text supervision, we open new possibilities for advancing domain transfer in scenarios with limited manual supervision. Although `LaGTran` achieves state-of-the-art performance across several datasets, it relies on external vision-language models for textual guidance in the absence of metadata, potentially constraining its applicability in scenarios where the pre-trained VLM models fail to offer discriminative text supervision. Additionally, while exhibiting fewer domain discrepancies, there remain non-trivial gaps even within the text modality that may reduce the accuracy of pseudo-labels in the target domain, which can be potentially addressed by additionally incorporating text-adaptation mechanisms.

Acknowledgements

We acknowledge support from NSF and a Google Award for Inclusion Research.

Impact Statement

Our paper presents an approach that can improve accuracy on domains facing label scarcity. Advancing this research area would enhance wider adoption of current AI technologies, and unlocks new capabilities in democratizing the progress in AI. Given that our proposed methodology only operates in the standard realm of image classification and our showcased results only use already publicly available datasets, we do not foresee any negative societal consequences specifically arising due to our method.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

Andreassen, A., Bahri, Y., Neyshabur, B., and Roelofs, R. The evolution of out-of-distribution robustness throughout fine-tuning. *arXiv preprint arXiv:2106.15831*, 2021. 3

Bain, M., Nagrani, A., Varol, G., and Zisserman, A. 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*, pp. 1728–1738, 2021. 3

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137–144, 2006. 2, 3

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010. 2, 3

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., and Erhan, D. Domain separation networks. In *Advances in neural information processing systems*, pp. 343–351, 2016. 2

Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568, 2021. 3

Chen, C., Xie, W., Huang, W., Rong, Y., Ding, X., Huang, Y., Xu, T., and Huang, J. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 627–636, 2019a. 2, 7

Chen, L., Chen, H., Wei, Z., Jin, X., Tan, X., Jin, Y., and Chen, E. Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7181–7190, June 2022. 1, 6

Chen, M.-H., Kira, Z., AlRegib, G., Yoo, J., Chen, R., and Zheng, J. Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6321–6330, 2019b. 2, 6, 7, 16

Chen, S., Zhang, Y., Jiang, W., Lu, J., and Zhang, Y. Large language models as visual cross-domain learners. *arXiv preprint arXiv:2401.03253*, 2024. 2

Choi, J., Sharma, G., Chandraker, M., and Huang, J.-B. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1717–1726, 2020. 2, 6

Dasgupta, A., Jawahar, C., and Alahari, K. Overcoming label noise for source-free unsupervised video domain adaptation. In *Proceedings of the Thirteenth Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 1–9, 2022. 2

Deng, Z., Luo, Y., and Zhu, J. Cluster alignment with a teacher for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9944–9953, 2019. 2

Desai, K. and Johnson, J. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11162–11173, 2021. 3

Desai, K., Kaul, G., Aysola, Z., and Johnson, J. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021. 3

Devillers, B., Choksi, B., Bielawski, R., and VanRullen, R. Does language help generalization in vision models? *arXiv preprint arXiv:2104.08313*, 2021. 3

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics*, 2019. doi: 10.18653/v1/N19-1423. 4, 8

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5, 15

Du, Z., Li, J., Su, H., Zhu, L., and Lu, K. Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3937–3946, 2021. 6, 7

Dunlap, L., Mohri, C., Guillory, D., Zhang, H., Darrell, T., Gonzalez, J. E., Raghunathan, A., and Rohrbach, A. Using language to extend to unseen domains. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=eR2dG8yjnQ. 3, 5

El Banani, M., Desai, K., and Johnson, J. Learning visual representations via language-guided sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19208–19220, 2023. 3

French, G., Mackiewicz, M., and Fisher, M. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208*, 2017. 2

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015. 1, 2

Girdhar, R., Singh, M., Ravi, N., van der Maaten, L., Joulin, A., and Misra, I. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16102–16112, 2022. 7, 16

Gokhale, T., Anirudh, R., Kailkhura, B., Thiagarajan, J. J., Baral, C., and Yang, Y. Attribute-guided adversarial training for robustness to natural perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7574–7582, 2021. 3

Goyal, P., Duval, Q., Seessel, I., Caron, M., Misra, I., Sagun, L., Joulin, A., and Bojanowski, P. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv preprint arXiv:2202.08360*, 2022. 3

Goyal, S., Kumar, A., Garg, S., Kolter, Z., and Raghunathan, A. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19338–19347, 2023. 3

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., 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*, pp. 18995–19012, 2022. 3

Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *arXiv preprint arXiv:2311.18259*, 2023. 3, 6, 15

Gu, X., Sun, J., and Xu, Z. Spherical space domain adaptation with robust pseudo-label loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9101–9110, 2020. 2

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 5, 6

Huang, Y., Chen, G., Xu, J., Zhang, M., Yang, L., Pei, B., Zhang, H., Dong, L., Wang, Y., Wang, L., et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. *arXiv preprint arXiv:2403.16182*, 2024. 3, 6

Huang, Z., Zhou, A., Ling, Z., Cai, M., Wang, H., and Lee, Y. J. A sentence speaks a thousand images: Domain generalization through distilling clip with language guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11685–11695, 2023. 3

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021. 3

Kalluri, T. and Chandraker, M. Cluster-to-adapt: Few shot domain adaptation for semantic segmentation across disjoint labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4121–4131, 2022. 2

Kalluri, T., Sharma, A., and Chandraker, M. Memsac: Memory augmented sample consistency for large scale domain adaptation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pp. 550–568. Springer, 2022. 2, 5, 6, 7

Kalluri, T., Xu, W., and Chandraker, M. Geonet: Benchmarking unsupervised adaptation across geographies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15368–15379, June 2023. 1, 2, 5, 6, 7

Kang, G., Jiang, L., Yang, Y., and Hauptmann, A. G. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4893–4902, 2019. 2

Kumar, A., Sattigeri, P., Wadhawan, K., Karlinsky, L., Feris, R., Freeman, B., and Wornell, G. Co-regularized alignment for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 9345–9356, 2018. 2

Kumar, A., Ma, T., and Liang, P. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, pp. 5468–5479. PMLR, 2020. 2

Kumar, A., Raghunathan, A., Jones, R., Ma, T., and Liang, P. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022. 3

Lai, Z., Vesdapunt, N., Zhou, N., Wu, J., Huynh, C. P., Li, X., Fu, K. K., and Chuah, C.-N. Padclip: Pseudo-labeling with adaptive debiasing in clip for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16155–16165, 2023. 2

Lai, Z., Bai, H., Zhang, H., Du, X., Shan, J., Yang, Y., Chuah, C.-N., and Cao, M. Empowering unsupervised domain adaptation with large-scale pre-trained vision-language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2691–2701, 2024. 2

Li, J., Li, D., Savarese, S., and Hoi, S. C. H. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International Conference on Machine Learning*, 2023. doi: 10.48550/arXiv.2301.12597. 3, 5

Li, S., Xie, M., Lv, F., Liu, C. H., Liang, J., Qin, C., and Li, W. Semantic concentration for domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9102–9111, 2021a. 2, 6, 7

Li, Y., Nagarajan, T., Xiong, B., and Grauman, K. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6943–6953, 2021b. 2, 6

Lin, K. Q., Wang, J., Soldan, M., Wray, M., Yan, R., XU, E. Z., Gao, D., Tu, R.-C., Zhao, W., Kong, W., et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022. 3, 7, 8, 15

Liu, G. and Wang, Y. Tdg: Text-guided domain generalization. *arXiv preprint arXiv:2308.09931*, 2023. 3, 5

Liu, H., Wang, J., and Long, M. Cycle self-training for domain adaptation. *Advances in Neural Information Processing Systems*, 34:22968–22981, 2021a. 2

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *NEURIPS*, 2023. 3

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021b. 5, 15

Long, M., Zhu, H., Wang, J., and Jordan, M. I. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pp. 2208–2217. PMLR, 2017. 2

Long, M., Cao, Z., Wang, J., and Jordan, M. I. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 1640–1650, 2018. 1, 2, 6

Luo, Y., Zheng, L., Guan, T., Yu, J., and Yang, Y. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2507–2516, 2019. 2

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and Van Der Maaten, L. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 181–196, 2018. 3

Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., and Sivic, J. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2630–2640, 2019. 3

Min, S., Park, N., Kim, S., Park, S., and Kim, J. Grounding visual representations with texts for domain generalization. In *European Conference on Computer Vision*, pp. 37–53. Springer, 2022. 3

Munro, J. and Damen, D. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 122–132, 2020. 6

Ohkawa, T., Yagi, T., Nishimura, T., Furuta, R., Hashimoto, A., Ushiku, Y., and Sato, Y. Exo2egodvc: Dense video captioning of egocentric procedural activities using web instructional videos. *arXiv preprint arXiv:2311.16444*, 2023. 2, 6

Park, C., Lee, J., Yoo, J., Hur, M., and Yoon, S. Joint contrastive learning for unsupervised domain adaptation. *arXiv preprint arXiv:2006.10297*, 2020. 2

Pei, Z., Cao, Z., Long, M., and Wang, J. Multi-adversarial domain adaptation. *arXiv preprint arXiv:1809.02176*, 2018. 2

Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., and Saenko, K. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017. 1

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1406–1415, 2019. 5

Pham, H., Dai, Z., Ghiasi, G., Kawaguchi, K., Liu, H., Yu, A. W., Yu, J., Chen, Y.-T., Luong, M.-T., Wu, Y., et al. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555:126658, 2023. 3

Plizzari, C., Perrett, T., Caputo, B., and Damen, D. What can a cook in italy teach a mechanic in india? action recognition generalisation over scenarios and locations. *arXiv preprint arXiv: 2306.08713*, 2023. 6

Prabhu, V., Selvaraju, R. R., Hoffman, J., and Naik, N. Can domain adaptation make object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3981–3988, 2022. 1, 2

Quattrocchi, C., Furnari, A., Di Mauro, D., Giuffrida, M. V., and Farinella, G. M. Synchronization is all you need: Exocentric-to-egocentric transfer for temporal action segmentation with unlabeled synchronized video pairs. *arXiv preprint arXiv:2312.02638*, 2023. 2, 6

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 4, 8, 16

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021. 3, 5, 6, 7, 8, 16

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 4, 8, 16

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021. 3

Reimers, N. and Gurevych, I. 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, 11 2019. URL http://arxiv.org/abs/1908.10084. 5

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022. 3

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pp. 213–226. Springer, 2010. 1

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022. 3

Sahoo, A., Shah, R., Panda, R., Saenko, K., and Das, A. Contrast and mix: Temporal contrastive video domain adaptation with background mixing. *Advances in Neural Information Processing Systems*, 34, 2021. 2

Saito, K. and Saenko, K. Ovanet: One-vs-all network for universal domain adaptation. *IEEE International Conference on Computer Vision*, 2021. doi: 10.1109/ICCV48922.2021.00887. 6, 7

Saito, K., Ushiku, Y., Harada, T., and Saenko, K. Adversarial dropout regularization. *arXiv preprint arXiv:1711.01575*, 2017. 2

Saito, K., Watanabe, K., Ushiku, Y., and Harada, T. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3723–3732, 2018. 1, 2, 6, 7

Saito, K., Kim, D., Sclaroff, S., and Saenko, K. Universal domain adaptation through self supervision. *Advances in neural information processing systems*, 33:16282–16292, 2020. 4, 5, 6, 7

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *NEURIPS*, 2019. 4, 8, 16

Sariyildiz, M. B., Perez, J., and Larlus, D. Learning visual representations with caption annotations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pp. 153–170. Springer, 2020. 3

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022. 3

Sharma, A., Kalluri, T., and Chandraker, M. Instance level affinity-based transfer for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5361–5371, 2021. 1, 2

Sigurdsson, G. A., Gupta, A., Schmid, C., Farhadi, A., and Alahari, K. Actor and observer: Joint modeling of first and third-person videos. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7396–7404, 2018. 6

Singh, M., Gustafson, L., Adcock, A., de Freitas Reis, V., Gedik, B., Kosaraju, R. P., Mahajan, D., Girshick, R., Dollár, P., and Van Der Maaten, L. Revisiting weakly supervised pre-training of visual perception models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 804–814, 2022. 3

Sun, B. and Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pp. 443–450. Springer, 2016. 2

Sun, T., Lu, C., Zhang, T., and Ling, H. Safe self-refinement for transformer-based domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7191–7200, 2022. 2, 6, 7

Tan, S., Peng, X., and Saenko, K. Class-imbalanced domain adaptation: an empirical odyssey. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 585–602. Springer, 2020. 2

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 3

Tzeng, E., Hoffman, J., Darrell, T., and Saenko, K. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE international conference on computer vision*, pp. 4068–4076, 2015. 2

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176, 2017. 2

Udandarao, V., Gupta, A., and Albanie, S. Sus-x: Training-free name-only transfer of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2725–2736, 2023. 3

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5018–5027, 2017. 1

Wang, R., Wu, Z., Weng, Z., Chen, J., Qi, G.-J., and Jiang, Y.-G. Cross-domain contrastive learning for unsupervised domain adaptation. *IEEE Transactions on Multimedia*, 2022. 2

Wang, Z., Zhang, L., Wang, L., and Zhu, M. Landa: Language-guided multi-source domain adaptation. *arXiv preprint arXiv:2401.14148*, 2024. 3

Wei, G., Lan, C., Zeng, W., Zhang, Z., and Chen, Z. Toalign: Task-oriented alignment for unsupervised domain adaptation. In *NeurIPS*, 2021. 1, 2, 5, 6

Wei, P., Kong, L., Qu, X., Ren, Y., Xu, Z., Jiang, J., and Yin, X. Unsupervised video domain adaptation for action recognition: A disentanglement perspective. In *Advances in Neural Information Processing Systems*, 2023. 2, 6, 7, 8, 16

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7959–7971, 2022. 3

Xie, S., Zheng, Z., Chen, L., and Chen, C. Learning semantic representations for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 5423–5432, 2018. 2

Xu, J., Huang, Y., Hou, J., Chen, G., Zhang, Y., Feng, R., and Xie, W. Retrieval-augmented egocentric video captioning. *arXiv preprint arXiv:2401.00789*, 2024. 3

Xu, R., Li, G., Yang, J., and Lin, L. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1426–1435, 2019. 1

Xu, T., Chen, W., Wang, P., Wang, F., Li, H., and Jin, R. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*, 2021. 2, 7

Xue, Z. S. and Grauman, K. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. *Advances in Neural Information Processing Systems*, 36, 2023. 2

You, K., Long, M., Cao, Z., Wang, J., and Jordan, M. I. Universal domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2720–2729, 2019. 4, 5, 6, 7

Zhang, X., Gu, S. S., Matsuo, Y., and Iwasawa, Y. Domain prompt learning for efficiently adapting clip to unseen domains. *Transactions of the Japanese Society for Artificial Intelligence*, 38(6):B–MC2_1, 2023. 2

Zhang, Y., Liu, T., Long, M., and Jordan, M. I. Bridging theory and algorithm for domain adaptation. *arXiv preprint arXiv:1904.05801*, 2019. 6, 7

Zhao, Y., Misra, I., Krähenbühl, P., and Girdhar, R. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6586–6597, 2023. 3, 7, 8

Zhu, J., Bai, H., and Wang, L. Patch-mix transformer for unsupervised domain adaptation: A game perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3561–3571, 2023. 1, 2, 5, 6, 7, 8, 15

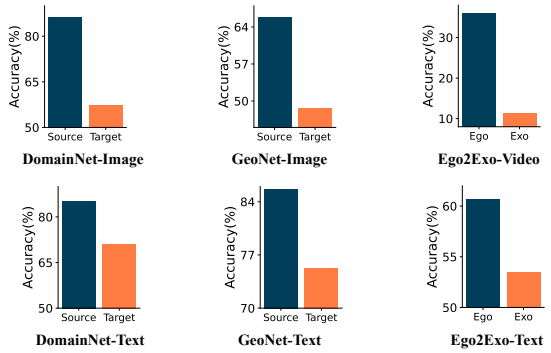


Figure 7: **Domain Robustness Text vs Image**: Cross-domain accuracy of image or video classifier transfer across domains compared to text modality. As opposed to significantly large domain drops when transferring image-based models across domains, text-classifier based models are surprisingly robust on all the studied benchmarks, leading to minimal domain drops and high accuracy.

A. Illustrating Cross-domain Robustness

We illustrate the cross-domain robustness properties of image vs text classifiers in Fig. 7. We show the remarkably powerful target-domain models obtained by transferring the text classifier, as opposed to image-based models which suffer high domain gap. This behavior is consistent across all the datasets studied, and forms the backbone of our motivations in leveraging textual guidance in performing unsupervised transfer across domains.

B. Construction of Ego2Exo benchmark

We curate Ego2Exo dataset from the larger Ego-Exo4D dataset (Grauman et al., 2023). Specifically, the *keystep annotations* provided along with Ego-Exo4D offer fine-grained action labels for several short video clips, called *segments*, that are manually trimmed from long procedural videos to focus on the keysteps to recognize. We restrict focus to videos from *cooking* activity, as they include the largest set of segments and labels capturing a diverse pool of actions. Moreover, to ease the difficult task of predicting very fine-grained action classes from short segments (eg: `add coffee beans` vs. `add coffee grounds`), we use the provided label hierarchy and manually remap the original 96 annotated actions labels into 24 labels by merging similar classes into a larger, common class. The final category list is as follows:

1. Cook
2. Serve
3. Clean up
4. Add water
5. Make dough
6. Make pasta
7. Make salad
8. Make chai tea
9. Make milk tea
10. Get Ingredients
11. Prepare dressing
12. Prepare a skillet
13. Add spring onions
14. Turn off the stove
15. Check paper recipe
16. Prepare ingredients
17. Prepare milk (boiled)
18. Construct undressed salad
19. Cook noodles in a skillet
20. Get kitchenware & utensils
21. Brew coffee (instant coffee)
22. Boil noodles in boiling water
23. Brew coffee (manual pour-over)
24. Mix noodles with sauce in a bowl

To provide text supervision to our algorithm, we use the *atomic action descriptions* provided in Ego-Exo4D dataset. These descriptions provide a narrative of the events in the video, presented in free-form text from the perspective of a third-party observer. Unlike keystep labels, which are defined between specific start and end times within a video, these text descriptions are associated with distinct timestamps, or a single point in time within the video. To create correspondence mapping between the keystep segments and text descriptions, we adopt the method outlined in EgoVLP (Lin et al., 2022) as follows: to generate a text description for a segment, we compile all text descriptions that fall within the timestamps defined by the start and end times of that segment. If multiple timestamps exist, we concatenate the corresponding texts; if no timestamps are available, we include no associated text with the segment. Furthermore, we concatenate the annotations provided by multiple annotators in creating the text description.

Our proposed Ego2Exo consists of video segments labeled with actions from one of the 24 keysteps, with corresponding text descriptions for each segment. We split these video segments into two equal groups classwise, and collect ego-videos from one group and exo-videos from the other to create our adaptation benchmark. The same procedure applied to the validation videos yields 3147 validation segments with both ego and exo views. The `json` file containing our complete set of videos (referenced from Ego-Exo4D dataset), along with annotations and text descriptions is available along with our publicly released code.

C. Training Details

Image Classifier We use a ViT-base (Dosovitskiy et al., 2020) backbone as the image encoder on the GeoNet dataset, and follow prior work (Zhu et al., 2023) and use Swin-base backbone (Liu et al., 2021b) for experiments on the DomainNet data. Both the backbones are pre-trained on

ImageNet-1k, and we add a 2-layer MLP on top of the computed features as the classifier head. Across all transfer settings, we train these backbones for 90,000 iterations using the objective function specified in Eq. (3), employing SGD with a learning rate of 3e-4 and batch size of 64 from each domain, along with a cosine decay schedule.

Text Classifier We use a pre-trained Distill-BERT (Sanh et al., 2019) model from HuggingFace as the sentence classification model $\mathcal{B}(;\phi)$, and fine-tune it for five epochs over the source domain data using AdamW optimizer with a learning rate of 5e-5 and cosine decay over the training schedule. We observed sub-optimal performance using other pre-trained backbones such as T5 (Raffel et al., 2020), GPT-2 (Radford et al., 2019) or text encoder in CLIP (Radford et al., 2021) (Sec. 4.4).

Video Classifier We use the pre-computed Omnivore-base (Girdhar et al., 2022) features provided along with the EgoExo4D dataset for training and evaluation. Since different keysteps may be represented by largely different timespans (Fig. 4a), we collect all features that fall within the start and end times of a segment, and pool these features together to form a 1536-dimensional feature representation of that segment. We then train a 2-layer MLP classifier on top of these features, using the labeled source feature as well as psuedo-labeled target features following Eq. (3). Note that this training strategy is equivalent to training an MLP classifier on top of frozen Omnivore backbone. For fair comparison, we follow the same strategy for training all the other baselines as well as prior adaptation methods. For methods that require a temporal sequence of features (Wei et al., 2023; Chen et al., 2019b), we sample 8 equally spaced features from the complete set of segment features, and use this feature sequence as input. We follow similar strategy for evaluation, and use features pre-extracted from the validation videos for testing. We use the top-1 accuracy on the validation set for evaluation.