APPENDIX
Bridging the Domain Gap in Small Multimodal Models: A Dual-level Alignment Perspective

Aveen Dayal[1] Peketi Divya[1] Nidhi Tiwari [2] Linga Reddy Cenkeramaddi[3]
C Krishna Mohan[1] Abhinav Kumar[1]
[1]Indian Institute of Technology Hyderabad, Hyderabad, India
[2] Microsoft, Hyderabad, India
[3]University of Agder, Grimstad, Norway
ai21resch11003@iith.ac.in, ai21resch01001@iith.ac.in, nidhitiwari@microsoft.com,
linga.cenkeramaddi@uia.no, ckm@cse.iith.ac.in, abhinavkumar@ee.iith.ac.in

A1. Proofs

Lemma 1. *Let $\mathcal{D}_S$ be the source domain, and let $\mathcal{D}_{\text{mid}}$ be an intermediate distribution constructed so that $\mathcal{D}^Z_{\text{mid}} = \mathcal{D}^Z_T$,$\Pr_{\mathcal{D}_{\text{mid}}}(x^Q \mid z) = \Pr_{\mathcal{D}_S}(x^Q \mid z)$ Then for any measurable event A in the space of (x^V, x^Q),*

$$\left|\Pr_{\mathcal{D}_S}[A] - \Pr_{\mathcal{D}_{\text{mid}}}[A]\right| \leq d_{\text{enc}}\big(\mathcal{D}^Z_S,\ \mathcal{D}^Z_T\big) \tag{1}$$

Proof. Step 1: Since $\mathcal{D}_{\text{mid}}$ has the same marginal distribution of z as $\mathcal{D}_T$, we write $\mathcal{D}^Z_{\text{mid}} = \mathcal{D}^Z_T$. We recall that $d_{\text{enc}}(\mathcal{D}^Z_S, \mathcal{D}^Z_T) = \sup_{\phi\in\Phi}\big|\Pr_{\mathcal{D}^Z_S}[\phi(z) = 1] - \Pr_{\mathcal{D}^Z_T}[\phi(z) = 1]\big|$.
Step 2: In $\mathcal{D}_{\text{mid}}$, the conditional distribution of $(x^Q \mid z)$ remains the same as in $\mathcal{D}_S$. Hence, for any event A expressible in terms of (x^V, x^Q) (equivalently (z, x^Q)), we can decompose as shown in Eqn. (2)

$$\begin{aligned}\Pr_{\mathcal{D}_S}[A] &= \int \Pr_{\mathcal{D}_S}\big[A \mid z\big]\, d\Pr_{\mathcal{D}^Z_S}(z),\\ \Pr_{\mathcal{D}_{\text{mid}}}[A] &= \int \Pr_{\mathcal{D}_S}\big[A \mid z\big]\, d\Pr_{\mathcal{D}^Z_{\text{mid}}}(z).\end{aligned} \tag{2}$$

only difference is that the outside integral for $\mathcal{D}_{\text{mid}}$ is taken w.r.t. $\mathcal{D}^Z_{\text{mid}} = \mathcal{D}^Z_T$.
Step 3: Thus,

$$\begin{aligned}\left|\Pr_{\mathcal{D}_S}[A] - \Pr_{\mathcal{D}_{\text{mid}}}[A]\right| \ =\\ \left|\int \Pr_{\mathcal{D}_S}[A \mid z]\, d\big(\Pr_{\mathcal{D}^Z_S} - \Pr_{\mathcal{D}^Z_{\text{mid}}}\big)(z)\right|.\end{aligned} \tag{3}$$

By definition, $\mathcal{D}^Z_{\text{mid}} = \mathcal{D}^Z_T$, so this difference in integrals is bounded by a measure of how $\mathcal{D}^Z_S$ differs from $\mathcal{D}^Z_T$. Specifically,

$$\left|\Pr_{\mathcal{D}^Z_S} - \Pr_{\mathcal{D}^Z_T}\right| \ \leq\ d_{\text{enc}}(\mathcal{D}^Z_S, \mathcal{D}^Z_T). \tag{4}$$

Hence,

$$\left|\Pr_{\mathcal{D}_S}[A] - \Pr_{\mathcal{D}_{\text{mid}}}[A]\right| \leq d_{\text{enc}}\big(\mathcal{D}^Z_S, \mathcal{D}^Z_T\big). \tag{5}$$

This completes the proof. □

Lemma 2. *Let $\mathcal{D}_{\text{mid}}$ be as defined in Lemma 1.Suppose $\mathcal{D}_T$ differs from $\mathcal{D}_{\text{mid}}$ only in the conditional distribution of x^Q given z. Then for any measurable event A,*

$$\left|\Pr_{\mathcal{D}_{\text{mid}}}[A] - \Pr_{\mathcal{D}_T}[A]\right| \ \leq\ d_{\text{fus}}\Big(\mathcal{D}^{(Z,Q)}_{\text{mid}},\ \mathcal{D}^{(Z,Q)}_T\Big) \tag{6}$$

Proof. Step 1: By assumption, $\mathcal{D}^Z_{\text{mid}} = \mathcal{D}^Z_T$. Hence, $(z, x^Q) \sim \mathcal{D}^{(Z,Q)}_{\text{mid}}$ differs from $(z, x^Q) \sim \mathcal{D}^{(Z,Q)}_T$ only through how x^Q is drawn given z.
Step 2: We recall that

$$\begin{aligned}d_{\text{fus}}\big(\mathcal{D}^{(Z,Q)}_1,\ \mathcal{D}^{(Z,Q)}_2\big) = \sup_{\psi\in\Psi}\Big|\Pr_{\mathcal{D}^{(Z,Q)}_1}[\psi(z, x^Q) = 1]\\ - \Pr_{\mathcal{D}^{(Z,Q)}_2}[\psi(z, x^Q) = 1]\Big|\end{aligned} \tag{7}$$

for some family Ψ of binary functions on (z, x^Q). It measures how distinguishable (z, x^Q) is between two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$.
Step 3: Because $\mathcal{D}_{\text{mid}}$ and $\mathcal{D}_T$ coincide on the marginal distribution of z, the probability that an event A occurs can differ only if A depends on how x^Q is paired with z. In other words,

$$\left|\Pr_{\mathcal{D}_{\text{mid}}}[A] - \Pr_{\mathcal{D}_T}[A]\right| \ \leq\ d_{\text{fus}}\big(\mathcal{D}^{(Z,Q)}_{\text{mid}}, \mathcal{D}^{(Z,Q)}_T\big) \tag{8}$$

by precisely the argument that if A can be rephrased as $\{(z, x^Q) \in \mathcal{S}\}$, then the difference in probabilities of A is at most how different the joint (z, x^Q)-distributions are. That difference is measured by d_{fus}.

□

Theorem 1. *Let $h(x^V, x^Q) = f(g(x^V), x^Q)$ be a multimodal model, and let $R_S(h)$ and $R_T(h)$ denote its 0–1 risk on the source $\mathcal{D}_S$ and target $\mathcal{D}_T$ respectively. Suppose $d_{\text{enc}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$ and $d_{\text{fus}}(\mathcal{D}_S^{(Z,Q)}, \mathcal{D}_T^{(Z,Q)})$ are the encoder- and fusion-level divergences. Then for any h,*

$$R_T(h) \le R_S(h) + d_{\text{enc}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z) + d_{\text{fus}}(\mathcal{D}_S^{(Z,Q)}, \mathcal{D}_T^{(Z,Q)}) + \lambda. \tag{9}$$

where $\lambda = \min_{h' \in \{f' \circ g'\}} \Big[R_S(h') + R_T(h')\Big]$ be the minimal combined error across source and target in the hypothesis class.

Proof. Step 1: Let

$$R_T(h) - R_S(h) = \Big[R_T(h) - R_T(h')\Big] - \Big[R_S(h) - R_S(h')\Big] + \big[R_T(h') - R_S(h')\big]. \tag{10}$$

After rearranging terms we get,

$$R_T(h) = R_S(h) + \Big[R_T(h) - R_T(h')\Big] - \Big[R_S(h) - R_S(h')\Big] + \big[R_T(h') - R_S(h')\big]. \tag{11}$$

Since $R_T(h') - R_S(h') \le R_T(h') + R_S(h') = \lambda$, we get

$$R_T(h) \le R_S(h) + \Big[R_T(h) - R_T(h')\Big] - \Big[R_S(h) - R_S(h')\Big] + \lambda. \tag{12}$$

Step2: Define $A = \{(x^V, x^Q) \mid h(x^V, x^Q) \neq h'(x^V, x^Q)\}$. Following Lemma 3 from [4] we show,

$$\big(R_T(h) - R_T(h')\big) - \big(R_S(h) - R_S(h')\big) = \Pr_T[A] - \Pr_S[A]. \tag{13}$$

Hence,

$$\big|\big(R_T(h) - R_T(h')\big) - \big(R_S(h) - R_S(h')\big)\big| = \big|\Pr_T[A] - \Pr_S[A]\big| \tag{14}$$

Step3: By constructing $\mathcal{D}_{\text{mid}}$ as in Lemma 1 and Lemma 2, we have

$$\big|\Pr_T[A] - \Pr_S[A]\big| \le \big|\Pr_{\text{mid}}[A] - \Pr_S[A]\big| + \big|\Pr_T[A] - \Pr_{\text{mid}}[A]\big| \le d_{\text{enc}} + d_{\text{fus}} \tag{15}$$

Thus,

$$\Big|\big(R_T(h) - R_T(h')\big) - \big(R_S(h) - R_S(h')\big)\Big| = \big|\Pr_T[A] - \Pr_S[A]\big| \le d_{\text{enc}} + d_{\text{fus}}. \tag{16}$$

Putting the above relation in Eqn. (12), we get the bound as shown in Theorem (1) □

A2. Dual-Divergence Risk for Generic Bounded Loss

Let $\ell\colon \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ be an arbitrary loss function bounded by 1.[1] For any hypothesis h and domain $\mathcal{D}$ we write $R_{\mathcal{D}}^{\ell}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell\big(h(x), y\big)\big]$. When ℓ is the sequence-level 0–1 loss $\ell(h(x), y) = \mathbb{I}[h(x) \neq y]$, this reduces to the risk $R_{\mathcal{D}}(h)$ used in the main paper.

Loss-based dual divergences. For a bounded loss ℓ define the discrepancy distance [A5] between two distributions P, Q over a space Ω with respect to a hypothesis class $\mathcal{H}$ as $\text{disc}_{\ell}^{\mathcal{H}}(P, Q) = \sup_{h,h' \in \mathcal{H}} \big|\mathbb{E}_P\big[\ell(h(\omega), h'(\omega))\big] - \mathbb{E}_Q\big[\ell(h(\omega), h'(\omega))\big]\big|$. Choosing $\mathcal{H} = \Phi$ and Ψ (as in Eqns. (1) - (2), in the main paper) yields the encoder-level and fusion-level divergences as shown.

$$d_{\text{enc}}^{\ell}(\mathcal{D}_S^Z, \mathcal{D}_T^Z) = \text{disc}_{\ell}^{\Phi}(\mathcal{D}_S^Z, \mathcal{D}_T^Z), \quad d_{\text{fus}}^{\ell}(\mathcal{D}_S^{(Z,Q)}, \mathcal{D}_T^{(Z,Q)}) = \text{disc}_{\ell}^{\Psi}(\mathcal{D}_S^{(Z,Q)}, \mathcal{D}_T^{(Z,Q)}). \tag{17}$$

Let $\mathcal{D}_{mid}$ be the "mix-and-match" distribution used in the 0–1 proof: it satisfies $\mathcal{D}_{mid}^Z = \mathcal{D}_T^Z$ and $\Pr_{\mathcal{D}_{mid}}(x^Q \mid z) = \Pr_{\mathcal{D}_S}(x^Q \mid z)$. Labels $(y \mid x)$ are unchanged. Hence, $\mathcal{D}_{mid}$ lives on the same sample space (x^V, x^Q, y) as $\mathcal{D}_S$ and $\mathcal{D}_T$.

Lemma 3. *For any hypotheses h, h'*

$$\Big|\mathbb{E}_{\mathcal{D}_S}\big[\ell(h(x), h'(x))\big] - \mathbb{E}_{\mathcal{D}_{mid}}\big[\ell(h(x), h'(x))\big]\Big| \le d_{enc}^{\ell}(\mathcal{D}_S^Z, \mathcal{D}_T^Z). \tag{18}$$

Proof. Because $\mathcal{D}_S$ and $\mathcal{D}_{mid}$ differ only in the marginal z, write the expectation as a two-stage integral:

$$\mathbb{E}_{\mathcal{D}_S}\big[\ell(h, h')\big] = \int_z \phi_{h,h'}(z)\, d\mathcal{D}_S^Z(z), \tag{19}$$

where $\phi_{h,h'}(z) = \mathbb{E}_{x^Q \sim \mathcal{D}_S(x^Q \mid z)}\big[\ell(h(x), h'(x))\big]$. Similarly, we have

$$\mathbb{E}_{\mathcal{D}_{mid}}\big[\ell(h, h')\big] = \int_z \phi_{h,h'}(z)\, d\mathcal{D}_{mid}^Z(z), \tag{20}$$

[1] The bounded-loss setting follows the discrepancy framework of Cortes et al. [A5], which generalizes the 0–1 analysis of Ben-David et al. [A2].

Since $\mathcal{D}^Z_{mid} = \mathcal{D}^Z_T$ we get

$$\Big|\mathbb{E}_{\mathcal{D}_S}\big[\ell(h,h')\big] - \mathbb{E}_{\mathcal{D}_{mid}}\big[\ell(h,h')\big]\Big| = \Big|\int_z \phi_{h,h'}(z)\, d(\mathcal{D}^Z_S - \mathcal{D}^Z_{mid})(z)\Big| \tag{21}$$

Since $\ell \in [0,1]$, each integrand $\phi_{h,h'}(z) \in [0,1]$ and is measurable in z. Hence, for every (h,h') the map $z \mapsto \phi_{h,h'}(z)$ is an admissible discriminator in the supremum that defines $\mathrm{disc}^{\Phi}_{\ell}(\mathcal{D}^Z_S, \mathcal{D}^Z_T) = d^{\ell}_{enc}(\mathcal{D}^Z_S, \mathcal{D}^Z_T)$, yielding the claim. □

Lemma 4. *For any hypotheses h, h'*

$$\Big|\mathbb{E}_{\mathcal{D}_{mid}}\big[\ell(h(x),h'(x))\big] - \mathbb{E}_{\mathcal{D}_T}\big[\ell(h(x),h'(x))\big]\Big| \le d^{\ell}_{\mathrm{fus}}(\mathcal{D}^{(Z,Q)}_S, \mathcal{D}^{(Z,Q)}_T). \tag{22}$$

Proof. $\mathcal{D}_{mid}$ and $\mathcal{D}_T$ share the same marginal over z but differ in the conditional $(x^Q \mid z)$. Therefore they differ only in the joint (z, x^Q) distribution. The inner expectation $\ell\big(h(x), h'(x)\big)$ depends on (x^V, x^Q) (and hence on (z, x^Q)) but not on any unobserved variable. Consequently the difference of expectations is upper-bounded by the ℓ-discrepancy $\mathrm{disc}_{\ell}\big(\mathcal{D}^{(Z,Q)}_{mid}, \mathcal{D}^{(Z,Q)}_T\big) = d^{\ell}_{\mathrm{fus}}\big(\mathcal{D}^{(Z,Q)}_{mid}, \mathcal{D}^{(Z,Q)}_T\big)$. □

Theorem 2. *Let $h(x^V, x^Q) = f\big(g(x^V), x^Q\big)$ be a multimodal hypothesis, and let $R^{\ell}_S(h)$, $R^{\ell}_T(h)$ denote its ℓ-risk on $\mathcal{D}_S$ and $\mathcal{D}_T$, respectively. Then, for every bounded loss ℓ and for every h,*

$$R^{\ell}_T(h) \le R^{\ell}_S(h) + d^{\ell}_{\mathrm{enc}}\big(\mathcal{D}^Z_S, \mathcal{D}^Z_T\big) + d^{\ell}_{\mathrm{fus}}\big(\mathcal{D}^{(Z,Q)}_S, \mathcal{D}^{(Z,Q)}_T\big) + \lambda^{\ell}, \tag{23}$$

where $\lambda^{\ell} = \min_{h' \in \{f' \circ g'\}} \big[R^{\ell}_S(h') + R^{\ell}_T(h')\big]$ is the ideal joint error in the hypothesis class.

Proof. Pick any hypothesis h and let

$$h^{\star} = argmin_{h' \in \{f' \circ g'\}} \big[R^{\ell}_S(h') + R^{\ell}_T(h')\big], \quad \lambda^{\ell} = R^{\ell}_S(h^{\star}) + R^{\ell}_T(h^{\star}). \tag{24}$$

Step 1:

$$R^{\ell}_T(h) = R^{\ell}_S(h) + \big[R^{\ell}_T(h) - R^{\ell}_T(h^{\star})\big] - \big[R^{\ell}_S(h) - R^{\ell}_S(h^{\star})\big] + \big[R^{\ell}_T(h^{\star}) - R^{\ell}_S(h^{\star})\big]. \tag{25}$$

Because $R^{\ell}_T(h^{\star}) - R^{\ell}_S(h^{\star}) \le R^{\ell}_T(h^{\star}) + R^{\ell}_S(h^{\star}) = \lambda^{\ell}$, we have

$$R^{\ell}_T(h) \le R^{\ell}_S(h) + \Delta_T - \Delta_S + \lambda^{\ell}, \tag{26}$$

where $\Delta_{\mathcal{D}} := R^{\ell}_{\mathcal{D}}(h) - R^{\ell}_{\mathcal{D}}(h^{\star}) = \mathbb{E}_{\mathcal{D}}\big[\ell\big(h(x), y\big) - \ell\big(h^{\star}(x), y\big)\big]$.

Step 2: By the triangle inequality,

$$\big|\Delta_T - \Delta_S\big| \le \big|\Delta_T - \Delta_{\mathcal{D}_{mid}}\big| + \big|\Delta_{\mathcal{D}_{mid}} - \Delta_S\big|. \tag{27}$$

Observe that

$$\Delta_{\mathcal{D}} = \mathbb{E}_{\mathcal{D}}\big[\ell\big(h(x), h^{\star}(x)\big)\big], \tag{28}$$

because $\ell(a,b) - \ell(c,b) = \ell(a,c)$ when ℓ is a metric in its first argument and we choose $b = h^{\star}(x)$. (For cross-entropy or squared error this identity holds by symmetry; see [A5] for the full set of admissible ℓ.) Hence no labels y appear and we may apply the two lemmas:

$$\big|\Delta_S - \Delta_{\mathcal{D}_{mid}}\big| \le d^{\ell}_{\mathrm{enc}}\big(\mathcal{D}^Z_S, \mathcal{D}^Z_T\big) \quad \text{(Lemma 3)}, \qquad \big|\Delta_T - \Delta_{\mathcal{D}_{mid}}\big| \le d^{\ell}_{\mathrm{fus}}\big(\mathcal{D}^{(Z,Q)}_S, \mathcal{D}^{(Z,Q)}_T\big) \quad \text{(Lemma 4)}. \tag{29}$$

Step 3: Insert the last two inequalities into (26) and drop the absolute-value bars:

$$R^{\ell}_T(h) \le R^{\ell}_S(h) + d^{\ell}_{\mathrm{enc}}\big(\mathcal{D}^Z_S, \mathcal{D}^Z_T\big) + d^{\ell}_{\mathrm{fus}}\big(\mathcal{D}^{(Z,Q)}_S, \mathcal{D}^{(Z,Q)}_T\big) + \lambda^{\ell}, \tag{30}$$

which is exactly Eqn. (23). □ □

A3. Implementation Details for Negation-flip example

The dual-divergence bound (Thm. 1) holds for any $h = f \circ g$; we demonstrate it with a logistic classifier on source-only $[z, x^Q]$, which isolates divergence effects without confounding factors. Although SMMs have a richer hypothesis class and often achieve lower errors, the same bound applies unchanged (Sec. 2), and Sec. 4 shows the fusion term remains critical for SMMs. Divergences are estimated by fitting logistic domain discriminators (the simplest model to approximate the supremum in Eqs. (1) – (2), in the main paper) and converting its area under the ROC curve AUC to empirical total-variation (TV) distance as follows.

$$\mathrm{TV}(P,Q) = \sup_A \big|\Pr_P[A] - \Pr_Q[A]\big| = \big|2 \cdot \mathrm{AUC} - 1\big|. \tag{31}$$

A4. Datasets

We conduct extensive experiments on the VQA task. We design 12 new cross-domain UDA experiments using four diverse datasets: Hateful Memes (H), Visual Spatial Reasoning (V), IconQA (I), and ScienceQA (S). For Hateful Memes, VSR, and IconQA, we use the Cauldron split [A12] and further divide it into an 80:10:10 train:validation:test split. For ScienceQA, we use the Cauldron split for training and validation (90:10) and adopt the test split from [20]. The following are detailed descriptions of each of the dataset used in this works

ScienceQA Dataset: The ScienceQA dataset [20] contains 21,208 multimodal multiple-choice questions from elementary and high school curricula, with 48.7% including image contexts and 48.2% including text contexts, and 30.8% featuring both. Annotated with grounded lectures (83.9%) and detailed explanations (90.5%), it covers three subjects: natural science, language science, and social science, spanning 26 topics, 127 categories, and 379 skills.

Visual Spatial Reasoning (VSR) Dataset: The VSR dataset [18] includes 10,972 caption-image pairs with true or false labels, designed to evaluate vision-language models' spatial understanding. Each caption describes the spatial relation between two objects in an image, requiring the model to determine accuracy. The dataset features two splits: a random split for training, development, and test sets, and a zero-shot split to ensure no concept overlap.

Hateful Memes Dataset: Created by Facebook AI, the Hateful Memes dataset [14] detects hate speech in memes with over 10,000 examples combining images and text. Designed to challenge models in understanding context, it includes images licensed from Getty Images. The task is evaluated as a binary classification problem, with "benign confounders" included to highlight the need for multimodal approaches.

IconQA Dataset: The IconQA dataset [19] is a large-scale benchmark for abstract diagram understanding and visual language reasoning, comprising 107,439 question–answer pairs drawn from real-world math word problems with rich iconography. It is divided into three subtasks—multi-image-choice, multi-text-choice, and filling-in-the-blank—each requiring not only perception skills such as object recognition and text understanding but also diverse cognitive reasoning abilities including geometric, commonsense, and arithmetic reasoning. The dataset was constructed by sourcing icon-based problems from publicly available educational materials, then having trained annotators filter, shuffle, and balance examples to ensure diversity and fairness across all question types

We also report the licensing terms for the four multimodal VQA datasets. ScienceQA and IconQA are published under Creative Commons licenses that permit non-commercial reuse with share-alike, Hateful Memes uses a proprietary research-only license with Getty Images content, and Visual Spatial Reasoning is released under the permissive Apache 2.0 license.

A5. Implementation Details

We apply the proposed DuAA framework on two recent SMM models, Phi-3.5-Vision [1] and LLaVA-OneVision 0.5B [17].

(i) Phi-3.5-Vision model: The Phi 3.5-Vision (Phi3.5V) model, featuring 4.2 billion parameters, is a multimodal model capable of processing both single and multiple images alongside textual prompts to generate textual outputs. This model comprises two primary components: an image encoder (CLIP ViT-L/14) and a transformer decoder (phi-3.5-mini). Visual tokens extracted by the image encoder are combined with text tokens in an interleaved manner, without a fixed order. To manage high-resolution images and varying aspect ratios, a dynamic cropping strategy is employed, splitting the input image into a 2D array of blocks, with tokens from these blocks concatenated to represent the entire image. For multi-image inputs, tokens from each image are concatenated together.

(ii) LLaVa-OneVision: The LLaVA-OneVision (LLaVa-OV) model adopts the minimalist design principles of the LLaVA series, aiming to effectively leverage the pre-trained capabilities of both language and visual models while enabling robust scaling with data and model size. The architecture includes Qwen-2 as the language model, chosen for its strong language capabilities and various model sizes, and SigLIP as the vision encoder, which converts input images into visual features. These visual features are then projected into the word embedding space using a 2-layer MLP, resulting in a sequence of visual tokens. The model capitalizes on empirical insights that stronger language models enhance multimodal capabilities, with SigLIP offering superior performance among open vision encoders. The visual input can vary depending on the scenario, such as individual image crops in a single-image sequence or frames in a video sequence, ensuring that the visual signal remains grounded for all answers. In this work, we use 0.5B variant of this model.

We train the proposed DuAA framework using LoRA-based [A9] fine-tuning with lora rank 16 and alpha value as 8. We use the huggingface face library to implement the proposed DuAA, and use the default weight initializing and random seed values. We use AdamW optimizer [A15] with a mini-batch size of 64 and ϵ value of 1e-7. We use a linear learning rate scheduler and maximum gradient norm of 1.0. We also use mixed precision values for the model parameters. We train the Stage 1 of DuAA for 1 epoch and Stage 2 of DuAA for 2 epochs. We train all the experiments on NVIDIA A100 GPU with 40GB GPU RAM. To accomodate for memory efficiency we use gradient accumulation strategy with accumulation step same as the mini batch size to perform each model update. Detailed hyperparameter values for the DuAA model, tailored to each dataset, are provided in Table A1, while Table A2 outlines the corresponding search space for each hyperparameter.

A6. Additional Results on Image Classification Task

Dataset For the image classification task, we conduct experiments on the Office-Home dataset [A23], which is a widely recognized benchmark in domain adaptation. This

Table A1. Hyperparameter values of our DuAA model implementation on all datasets and SMMs.

Hyperparameter	VQA	Office-Home
LLaVA-OV		
τ	0.85	0.95
γ_1	0.25	0.25
γ_2	0.25	0.25
Learning rate	0.0002	0.0002
Weight Decay	0.01	0.001
Adam β_1	0.9	0.9
Adam β_2	0.95	0.95
Phi3.5V		
τ	0.85	0.95
γ_1	0.25	0.25
γ_2	0.25	0.25
Learning rate	0.0002	0.0002
Weight Decay	0.01	0.001
Adam β_1	0.9	0.9
Adam β_2	0.95	0.95

Table A2. Ranges of values for hyperparameter tuning.

Hyperparameter	Search Values
τ	[0.85,0.9,0.95]
γ_1	[0.25,0.5,1]
γ_2	[0.25,0.5,1]
Learning rate	[0.002,0.0002,0.00002]
Weight Decay	[0.01,0.05]
Adam β_1	[0.9,0.85]
Adam β_2	[0.95, 0.99]

dataset comprises 15,500 images spanning four distinct domains: Art, Clipart, Product, and Real-World, each containing 65 categories. The Art domain includes sketches, paintings, and ornamentation, while the Clipart domain features a collection of clipart images. The Product domain contains images of objects devoid of background context, and the Real-World domain consists of images captured by a standard camera. Each category averages approximately 70 images, with a maximum of 99 images per class. The structured nature of this dataset facilitates the evaluation of domain adaptation algorithms, particularly in object recognition and classification using deep learning techniques.

Baselines. For the image classification task, we compare the proposed DuAA framework with several recent benchmark methods. These include ViT-based approaches such as, SSRT [A22], PMTrans [A28], and SKD [A25], which utilize the Vision Transformer (ViT) architecture as their backbone. In addition, we consider CLIP-based methods, including PADCLIP [A10], VFR [A11], and HVCLIP [A24], which leverage the strong multimodal alignment capabilities of the CLIP framework. We also compare against the LLaVO method [A4], which incorporates a large language model (LLM) decoder.

Results. Table A3 shows that the proposed DuAA achieves competitive performance against state-of-the-art UDA methods on the Office-Home dataset. DuAA with LLaVA-OV attains the highest average accuracy of 93.1%, outperforming recent benchmark methods such as HVCLIP (92.0%) and VLLaVO (91.6%), the latter of which incorporates an LLM-based decoder. Similarly, DuAA with Phi3.5V achieves an average accuracy of 88.7%, on par with other leading methods. These improvements are meaningful given that the Office-Home dataset involves only visual domain shifts, making the task relatively easier compared to VQA, which requires handling both visual and linguistic variations. Despite the good performance of the Source-only baselines, 88.5% (Phi3.5V) and 92.5% (LLaVA-OV), the proposed DuAA consistently improves performance, further demonstrating its ability to reduce domain discrepancies even in settings with more limited domain complexity.

A7. More Analysis

τ Hyperparameter. Table A4 analyzes the impact of the confidence threshold (τ) used for pseudo-label selection in DuAA where each $\rightarrow$A column reports the average accuracy across experiments where A is target domain. As τ increases, only the most confident target samples are retained, reducing potential label noise but also limiting the number of pseudo-labeled samples. At $\tau = 0.95$, the strictest filtering leads to lower overall performance (54.1%), likely due to insufficient pseudo-labeled data for adaptation. Conversely, at $\tau = 0.85$, a balance between confidence and sample quantity is achieved, yielding the highest accuracy (57.9%). This suggests that overly restrictive thresholds may hinder adaptation by discarding useful target samples, whereas a moderately relaxed threshold provides more effective supervision.

Training and Inference Analysis. Table A5 reports the average time and GPU memory usage of DuAA with Phi3.5V and LLaVA-OV. As expected, DuAA (Phi3.5V) has a higher training time (70s vs. 32s) due to its larger model size (4B parameters vs. 0.5B). However, LLaVA-OV consumes more GPU memory during training (29GB vs. 13GB), likely due to additional memory-intensive preprocessing steps specific to its architecture. In contrast, during inference—where such preprocessing is absent—LLaVA-OV consistently requires less GPU memory (5.7GB vs. 11.8GB), while both models achieve comparable inference times (0.5s). This highlights a trade-off: Phi3.5V is more compute-intensive during training, while LLaVA-OV re-

Table A3. Accuracies (%) on the Office-Home dataset. *Only VLLaVO [A4] uses the , LLaMA-7B LLM. Avg. = Average Accuracy.

Method	Base Model	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg.
SSRT (2022)	ViT	75.2	89.0	91.1	85.1	88.3	89.9	85.0	74.2	91.3	85.7	78.6	91.8	85.4
PMTrans (2023)	ViT	81.2	91.6	92.4	88.9	91.6	93.0	88.5	80.0	93.4	89.5	82.4	94.5	88.9
SKD (2023)	ViT	79.6	93.7	92.7	89.5	93.7	92.9	89.1	81.1	92.6	90.2	81.6	94.2	89.2
PADCLIP (2023)	CLIP	76.4	90.6	90.8	86.7	92.3	92.0	86.0	74.5	91.5	86.9	79.1	93.1	86.7
VFR (2024)	CLIP	78.2	90.4	91.0	87.5	91.9	92.3	86.7	79.7	90.9	86.4	79.4	93.5	87.3
HVCLIP (2024)	CLIP	86.3	96.4	94.0	91.6	97.9	94.6	87.5	85.3	94.8	89.9	88.1	97.0	92.0
VLLaVO* (2024)	LLM	85.4	96.6	94.1	90.3	97.1	94.4	87.9	85.7	94.5	90.1	85.5	97.3	91.6
Source-only (2024)	SMM (Phi3.5V)	81.2	92.7	91.7	86.8	93.4	90.8	86.1	82.8	91.9	88.4	79.6	92.0	88.5
DuAA (Ours)	SMM (Phi3.5V)	81.9	91.7	91.1	87.1	93.3	90.7	87.5	83.1	91.8	88.9	80.1	92.3	88.7
Source-only (2024)	SMM (LLaVA-OV)	88.0	95.6	95.0	90.1	96.6	93.8	89.7	87.0	93.8	92.8	86.4	96.1	92.5
DuAA (Ours)	SMM (LLaVA-OV)	87.5	94.8	94.1	90.9	97.2	94.9	92.3	88.8	94.7	93.7	87.5	96.4	**93.1**

Table A4. Accuracy (%) of DuAA on VQA datasets, assessing the impact of different thresholds (τ).

τ	→H	→V	→I	→S	Avg.
0.95	43.8	57.7	54.3	60.7	54.1
0.90	45.5	59.8	55.4	61.7	55.6
0.85	53.1	60.8	55.6	62.1	**57.9**

quires more memory, with similar inference performance.

Table A5. Training and inference time (seconds) and GPU memory (GB) comparison of DuAA with Phi3.5V and LLaVA-OV.

Method	Avg. Time (s)	Avg. GPU RAM (GB)
Training		
Src-only (Phi3.5V)	60	11
DuAA (Phi3.5V)	70	14
Src-only (LLaVA-OV)	27	22
DuAA (LLaVA-OV)	32	29
Inference		
Src-only (Phi3.5V)	0.51	11.8
DuAA (Phi3.5V)	0.51	11.8
Src-only (LLaVA-OV)	0.50	5.7
DuAA (LLaVA-OV)	0.50	5.7

Experimental Consistency: Table A7 presents DuAA's mean accuracy and standard deviation across three independent runs for each of the 12 source→target VQA pairs. For nine of the twelve pairs, the standard deviation is under 1 percentage point, demonstrating that DuAA's improvements are highly reproducible rather than one-off gains. The larger variance observed in the S→H transfer reflects sensitivity in that domain pair, likely due to its extreme shift and resulting pseudo-label confidence fluctuations. Overall, the consistently low variability underscores DuAA's robustness in aligning both encoder and fusion divergences across diverse domain shifts.

Table A6. Accuracy (%) of DuAA on VQA datasets, assessing the impact of 'γ_1' and 'γ_2'.

γ_1	γ_2	Avg.
1	1	50.0
0.5	0.5	57.4
0.25	0.5	57.0
0.5	0.25	57.2
0.25	0.25	**57.9**

Table A7. DuAA (LLaVA-OV) accuracy averaged over three runs, with standard deviation, for each source→target VQA pair.

Pair	DuAA Mean ± Std (%)
I→H	41.9±2.1
V→H	70.1±0.9
S→H	54.1±6.1
H→V	63.0±1.9
I→V	58.9±0.9
S→V	61.1±0.1
H→I	54.9±0.7
V→I	55.7±0.1
S→I	55.6±0.3
H→S	62.3±0.6
V→S	61.0±0.2
I→S	61.0±1.1

Hyperparameters γ_1 and γ_2:Table A6 analyzes the impact of the weighting factors γ_1 and γ_2 in DuAA's Stage 2 objective (Eqn.(8)). γ_1 regulates the entropy loss, encouraging confident predictions, while γ_2 controls the domain alignment loss from the DDA modules. The results indicate that equal weighting ($\gamma_1 = 1$, $\gamma_2 = 1$) leads to suboptimal performance (50.0%), suggesting that excessive entropy regularization or domain alignment can be detrimental. The best performance (57.9%) is achieved with $\gamma_1 = 0.5$ and $\gamma_2 = 0.25$, striking a balance between confident target predictions and effective domain adaptation.

Effect of target sample selection Table A8 analyzes the impact of target sample selection in Stage 2 of DuAA. When using only confident target samples with pseudo-labels, DuAA achieves a higher average accuracy of 57.9%, compared to 53.8% when using all target samples. The

Table A8. Accuracy (%) of DuAA (LLaVA-OV) on VQA, analyzing the impact of target sample selection.

Sample Selection	Avg.
All Samples	53.8
Only Confident Samples	**57.9**

Table A9. Effect of unsupervised fine-tuning of target data vs DuAA.

Setting	Avg.
Zero-shot	48.8
Src-only	54.9
Complete Unsupervised FT	55.1
DuAA	**57.9**

performance drop when including all samples likely stems from the presence of low-confidence, unlabeled target examples, which contribute only to the DDA loss without task supervision, potentially introducing noise into the adaptation process. In contrast, restricting training to confident pseudo-labeled samples ensures that target supervision is applied to reliable data, leading to more effective adaptation.

Effect of complete unsupervised fine-tuning. Table A9 shows that including unsupervised target data along with labeled source data during stage 1 of training (Complete Unsupervised FT, 55.1) improves over Zero-shot (+6.3) and narrowly over Src-only (+0.2), but still trails DuAA (57.9) by a clear margin (+2.8). This gap supports our formulation: a brief source warm-up that stabilizes task predictions, followed by selective pseudo-labeling and dual-level adversarial alignment, yields more reliable target adaptation than unsupervised alignment alone. Notably, DuAA's +3.0 over Src-only and +9.1 over Zero-shot indicate that it leverages both domains effectively rather than over-specializing to the source.

Effect of multiple pseudo-label generation steps. Table A10 evaluates whether regenerating pseudo-labels multiple times improves adaptation. Despite the intuition that pseudo-label quality should rise as domain shift narrows, our attempt to regenerate once more (after Stage 2, epoch 1) yields a lower average accuracy (56.4) than the proposed two-stage DuAA pipeline (57.9). We attribute this drop to (i) confirmation bias introduced while the decision boundary is still moving under adversarial alignment, (ii) precision degradation of the high-confidence pool when feature distributions shift after adapter updates, and (iii) the destabilizing effect of repeatedly rewriting labels, which dilutes the stable source-anchored supervision established in Stage 1.

Table A10. Effect of multiple pseudo-label generation steps vs DuAA.

Setting	Avg.
Multiple pseudo-label steps	56.4
DuAA	**57.9**

Under identical data and compute budgets, these results support our design choice of a single, selective pseudo-label pass following a short source warm-up.

Adaptation to Open-Ended VQA: We show the qualitative examples of different datasets used in this work in Fig. A1. We can observe from Fig. A1, that open-ended questions are much harder to solve and require more reasoning capabilities than other closed-ended datasets. The two open-ended datasets used in this work are Infographic VQA [23] and DocVQA [22] datasets.

Infogrpahic VQA: This is a large-scale benchmark for evaluating visual question answering on infographics—documents that combine textual, graphical, and data-visualization elements. The dataset contains over 5,000 infographic images paired with 30,000 natural-language questions and answers, curated to require joint reasoning over layout, text, graphics, and basic arithmetic operations. Questions span multiple categories including extractive (text lookup), multiple-choice, and simple reasoning, making InfographicVQA a comprehensive testbed for multimodal understanding that goes beyond conventional scene-based VQA. The InfographicVQA dataset is released under the Apache License 2.0, a permissive open-source license that allows unrestricted use, modification, and redistribution provided that the license and copyright notices are preserved. All infographic images and annotations are covered by this license, facilitating both academic research and commercial applications without additional restrictions.

DocVQA: The DocVQA is a document-centric VQA dataset consisting of 50,000 questions defined on over 12,000 document page images sourced from the UCSF Industry Documents Library. Each example includes both the document image and its OCR transcription, with question types ranging from table understanding to form-based queries, designed to benchmark models on structured and unstructured text comprehension in natural documents. The dataset provides train, validation, and test splits matching the CVPR 2020 Challenge Task 1, facilitating fair comparison across methods. The DocVQA dataset is distributed under the MIT License, a permissive license that allows commercial and non-commercial use, modification, and redistribution as long as copyright and license notices are preserved.

In this analysis we adapt from the closed-ended VQA

Hateful Memes (H)

Input Image:

Input Question:

Does this meme support discrimination?\nAnswer yes or no.

Output Answer

No

ScienceQA (S)

Input Image:

Input Question:

Lecture: An object has different properties. A property of an object can tell you how it looks, feels, tastes, or smells. Properties can also tell you how an object will behave when something happens to it.\nQuestion: Which property matches this object?\nHint: Select the better answer.\nChoices:\nA. flexible\nB. sticky\nAnswer with the letter.

Output Answer

Answer: C

Visual Spatial Reasoning (V)

Input Image:

Input Question:

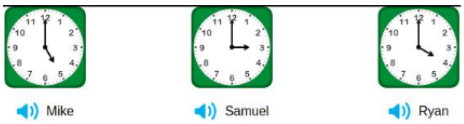
Evaluate: Does the caption \"The horse is next to the bus.\" match the image?\nAnswer yes or no.

Output Answer

No

IconQA (I)

Input Image:



Input Question:

Question: The clocks show when some friends did homework Wednesday afternoon. Who did homework last?\nChoices:\nA. Samuel\nB. Ryan\nC. Mike\nAnswer with the letter.

Output Answer

Answer: C

DocVQA (D)(Open-Ended)

Input Image:

Input Question:

What is the subject of the document/letter? \nGive a very brief answer.

Output Answer

Flavor Development Monthly Summary for April, 1990.

Infographic VQA (In)(Open-Ended)

Input Image:

Input Question:

What is the truth of reading under low light? \nBe succinct.

Output Answer

It strains your eyes, making them tired and achy.

Figure A1. Example input images, input questions, and ground truth output answer for different VQA datasets.

dataset V to two open-ended VQA datasets: DocVQA (D) [A17] and InfographicVQA (In) [A18]. Closed-ended VQA restricts answers to binary (yes/no) or multiple-choice options, whereas open-ended VQA requires generating free-form textual answers. Table A11 reports the ANLS score, the standard metric for evaluating answer overlap in these open-ended benchmarks, showing that DuAA improves Phi3.5V performance from 55.9 to 58.8 (+2.9%, 15.9% gap closed) on V→In and from 79.1 to 79.5 (+0.4%, 11.4% gap closed) on V→D. Adapting to open-ended tasks is more challenging due to the need for coherent language generation, extraction of fine-grained information, and integration of external knowledge beyond fixed answer vocabularies.

Table A11. Source-only (Src), DuAA (Ours), and oracle (OR) ANLS score for adaptation from the close-ended VQA domain (V) to open-ended domains In and D for Phi3.5V SMM.

Task	Src	DuAA	Δ	Gap(%)	OR
V→In	55.9	58.8	+2.9	15.9	74.1
V→D	79.1	79.5	+0.4	11.4	82.6
Average	67.5	69.2	**1.7**	**13.7**	78.4

A8. More Related Works

UDA for Single-Encoder and Dual-Encoder models: In the context of encoder-only models, two primary approaches have been widely adopted: adversarial learning, which extracts domain-invariant features through min-max optimization [A1, A3, A7, A16, A20, A27], and self-training, which refines pseudo-labels to improve target performance [A13, A14, A21, A22, A26]. With the introduction of vision-language models (dual-encoder) such as CLIP [A19], UDA has extended into multimodal settings with two popular strategies: prompt learning and pseudo-labeling. Prompt learning methods adapt domain-specific information via optimized textual or visual prompts [A6, A8]. Pseudo-labeling methods enhance adaptation by generating target pseudo-labels or textual descriptions for domain transfer [A4, A24]. In contrast to these CLIP-based approaches, the proposed DuAA is built on SMMs, which introduce distinct domain shift challenges due to their architecture characteristics. Furthermore, all CLIP-based methods are evaluated only on image classification task, while DuAA extends to the more complex VQA task.

A9. Limitations and Future Directions

Our experiments instantiate DuAA with LoRA-based adapters, which are a widely adopted and efficient choice for adapting SMMs. While our formulation does not rely on LoRA-specific assumptions, a systematic empirical study of how fusion divergence behaves under alternative fine-tuning strategies (e.g., prefix-tuning, full-parameter fine-tuning) can be explored in future works. In terms of compute, DuAA increases training time without affecting inference latency. This one-time training overhead is acceptable in scenarios where a single adapted model serves many downstream queries or operates in high-stakes, domain-shifted settings. However, it may be less attractive for workflows that require frequent re-training under tight compute budgets, which motivates future work on more lightweight or partially applied variants of DuAA.

Additional References

[A1] Divya Jyoti Bajpai and Manjesh Kumar Hanawal. DAdEE: Unsupervised domain adaptation in early exit PLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6389–6400, Miami, Florida, USA, 2024. Association for Computational Linguistics. 9

[A2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010. 2

[A3] Lin Chen, Huaian Chen, Zhixiang Wei, Xin Jin, Xiao Tan, Yi Jin, and Enhong Chen. 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*, pages 7181–7190, 2022. 9

[A4] Shuhao Chen, Yulong Zhang, Weisen Jiang, Jiangang Lu, and Yu Zhang. Vllavo: Mitigating visual gap through llms. *arXiv preprint arXiv:2401.03253*, 2024. 5, 6, 9

[A5] Corinna Cortes, Mehryar Mohri, et al. Domain adaptation with sample bias correction. In *ACM COLT*, 2014. 2, 3

[A6] Zhekai Du, Xinyao Li, Fengling Li, Ke Lu, Lei Zhu, and Jingjing Li. Domain-agnostic mutual prompting for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23375–23384, 2024. 9

[A7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016. 9

[A8] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 9

[A9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 4

[A10] Zhengfeng Lai, Noranart Vesdapunt, Ning Zhou, Jun Wu, Cong Phuoc Huynh, Xuelu Li, Kah Kuen Fu, and Chen-Nee Chuah. Padclip: Pseudo-labeling with adaptive debiasing in clip for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16155–16165, 2023. 5

[A11] Zhengfeng Lai, Haoping Bai, Haotian Zhang, Xianzhi Du, Jiulong Shan, Yinfei Yang, Chen-Nee Chuah, and Meng Cao. 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*, pages 2691–2701, 2024. 5

[A12] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024. 3

[A13] Jianling Li, Meishan Zhang, Peiming Guo, Min Zhang, and Yue Zhang. Llm-enhanced self-training for cross-domain

constituency parsing. *arXiv preprint arXiv:2311.02660*, 2023. 9

[A14] Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle self-training for domain adaptation. *Advances in Neural Information Processing Systems*, 34:22968–22981, 2021. 9

[A15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[A16] Bhavitvya Malik, Abhinav Ramesh Kashyap, Min-Yen Kan, and Soujanya Poria. Udapter-efficient domain adaptation using adapters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2249–2263, 2023. 9

[A17] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 8

[A18] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 8

[A19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 9

[A20] Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, Arihant Jain, and Venkatesh Babu Radhakrishnan. A closer look at smoothness in domain adversarial training. In *International conference on machine learning*, pages 18378–18399. PMLR, 2022. 9

[A21] Jon Saad-Falcon, Omar Khattab, Keshav Santhanam, Radu Florian, Martin Franz, Salim Roukos, Avirup Sil, Md Arafat Sultan, and Christopher Potts. Udapdr: unsupervised domain adaptation via llm prompting and distillation of rerankers. *arXiv preprint arXiv:2303.00807*, 2023. 9

[A22] Tao Sun, Cheng Lu, Tianshuo Zhang, and Haibin Ling. Safe self-refinement for transformer-based domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7191–7200, 2022. 5, 9

[A23] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 4

[A24] Noranart Vesdapunt, Kah Kuen Fu, Yue Wu, Xu Zhang, and Pradeep Natarajan. Hvclip: High-dimensional vector in clip for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 36–54. Springer, 2024. 5, 9

[A25] Thomas Westfechtel, Dexuan Zhang, and Tatsuya Harada. Combining inherent knowledge of vision-language models with unsupervised domain adaptation through self-knowledge distillation. *CoRR*, 2023. 5

[A26] Tongkun Xu, Weihua Chen, Pichao WANG, Fan Wang, Hao Li, and Rong Jin. CDTrans: Cross-domain transformer for unsupervised domain adaptation. In *International Conference on Learning Representations*, 2022. 9

[A27] Yufeng Zhang, Jianxing Yu, Yanghui Rao, Libin Zheng, Qinliang Su, Huaijie Zhu, and Jian Yin. Domain adaptation for subjective induction questions answering on products by adversarial disentangled learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9074–9089, Bangkok, Thailand, 2024. Association for Computational Linguistics. 9

[A28] Jinjing Zhu, Haotian Bai, and Lin Wang. Patch-mix transformer for unsupervised domain adaptation: A game perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3561–3571, 2023. 5