DSSG: Dual-Stream Semantic Guidance for Source-Fully-Free Adaptation of Vision-Language Models

Weiwei Xiang[1,2], **Shun Peng**[1], **Guangyi Xiao**[1,*], **Hao Chen**[1] and **Lei Yang**[1,*]
[1]Hunan University
[2]Huaihua University
{mr_menand, pengshun, gyxiao, chenhao, jt_yl}@hnu.edu.cn

Abstract

Source-Fully-Free Domain Adaptation (SFF-DA) has emerged as a strategic paradigm to adapt Vision-Language Models (VLMs) without any access to source data or task-specific source models. However, we identify a critical *Dual Semantic Drift* that hinders this process: static drift arising from the rigidity of frozen class anchors, and dynamic drift stemming from the divergence of generated captions, causing severe semantic misalignment that intensifies the stability-plasticity dilemma.

To address this, we propose DSSG (Dual-Stream Semantic Guidance), an end-to-end framework that reconciles fine-grained plasticity with global stability. Our core contribution is the Dual Semantic Guidance (DSG) module, which integrates a caption stream for domain-specific knowledge with a class-anchor stream to anchor global categorical consistency. Furthermore, a Dynamic Cross-Modal Knowledge Distillation (CMKD) module is introduced to leverage the evolving teacher distribution for calibrating teacher-student consistency. Extensive experiments demonstrate that DSSG significantly outperforms current state-of-the-art methods across multiple benchmarks, providing a robust solution for the unsupervised transfer of foundation models. The code is available on https://github.com/mrmenand/DSSG.

1 Introduction

Traditional unimodal source-free domain adaptation (SFDA) follows a two-stage pipeline of source-domain pre-training and target-domain adaptation. While effective, this paradigm heavily relies on task-specific source priors and is restricted by practical deployment bottlenecks, such as data privacy concerns and model black-box constraints (e.g., API-only access). Recently, Vision-Language Models (VLMs) represented by CLIP, pre-trained on large-scale image-text pairs, exhibit strong zero-shot generalization, serving as generalized source models rich in transferable knowledge. To eliminate dependency on task-specific source training, we focus on Source-Fully-Free Domain Adaptation (SFF-DA)—also known as Unsupervised Fine-Tuning (UFT) [Li *et al.*, 2025]. We directly leverage VLMs for adaptation on unlabeled target domains without accessing any source data or models, aiming to transfer their inherent universal knowledge to domain-specific contexts under distribution shifts.

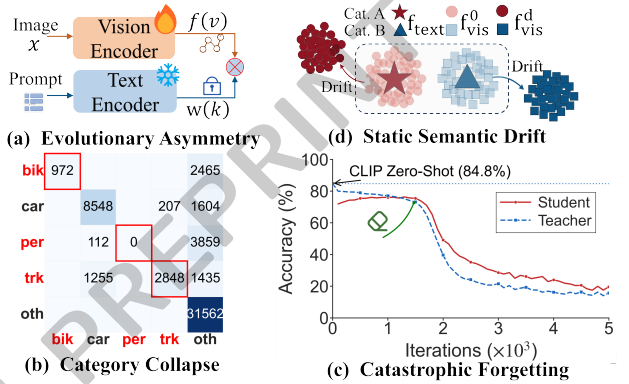


Figure 1: Challenges in SFF-DA. (a) Evolutionary Asymmetry causes feature misalignment, leading to (b) Category Collapse and (c) Catastrophic Forgetting. We term this phenomenon (d) Static Semantic Drift, where $0/d$ denote initial/drifted visual features.

While parameter-efficient methods (e.g., Prompt Learning [Zhang *et al.*, 2022b], Adapters [Zhang *et al.*, 2022a]) preserve VLM priors by freezing backbones, their limited learnable parameters hinder the acquisition of intricate domain-specific knowledge. In contrast, directly fine-tuning foundation models in SFF-DA encounters a twofold challenge: *(i) error accumulation and catastrophic forgetting* (Fig. 1(b,c)), where the lack of source guidance accumulates pseudo-label errors, exacerbating confirmation bias and category collapse, thereby corrupting large-scale pre-trained knowledge; and *(ii) evolutionary asymmetry and static semantic drift* (Fig. 1(a, d)), where the vision encoder evolves toward a domain-specific manifold while template-derived embeddings w_k remain fixed in the frozen text space, inducing a spatiotemporal misalignment that undermines cross-modal consistency, termed static semantic drift.

Existing CLIP-based SFDA methods generally follow two paradigms. Firstly, CLIP-assisted approaches (e.g., DIFO,

*Corresponding authors.

ProDe [Tang *et al.*, 2024b; Tang *et al.*, 2025]) employ CLIP as an external teacher to calibrate task-specific source models. However, the reliance on pre-trained source weights limits their applicability in SFF-DA. Secondly, the unsupervised fine-tuning paradigm adapts VLMs by optimizing prompt variables [Long *et al.*, 2024], the vision encoder [Mirza *et al.*, 2023], or joint tuning [Hu *et al.*, 2024]. While updating encoders improves adaptability, it is prone to damaging universal priors. This leads to a fundamental stability-plasticity dilemma: prompt-based tuning lacks the plasticity for target-specific distributions, while encoder-based tuning damages universal pretrained knowledge.

Motivation: Drawing inspiration from ImCapDA [Xiang *et al.*, 2025], we consider generating descriptive captions as semantic proxies to jointly fine-tune encoders, which has proven effective in enforcing cross-modal alignment. However, in the rigorous SFF-DA scenario lacking source label supervision, this introduces a grave risk of dynamic semantic drift. Generated captions exhibit an inherent bias towards non-discriminative instance-level details or background noise, inducing an unconstrained semantic flow that causes visual features to diverge from their intrinsic class manifolds. This semantic divergence, coupled with the rigidity of fixed class embeddings w_k, constitutes a *Dual Semantic Drift* dilemma. It hinders the model from reconciling the capture of domain-specific with the preservation of global categorical structures.

To address this, we present the Dual-Stream Semantic Guidance (DSSG) framework. First, the DSG module integrates a caption stream and a class-anchor stream to construct a complementary semantic space—utilizing captions to provide instance-specific context for mitigating forgetting, while employing anchors to preserve intrinsic category semantics. Second, the Dynamic CMKD module calibrates the evolving teacher distribution via self-distillation to enforce teacher–student consistency. This coupled and complementary design ensures robust cross-modal alignment and global categorical consistency via dual semantic guidance.

Our main contributions are summarized as follows:

1. We identify Dual Semantic Drift—the conflict between static embedding rigidity and caption divergence—as a core stability-plasticity dilemma in SFF-DA.
2. We introduce the DSSG framework, which integrates parallel semantic guidance with a dynamic distillation module to reconcile fine-grained adaptation and global consistency.
3. Extensive experiments on mainstream benchmarks demonstrate that our method significantly outperforms existing approaches, establishing a new state-of-the-art.

2 Related Work

Unimodal Source-Free Domain Adaptation. Unimodal SFDA approaches typically rely solely on visual features, leveraging source prior models for self-training on unlabeled targets. SHOT [Liang *et al.*, 2020] and AaD [Yang *et al.*, 2022] employ mutual information and clustering for feature alignment. For robustness, TPDS [Tang *et al.*, 2024a] and DIPE [Wang *et al.*, 2022] investigate distribution flow search and domain-invariant parameter mining, respectively. Recent strategies explore Transformer adaptability [Sanyal *et al.*, 2023] and advanced augmentation [Hwang *et al.*, 2024], while TiGM [Zhu *et al.*, 2025] uses multi-dimensional metrics for pseudo-label selection. However, this unimodal limitation leads to a lack of explicit semantic guidance, limiting generalization across complex domain shifts.

Vision-Language Model Assisted SFDA. Recent works leverage CLIP as an external teacher to guide knowledge transfer via two primary paradigms: *(1) Prediction Calibration:* Co-learn++ [Zhang *et al.*, 2025] and DPC [Zhan *et al.*, 2024] ensemble zero-shot priors within dual-branch frameworks, while ProDe [Tang *et al.*, 2025] utilizes CLIP as a noisy agent for probabilistic denoising. *(2) Structural Alignment:* DIFO [Tang *et al.*, 2024b] maximizes mutual information for alignment, while MMGA [Chen *et al.*, 2025] and DTKI [Zhan *et al.*, 2026] enforce topological consistency to anchor features to semantic prototypes. Yet, requiring source supervision and complex auxiliary deployment restricts their SFF-DA feasibility and yields suboptimal performance compared to direct VLM adaptation

Multimodal Source-Fully-Free Domain Adaptation (SFF-DA). Multimodal SFF-DA directly fine-tunes VLMs on unsupervised target domains. Existing works follow three paradigms: *(1) Prompt Tuning*, which optimizes learnable prompts with frozen backbones (e.g., UPL [Huang *et al.*, 2022], TFUP [Long *et al.*, 2024]); *(2) Visual Tuning*, which fine-tunes only the visual encoder to align with fixed textual features (e.g., LaFTer [Mirza *et al.*, 2023], DPA [Ali *et al.*, 2025]); and *(3) Joint Fine-tuning*, which optimizes both visual and text encoders for cross-modal feature alignment (e.g., ReCLIP [Hu *et al.*, 2024], POUF [Tanwisuth *et al.*, 2023], ImCapDA [Xiang *et al.*, 2025]). However, these methods face the Dual Semantic Drift dilemma: visual-only tuning leads to static semantic drift due to fixed class embeddings, whereas caption-based joint tuning induces dynamic semantic drift under unconstrained semantic flow.

3 DSSG Framework

We propose the DSSG framework, which integrates parallel semantic guidance with a dynamic distillation module to address the *Dual Semantic Drift* dilemma in SFF-DA (Fig. 2). The framework consists of VLM-based caption generation (Sec. 3.2), the Dual-Stream Guidance (DSG) module (Sec. 3.3), and Dynamic CMKD (Sec. 3.4).

3.1 Preliminaries and Problem Definition

SFF-DA Setting: Given an unlabeled target dataset $\mathcal{D}_t = \{x_i\}_{i=1}^{N_t}$, S SFF-DA adapts a pre-trained vision-language model Θ_{clip} via unsupervised fine-tuning, without access to either source data $\mathcal{D}_s$ or source model Θ_s.

Revisiting CLIP: CLIP comprises a visual encoder E_V and a text encoder E_T, jointly trained to align images and text into a shared feature space via large-scale contrastive pre-training. During inference, given a category set $\mathcal{C} = \{c_k\}_{k=1}^{K}$, fixed class embeddings $w_k = E_T(t_k)$ are computed from

prompt templates t_k (e.g., "a photo of a c_k"). For an image x_i with feature $f_i = E_V(x_i)$, the prediction probability is defined as:

$$p(y_i = k|x_i) = \frac{\exp(\text{sim}(f_i, w_k)/\tau)}{\sum_{j=1}^{K} \exp(\text{sim}(f_i, w_j)/\tau)}, \quad (1)$$

where $\text{sim}(\cdot,\cdot)$ is the cosine similarity and τ is the temperature parameter.

3.2 Data Priors: Generative Instance-level Captions

In SFF-DA, fixed class embeddings $\boldsymbol{w}_k$ derived from templates t_k remain anchored in the static semantic space, failing to track dynamic visual manifolds and inducing evolutionary asymmetry. Furthermore, fine-tuning the vision and text encoders $\{\theta_V, \theta_T\}$ without source guidance is prone to category collapse and catastrophic forgetting. To mitigate these issues, we leverage the generative captioning strategy from ImCapDA [Xiang *et al.*, 2025], proven to facilitate domain-specific knowledge learning and alleviating such forgetting. Specifically, a frozen VLM generates a caption stream for each unlabeled target sample $x_i \in \mathcal{D}_t$ to construct instance-level priors:

$$T_{cap,i} = \text{VLM}(x_i), \quad (2)$$

where advanced VLMs (e.g., BLIP-3 [Xue *et al.*, 2024]) capture rich instance-level context, such as background textures and object attributes. Crucially, contrastive alignment of target images with the generated caption stream is essential to drive parameter updates in $E_T(\cdot;\theta_T)$. This process enables coupled evolution of the semantic space with the visual manifold to mitigate static drift, acting as a semantic soft constraint to counteract confirmation bias. Consequently, $T_{cap,i}$ serves as the critical input for our DSSG framework.

3.3 Dual-Stream Guidance

While T_{cap} introduces domain context, relying solely on its unconstrained semantics inherently induces dynamic divergence, thereby trapping the model in a dual semantic drift dilemma. Diverging from prior paradigms constrained to either fixed prompt templates [Zhou and Zhou, 2024] or caption-only guidance [Xiang *et al.*, 2025], we introduce a Dual-Stream Guidance (DSG) mechanism that integrates a caption stream and a class-anchor stream in parallel. Specifically, for a target image x_i, its visual representation $f_i = E_V(x_i;\theta_V)$ is aligned with two complementary textual embeddings encoded by a shared text encoder θ_T:

$$g_{anc} = E_T(T_{anc};\theta_T), \quad g_{cap,i} = E_T(T_{cap,i};\theta_T), \quad (3)$$

where g_{anc} is derived from class-generic prompt templates $\{t_k\}_{k=1}^K$ and preserves intrinsic category semantics, thereby providing global semantic rigidity against dynamic semantic drift. Conversely, $g_{cap,i}$ captures instance-specific knowledge from VLM-generated descriptions and introduces local semantic plasticity to alleviate static semantic drift. This dual-stream design maintains cross-modal alignment within a stable semantic manifold.

To jointly fine-tune θ_V, θ_T, we adopt a self-distillation paradigm following prior works [Zhou and Zhou, 2024; Xiang *et al.*, 2025], which employs a similarity-based teacher predictor and a linear student classifier $h(\cdot)$. Within this framework, we formulate objectives that exploit the dynamic adaptability of the caption stream while enforcing the semantic rigidity of the anchor stream:

Caption-guided Contrastive Alignment. To promote fine-grained cross-modal feature alignment, we employ a symmetric InfoNCE loss that maximizes the similarity of matched image–caption pairs $(f_i, g_{cap,i})$ while minimizing similarities to mismatched pairs. The objective is formulated with an image-to-text loss $\mathcal{L}_{i2t}$ and a text-to-image loss $\mathcal{L}_{t2i}$:

$$\mathcal{L}_{i2t} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\text{sim}(f_i, g_{cap,i})/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(f_i, g_{cap,j})/\tau)}, \quad (4)$$

$$\mathcal{L}_{t2i} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\text{sim}(g_{cap,i}, f_i)/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(g_{cap,i}, f_j)/\tau)}. \quad (5)$$

The overall objective is $\mathcal{L}_{con} = 0.5(\mathcal{L}_{i2t} + \mathcal{L}_{t2i})$.

Dynamic Anchor-based Self-training. Since gradients from T_{cap} propagate to update θ_T, retaining fixed class embeddings w_k induces misalignment with the evolving space (static drift), while instance-level captions cause visual features to diverge from their intrinsic class manifolds (dynamic drift). To address this dual-semantic drift, we recalibrate class anchors $\{g_{\text{anc}}^k\}_{k=1}^K$ under the current θ_T, formulating the dynamic teacher distribution:

$$\begin{aligned} P^{tea} &= \text{Softmax}(\text{sim}(f_i, w_k)/\tau) \quad \text{(Static)} \\ P^{tea*} &= \text{Softmax}(\text{sim}(f_i, g_{anc})/\tau) \quad \text{(Dynamic)}, \end{aligned} \quad (6)$$

where P^{tea*} provides calibrated guidance compared to the static baseline P^{tea}. Consequently, we derive pseudo-labels from the dynamic teacher: $\hat{y}_i = \arg\max_k P^{tea*}(y = k|x_i)$. This dynamic recalibration is performed during training with marginal computational overhead and introduces no additional cost at inference stage (see Appendix).

Following ImCapDA [Xiang *et al.*, 2025], we compute the student prediction P^{stu} by feeding the concatenated visual and caption features into a linear classifier $h(\cdot)$

$$P^{stu}(y|x_i) = \text{Softmax}\big(h([f_i; g_{cap,i}])\big). \quad (7)$$

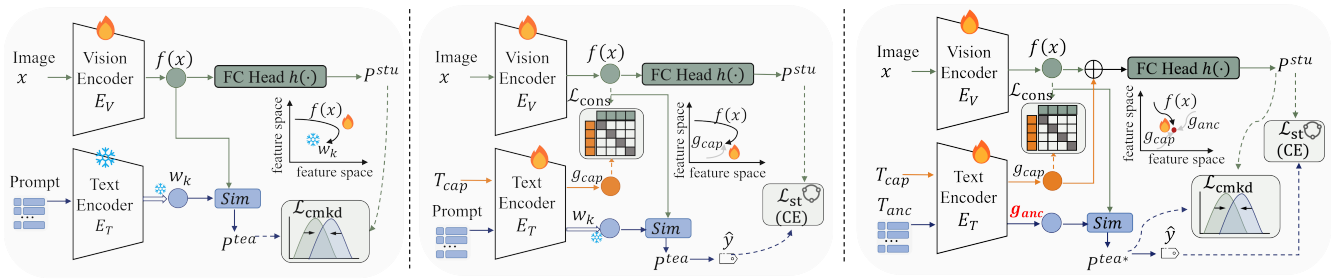
To guide and accelerate student training, we filter high-confidence pseudo-labels from the dynamic teacher via a threshold γ as supervision signals, minimizing:

$$\mathcal{L}_{st} = -\frac{1}{B}\sum_{i=1}^{B} \mathbb{I}(P^{tea*}(\hat{y}_i|x_i) \geq \gamma) \log P^{stu}(\hat{y}_i|x_i), \quad (8)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function.

3.4 Dynamic Cross-Modal Knowledge Distillation

To harness the robust generalization capability of vision-language pre-training models for UDA, CMKD [Zhou and Zhou, 2024] leverages the static semantic guidance from frozen text encoders to transfer generic knowledge to the unlabeled target domain. However, a critical limitation remains: fixed class embeddings w_k fail to track the continuously evolving visual representations during training, inevitably leading to severe cross-modal misalignment. To



(a) Conventional CMKD (b) ImCapDA (c) Ours DSSG

Figure 2: Illustration of (a) Conventional CMKD, (b) ImCapDA, and (c) Our DSSG framework.

bridge this gap, we propose Dynamic CMKD, which utilizes the refined dynamic teacher (P^{tea*}) and student (P^{stu}) distributions derived in Sec. 3.3 to enforce distribution consistency via a dynamic distillation loss.

Optimization Objective. To quantify the prediction consensus between the student and the dynamic teacher, a consistency coefficient $coe = \text{sg}(\exp(-\text{KL}(P^{stu} \parallel P^{tea*})))$ is computed, where $\text{sg}(\cdot)$ denotes the stop-gradient operation. Furthermore, to mitigate the over-confidence risk inherent in standard entropy minimization, the joint objective is constructed based on Gini Impurity:

$$\mathcal{L}^*_{cmkd} = \alpha \Big[\underbrace{coe \cdot \mathcal{G}(P^{stu})}_{\mathcal{L}_{task}} + \underbrace{(1-coe) \cdot \mathcal{G}(P^{mix})}_{\mathcal{L}_{distill}} \Big] + \underbrace{\beta \cdot \mathcal{G}(P^{tea*})}_{\mathcal{L}_{reg}}, \tag{9}$$

where $\mathcal{G}(P) = 1 - \|P\|_2^2$ denotes the Gini Impurity, and $P^{mix} = 0.5(P^{stu} + P^{tea*})$ represents the averaged mixture distribution. Specifically, when consensus is high ($coe \to 1$), $\mathcal{L}_{task}$ drives the student to maximize prediction certainty. Conversely, when predictions diverge ($coe \to 0$), $\mathcal{L}_{distill}$ steers the student towards P^{mix} reduces bias through dynamic teacher priors. Furthermore, $\mathcal{L}_{reg}$ encourages the sharpness of the teacher's distribution to prevent over-smoothing. We adhere to the original hyperparameter settings of $\alpha = 0.25$ and $\beta = 0.025$, and employ λ_3 to balance the overall objective.

3.5 Overall Optimization Objective

To achieve robust SFF-DA, we propose an end-to-end DSSG framework that integrates the generative caption stream T_{cap} and class anchor stream T_{anc} to jointly fine-tune the visual encoder θ_V and textual encoder θ_T via the total objective:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{con} + \lambda_2 \mathcal{L}_{st} + \lambda_3 \mathcal{L}^*_{cmkd}, \tag{10}$$

where $\mathcal{L}_{con}$ ensures caption-guided alignment, while $\mathcal{L}_{st}$ and $\mathcal{L}^*_{cmkd}$ leverage dynamic self-training and consensus-aware distillation to mitigate dual semantic drift in a unified objective. $\lambda_{\{1,2,3\}}$ are trade-off hyperparameters.

Total Gradient. The model parameters are updated via the aggregated gradient dynamics: $\nabla_{\theta_V,\theta_T}\mathcal{L}_{total} = \lambda_1 \nabla_{\theta_V,\theta_T}\mathcal{L}_{con} + \lambda_2 \nabla_{\theta_V}\mathcal{L}_{st} + \lambda_3 \nabla_{\theta_V,\theta_T}\mathcal{L}^*_{cmkd}$.

4 Experiments

4.1 Experiments Setting

Datasets. Experiments are evaluated on four benchmarks: (i) *Office-31*, 4,110 images in 31 classes across three domains: Amazon (A), Webcam (W), and DSLR (D); (ii) *Office-Home*, 15,588 images in 65 classes across four domains: Art (A), Clipart (C), Product (P), and Real-World (R); (iii) *Mini-DomainNet*, a DomainNet subset with 140,006 images in 126 classes across four domains: Clipart (C), Painting (P), Real (R), and Sketch (S); (iv) *VisDA-2017*, a synthetic-to-real benchmark with 55,388 target images across 12 classes.

Baselines. We evaluate our method against representative baselines from three categories: (i) *Unimodal Source-Free DA* (U), including SHOT [Liang *et al.*, 2020], AaD [Yang *et al.*, 2022], DIPE [Wang *et al.*, 2022], DSiT [Sanyal *et al.*, 2023], SF(DA)^2 [Hwang *et al.*, 2024], and TPDS [Tang *et al.*, 2024a]; (ii) *Multimodal Source-Free DA* (U+M), employing an external multimodal teacher to guide the source-supervised unimodal student, represented by DIFO-C [Tang *et al.*, 2024c], ProDe [Tang *et al.*, 2025], MMGA [Chen *et al.*, 2025], Co-Learn [Zhang *et al.*, 2025], DTKI [Zhan *et al.*, 2026], and DPC [Zhan *et al.*, 2024]; and (iii) *Multimodal Source-Fully-Free DA* (M), such as UPL [Huang *et al.*, 2022], POUF [Tanwisuth *et al.*, 2023], TFUP-T [Long *et al.*, 2024], ReCLIP [Hu *et al.*, 2024], and ImCapFusionSFDA [Xiang *et al.*, 2025].

Implementation Details. We adopt CLIP pre-trained on WIT-400M with ViT-B/16 and ResNet50 backbones (RN101 for VisDA-2017). We use SGD with momentum 0.9, weight decay 5e-4, and batch size 32. The learning rate is 3e-6 for the encoders and $1000\times$ higher for the student head. The confidence threshold γ is empirically set to 0.9 across all benchmarks. All experiments are implemented in PyTorch on an NVIDIA RTX 4090D GPU. See Appendix for more details.

4.2 Experimental Results

We compare DSSG with representative baselines from unimodal (U), external-teacher-guided multimodal (U+M), and directly fine-tuned multimodal (M) settings.

Results on Standard Benchmarks: Office-Home and Office-31. As shown in Table 1 for Office-Home, DSSG achieves **89.3%** with ResNet-50, exceeding the 84.5% of

Methods	Veneus	SFF	Mark	C	P	R	A	P	R	A	C	R	A	C	P	Avg.
				A			C			P			R			
ResNet-50:																
SHOT [Liang *et al.*, 2020]	*ICML*	✘	U	68.0	67.4	73.3	57.1	54.9	58.8	78.1	78.2	84.3	81.5	78.1	82.2	71.8
SF(DA)^2 [Hwang *et al.*, 2024]	*ICLR*	✘	U	69.5	66.5	73.3	57.8	57.2	60.2	80.2	79.2	83.8	81.5	79.4	82.1	72.6
TIGM [Zhu *et al.*, 2025]	*CVPR*	✘	U	69.3	68.1	74.6	61.2	58.8	62.4	80.9	81.2	85.7	82.7	81.4	83.4	74.1
DIFO-C-RN$^{\dagger}{}_{R/50}$ [Tang *et al.*, 2024c]	*CVPR*	✘	U+M	79.5	78.3	80.0	62.6	63.4	63.3	87.5	87.9	87.7	87.1	87.4	88.1	79.4
DIFO-C-B32$^{\dagger}{}_{B/32}$ [Tang *et al.*, 2024c]	*CVPR*	✘	U+M	82.5	80.9	83.4	70.6	70.1	70.5	90.6	90.6	91.2	88.8	88.8	88.9	83.1
ProDe-R$^{\dagger}{}_{R/50}$ [Tang *et al.*, 2025]	*ICLR*	✘	U+M	81.1	79.8	80.9	64	65.4	65.5	90.0	90.1	90.2	88.3	88.6	89.0	81.1
ProDe-V$^{\dagger}{}_{B/32}$ [Tang *et al.*, 2025]	*ICLR*	✘	U+M	82.5	82.5	83.0	72.7	72.5	72.6	92.3	91.5	92.2	90.5	90.7	90.8	84.5
MMGA$^{\dagger}{}_{R/50}$ [Chen *et al.*, 2025]	*PR*	✘	U+M	77.2	75.9	79.4	65.8	65.3	66.1	84.5	86.9	88.7	85.9	86.1	86.1	79.0
Co-learn$^{\dagger}{}_{L/14}$ [Zhang *et al.*, 2025]	*IJCV*	✘	U+M	77.1	76.6	82.0	77.2	72.5	79.6	90.4	88.1	93.0	91.0	90.0	90.1	84.0
DTKI$^{\dagger}{}_{B/32}$ [Zhan *et al.*, 2026]	*IPM*	✘	U+M	82.3	82.8	82.7	72.0	71.1	71.7	91.7	91.9	91.6	90.0	90.0	90.3	84.0
CLIP-RN50 [Radford *et al.*, 2021]	*ICML*	✓	M		71.5			54.5			80.0			82.3		72.1
ImCapFusionSFDA [Xiang *et al.*, 2025]	*ESWA*	✓	M		81.9			77.3			88.9			90.0		84.5
DSSG	*Ours*	✓	M		**85.4**			**84.1**			**94.8**			**92.8**		**89.3**
ViT-B/16:																
SHOT [Liang *et al.*, 2020]	*ICML*	✘	U	76.6	76.3	80.4	67.1	65.3	66.7	83.5	83.4	83.4	85.5	83.7	85.3	78.1
DIPE [Wang *et al.*, 2022]	*CVPR*	✘	U	77.1	75.3	81.6	66.0	63.3	67.7	80.8	83.5	89.6	85.6	83.4	85.1	78.2
DSiT [Sanyal *et al.*, 2023]	*ICCV*	✘	U	80.7	77.9	82.4	69.2	67.9	68.3	83.5	86.1	89.8	87.3	86.2	86.6	80.5
DPC$^{\dagger}{}_{B/16}$ [Zhan *et al.*, 2024]	*IJCAI*	✘	U+M	86.3	86.6	87.6	71.1	74.1	75.0	87.1	90.9	91.8	91.3	91.6	91.8	85.4
CLIP-ViT/B-16 [Radford *et al.*, 2021]	*ICML*	✓	M		82.3			71.3			89.2			89.9		83.2
UPL [Huang *et al.*, 2022]	*ArXiv*	✓	M		83.3			67.7			91.5			90.7		83.3
POUF [Tanwisuth *et al.*, 2023]	*ICML*	✓	M		86.2			73.8			92.7			91.7		86.1
TFUP-T [Long *et al.*, 2024]	*ArXiv*	✓	M		86.0			74.2			93.1			91.7		86.3
ReCLIP [Hu *et al.*, 2024]	*WACV*	✓	M		86.1			76.0			93.9			92.0		87.0
ImCapFusionSFDA [Xiang *et al.*, 2025]	*ESWA*	✓	M		87.8			87.1			94.5			94.3		90.9
DSSG	*Ours*	✓	M		**90.1**			**89.4**			**97.4**			**94.6**		**92.9**

Table 1: SFDA results (%) on the Office-Home dataset. SFF: Source-Fully-Free. U: Unimodal; M: Multimodal; U+M: Unimodal Student + Multimodal Teacher. †: Auxiliary model (e.g., $_{B/32}$ for CLIP-ViT-B/32).

Method	Veneus	SFF	Mark	D	W	A	W	A	D	Avg.
				A		D		W		
ResNet-50:										
SHOT [Liang *et al.*, 2020]	*ICML*	✘	U	74.7	77.8	94	99.8	90.1	98.4	88.6
SF(DA)^2 [Hwang *et al.*, 2024]	*ICLR*	✘	U	75.7	76.8	95.8	99.8	92.1	99.0	89.9
DIFO-C-RN$^{\dagger}{}_{R/50}$ [Tang *et al.*, 2024c]	*CVPR*	✘	U+M	78.5	78.8	93.6	97.0	92.1	95.7	89.3
DIFO-C-B32$^{\dagger}{}_{B/32}$ [Tang *et al.*, 2024c]	*CVPR*	✘	U+M	83.0	83.2	97.2	98.8	95.5	97.2	92.5
ProDe-R$^{\dagger}{}_{R/50}$ [Tang *et al.*, 2025]	*ICLR*	✘	U+M	79.8	79.0	94.4	98.6	92.1	95.6	89.9
ProDe-V$^{\dagger}{}_{B/32}$ [Tang *et al.*, 2025]	*ICLR*	✘	U+M	83.1	82.5	96.8	99.8	96.4	97.0	92.6
Co-learn$^{\dagger}{}_{L/14}$ [Zhang *et al.*, 2025]	*IJCV*	✘	U+M	**85.3**	83.2	**99.2**	**100.0**	**99.7**	**99.1**	**94.4**
DTKI$^{\dagger}{}_{B/32}$ [Zhan *et al.*, 2026]	*IPM*	✘	U+M	83.2	83.1	98.2	99.2	95.1	98.8	92.9
CLIP-RN50 [Radford *et al.*, 2021]	*ICML*	✓	M	73.2		72.5		70.1		71.9
ImCapFusionSFDA [Xiang *et al.*, 2025]	*ESWA*	✓	M	78.1		88.6		86.0		84.2
DSSG	*Ours*	✓	M	**85.3**		92.6		94.1		90.6
ViT-B/16:										
SHOT [Liang *et al.*, 2020]	*ICML*	✘	U	79.4	80.2	95.3	**100.0**	94.3	99.0	91.4
DIPE [Wang *et al.*, 2022]	*CVPR*	✘	U	77.5	77.1	94.8	**100.0**	95.5	98.5	90.5
DSiT [Sanyal *et al.*, 2023]	*ICCV*	✘	U	81.7	81.9	98.0	**100.0**	97.2	**99.1**	93.0
DPC$^{\dagger}{}_{B/16}$ [Zhan *et al.*, 2024]	*IJCAI*	✘	U+M	83.9	83.7	96.9	**100.0**	97.0	98.2	93.3
CLIP-ViT-B/16 [Radford *et al.*, 2021]	*ICML*	✓	M	78.9		79.9		76.9		78.6
UPL [Huang *et al.*, 2022]	*ArXiv*	✓	M	81.4		82.6		83.6		82.5
POUF [Tanwisuth *et al.*, 2023]	*ICML*	✓	M	84.4		91.1		91.3		88.9
TFUP-T [Long *et al.*, 2024]	*ArXiv*	✓	M	84.8		90.8		93.2		89.6
ImCapFusionSFDA [Xiang *et al.*, 2025]	*ESWA*	✓	M	85.2		92.2		93.1		90.2
DSSG	*Ours*	✓	M	**87.8**		**98.6**		**98.4**		**94.9**

Table 2: SFDA results (%) on the Office-31 dataset.

both ImCapFusionSFDA and ProDe-V by a **4.8%** margin. With the ViT-B/16 backbone, DSSG improves accuracy to **92.9%** against ImCapFusionSFDA's 90.9%.

Table 2 exhibits consistent gains on Office-31. While U+M methods leverage external priors on ResNet-50, DSSG secures **90.6%**, significantly surpassing ImCapFusionSFDA at 84.2%. With ViT-B/16, DSSG attains **94.9%**, outperforming both ImCapFusionSFDA (90.2%) and DPC (93.3%), demonstrating robustness across architectures.

Results on Large-Scale Benchmarks: Mini-DomainNet and VisDA-2017. Table 3 shows that DSSG consistently outperforms the baseline ImCapFusionSFDA across various backbones, achieving **84.1%** and **90.0%** accuracy on ResNet-50, and **89.3%** and **92.1%** on ViT-B/16 for Mini-DomainNet and VisDA-2017, respectively. This sustained improvement underscores the effectiveness and generalization ability of DSSG. Detailed per-category results for VisDA-2017 are provided in Appendix A.2.

Trade-offs and Fair Comparison: U+M vs. M Paradigms. On ResNet-50, U+M methods like DIFO-R and ProDe-R (with same-level teachers) lag behind DSSG. Competitive performance for U+M emerges only when employing external teachers with significantly stronger backbones (e.g., CLIP-ViT-B/16 or L/14), resulting in a marginal edge on Office-31 ($\approx$ 3.8%) and Mini-DomainNet/VisDA-2017 ($\approx$ 1.0%), whereas DSSG maintains a significant lead on Office-Home (+4.8%). Conversely, on the ViT-B/16 backbone, DSSG consistently outperforms U+M methods across all datasets, demonstrating superior generalization without external teacher guidance. These results indicate that U+M's edge on the weaker ResNet-50 backbone is contingent upon robust weight initialization via supervised source pre-training and strong external teacher guidance. In contrast, DSSG directly fine-tunes a single VLM to capture domain-specific knowledge applicable to various downstream tasks, without the strict source dependency and dual-model deployment overhead inherent in U+M methods.

4.3 Ablation and Analysis

Ablation Study on DSSG Framework. We present a component-wise analysis in Table 4 to verify the effectiveness of the proposed framework, where ① serves as the optimized baseline ImCapFusionSFDA. We use Office-Home

Methods	Venue	SFF	Mark	Mini-DomainNet												Avg.	VisDA
				P	R	S	C	R	S	C	P	S	C	P	R		Real
				C			P			R			S				
ResNet-50:																	
SHOT [Liang *et al.*, 2020]	*ICML*	✘	U	67.9	67.7	70.2	63.5	67.6	64.0	78.2	81.3	78.0	59.5	61.7	57.8	68.1	82.9
TPDS [Tang *et al.*, 2024a]	*IJCV*	✘	U	65.6	66.4	68.6	62.9	67.0	64.3	77.1	79.0	75.3	59.8	61.5	68.2	67.1	–
DIFO-C-RN$^{\dagger}{}_{R/50}$ [Tang *et al.*, 2024c]	*CVPR*	✘	U+M	74.0	74.8	74.7	73.8	74.6	74.3	89.0	88.7	88.0	69.4	70.1	69.6	76.7	88.6
DIFO-C-B32$^{\dagger}{}_{B/32}$ [Tang *et al.*, 2024c]	*CVPR*	✘	U+M	80.0	80.8	80.5	76.6	77.3	76.7	87.2	87.4	87.3	74.9	75.6	75.5	80.0	90.3
ProDe-R$^{\dagger}{}_{R/50}$ [Tang *et al.*, 2025]	*ICLR*	✘	U+M	80.0	80.4	80.4	79.3	78.9	79.2	91.0	90.9	91.0	75.3	75.6	75.4	81.5	88.7
ProDe-V$^{\dagger}{}_{B/32}$ [Tang *et al.*, 2025]	*ICLR*	✘	U+M	85.0	85.5	85.5	83.2	83.1	**83.4**	92.4	92.3	92.4	79.0	79.3	79.1	**85.0**	**91.0**
MMGA$^{\dagger}{}_{R/50}$ [Chen *et al.*, 2025]	*PR*	✘	U+M	74.6	75.9	–	–	76.1	75.6	–	88.3	–	69.9	–	70.9	75.9	88.6
Co-learn$^{\dagger}{}_{L/14}$ [Zhang *et al.*, 2025]	*IJCV*	✘	U+M	78.9	85.4	81.2	75.1	79.1	73.8	86.5	86.7	84.4	78.5	76.8	76.7	80.3	88.6
DTKI$^{\dagger}{}_{B/32}$ [Zhan *et al.*, 2026]	*IPM*	✘	U+M	82.4	80.7	82.3	77.5	78.7	78.9	88.9	88.0	88.3	76.1	76.9	74.6	81.1	90.8
CLIP-RN [Radford *et al.*, 2021]	*ICML*	✓	M		69.1			67.9			84.8			62.9		71.2	84.4
ImCapFusionSFDA [Xiang *et al.*, 2025]	*ESWA*	✓	M		83.2			77.6			91.0			78.9		82.7	89.0
DSSG	*Ours*	✓	M		**85.6**			77.9			**92.9**			**80.0**		84.1	90.0
ViT-B/16:																	
SHOT [Liang *et al.*, 2020]	*ICML*	✘	U	68.9	72.3	74.0	64.8	70.6	69.2	82.3	84.0	83.6	63.1	62.7	61.7	71.4	–
AaD [Yang *et al.*, 2022]	*NIPS*	✘	U	70.4	74.6	76.4	66.8	72.1	71.2	81.0	84.0	82.8	63.8	65.4	63.8	72.7	–
DPC$^{\dagger}{}_{B/16}$ [Zhan *et al.*, 2024]	*IJCAI*	✘	U+M	89.7	86.1	87.2	82.5	82.9	**84.8**	89.6	91.2	89.2	82.1	80.9	81.4	85.6	–
CLIP-ViT/B-16 [Radford *et al.*, 2021]	*ICML*	✓	M		84.1			77.6			89.0			79.0		82.4	88.0
ImCapFusionSFDA [Xiang *et al.*, 2025]	*ESWA*	✓	M		90.2			83.3			94.3			86.8		88.7	91.8
DSSG	*Ours*	✓	M		**91.1**			83.7			**94.9**			**87.6**		**89.3**	**92.1**

Table 3: SFDA results (%) on the Mini-DomainNet and VisDA-2017. '–': Unreported results.

with ResNet-50 for representative analysis. **(1) Contribution of Individual Modules.** Independently integrating the DSG module (②) or static CMKD (③) boosts the 84.50% baseline to 87.38% and 87.09%, respectively. This indicates that DSG introduces structural constraints to alleviate semantic drift, whereas CMKD leverages self-distillation to transfer CLIP priors to the student task head. **(2) Impact of Dynamic Calibration.** Removing the DSG module (④†) causes severe performance degradation, verifying the inherent dependence of dynamic CMKD on the co-evolutionary semantic space. Conversely, the integration of DSG with static CMKD (⑤) yields 89.04%, validating the effectiveness of their joint optimization. Furthermore, incorporating dynamic CMKD (⑥) advances performance to **89.28%**, suggesting that dynamic teacher–student calibration alleviates feature misalignment inherent in static class embeddings. **(3) Cross-Dataset Generalization.** The progressive gains from individual modules to the full DSSG framework are consistently observed across all benchmarks, including Office-31, Mini-DomainNet, and VisDA-2017, demonstrating the framework's robust generalization across diverse datasets and backbones.

Ablation Study on DSG Module. We investigate the impact of distinct guidance mechanisms in the DSG module, as summarized in Table 5. **(1) Catastrophic Forgetting of Fine-tuning VLM.** Standard fine-tuning (FT w/o Stream) and anchor-only alignment ($\mathcal{T}_{anc}$) suffer from severe catastrophic forgetting. Even when their peak performance is obtained via a lower learning rate and 100-iteration evaluation (denoted as ‡), accuracy drops to 66.23% and 65.32% (vs. 74.90% baseline), indicating a severe loss of pretrained knowledge. **(2) Semantic Alignment and Divergence.** VLM-only alignment ($\mathcal{T}_{cap}$) recovers 85.09% accuracy via cross-modal alignment to alleviate forgetting. However, without source supervision, relying on unconstrained caption semantics risks semantic divergence. **(3) Effect of Dual-Stream Integration.** DSG module consistently surpasses the $\mathcal{T}_{cap}$ baseline, yielding +1.86% on RN50 and +0.94% on ViT-B/16. These gains demonstrate that dynamic class anchors provide structural regularization to stabilize the unconstrained caption stream, effectively preventing semantic divergence by balancing stability and plasticity.

Hyper-Parameter Sensitivity Analysis. We investigate the sensitivity of DSSG to hyper-parameters $\lambda_{\{1,2,3\}}$ in Eq. 10. Since ImCapFusionSFDA lacks reported results on these benchmarks, we re-implemented and optimized it via grid search on λ_1 and λ_2, establishing a rigorous lower bound (see Appendix A.3). Given identical loss definitions ($\mathcal{L}_{con}$ and $\mathcal{L}_{st}$), these optimal hyperparameters are directly inherited by DSSG. We then vary λ_3 and observe consistent outperformance over the baseline across a wide range, confirming the robustness of λ_3 (see Appendix A.4). This stability further validates that performance gains stem from structural designs, not specific tuning.

Computational Efficiency. We benchmark the efficiency of DSSG across Params, GFLOPs, Iteration Speed, and Latency. Although the dual-stream architecture slightly increases training costs, DSSG maintains inference efficiency comparable to the baseline ($\approx$ 4.4 ms latency) by caching fixed semantic anchors (detailed in Appendix A.5).

Visualization Analysis. t-SNE visualizations in Fig. 3 show that DSSG yields denser target clusters with clearer boundaries across all benchmarks, demonstrating superior discriminative capability. Grad-CAM results are available in Appendix A.6.

4.4 Discussion

Since the DSSG framework is built upon VLM-generated textual descriptions, several methodological questions naturally arise regarding this paradigm. These include the use of stronger captioning VLMs to guide fine-tuning, the need for an auxiliary text classifier, the presence of class-name tokens (and their removal/masking), and the scalability to larger

ID	DSG Module	$\mathcal{L}_{cmkd}$	CLIP-RN50														RN101
			Office31(31 classes)				Office-Home(65 classes)					MiniDomainNet(126 classes)					VisDA(12 classes)
			A	D	W	Avg.	A	C	P	R	Avg.	C	P	R	S	Avg.	Real
①	-	-	78.13	88.55	86.04	84.24	81.87	77.25	88.92	89.95	84.50	83.17	77.62	90.95	78.89	82.66	88.95
②	✓	-	81.54	90.16	92.33	88.01	82.94	82.57	92.25	91.74	87.38	83.49	76.35	92.06	78.89	82.70	89.72
③	-	✓	82.00	90.16	92.70	88.29	83.48	80.38	92.43	92.06	87.09	**85.87**	77.62	92.22	80.32	84.01	89.12
④	-†	✓⋆	84.06	81.12	82.64	83.35	83.07	69.19	91.24	87.80	82.83	84.76	76.67	91.75	76.51	82.42	84.07
⑤	✓	✓	83.95	91.77	93.58	89.77	85.25	83.96	94.30	92.66	89.04	84.92	76.35	92.70	**80.48**	83.61	89.60
⑥	✓	✓⋆	**85.27**	**92.57**	**94.09**	**90.64**	**85.41**	**84.08**	**94.80**	**92.82**	**89.28**	85.56	**77.94**	**92.86**	80.00	**84.09**	**89.98**

ID	DSG Module	$\mathcal{L}_{cmkd}$	CLIP-ViT-B/16														VisDA(12 classes)
			Office31(31 classes)				Office-Home(65 classes)					MiniDomainNet(126 classes)					
			A	D	W	Avg.	A	C	P	R	Avg.	C	P	R	S	Avg.	Real
①	-	-	85.20	92.17	93.08	90.15	87.76	87.06	94.50	94.29	90.90	90.16	83.33	94.29	86.83	88.65	91.75
②	✓	-	85.55	95.38	95.35	92.09	89.62	88.06	96.91	94.38	92.24	89.68	82.86	**94.92**	87.30	88.69	92.08
③	-	✓	86.16	92.97	92.20	90.44	88.75	88.04	94.66	94.22	91.42	90.32	83.33	94.13	87.14	88.73	91.78
④	-†	✓⋆	87.50	87.95	91.19	88.88	87.80	82.96	94.32	94.15	89.81	90.16	81.75	94.92	86.19	88.26	90.68
⑤	✓	✓	86.01	96.18	94.97	92.39	90.07	89.21	97.25	94.58	92.78	**91.11**	83.17	**94.92**	**87.62**	89.21	92.07
⑥	✓	✓⋆	**87.79**	**98.59**	**98.36**	**94.91**	**90.11**	**89.39**	**97.36**	**94.61**	**92.87**	**91.11**	**83.65**	**94.92**	**87.62**	**89.33**	**92.13**

Table 4: Ablation study of DSSG. 'DSG Module' indicates the dual-stream guiding mechanism. $\mathcal{L}_{cmkd}$ represents cross-modal knowledge distillation loss (⋆ for dynamic). † isolates dynamic CMKD without DSG ($\lambda_2\mathcal{L}_{st} + \lambda_3\mathcal{L}^{*}_{cmkd}$) to validate its dependency.

Method	O-31	O-Home	MiniDN	VisDA	Mean
Backbone: CLIP-RN50					
CLIP-ZS	71.90	72.10	71.20	84.40	74.90
FT w/o Stream‡	71.35	66.43	50.80	76.33	66.23
$\mathcal{T}_{anc}$ (Anchor-only)‡	71.95	64.80	48.29	76.24	65.32
$\mathcal{T}_{cap}$ (VLM-only)	84.24	84.50	82.66	88.95	85.09
DSG (Ours)	**88.01**	**87.38**	**82.70**	**89.72**	**86.95**
Backbone: CLIP-ViT-B/16					
CLIP-ZS	78.60	83.30	82.40	88.00	83.08
FT w/o Stream	85.92	85.77	85.64	90.39	86.93
$\mathcal{T}_{anc}$ (Anchor-only)	86.03	86.49	86.47	90.31	87.33
$\mathcal{T}_{cap}$ (VLM-only)	90.15	90.90	88.65	91.75	90.34
DSG (Ours)	**92.09**	**92.24**	**88.69**	**92.08**	**91.28**

Table 5: Ablation study of the DSG module. 'FT w/o Stream': fine-tuning w/o guidance. ‡: Catastrophic forgetting.

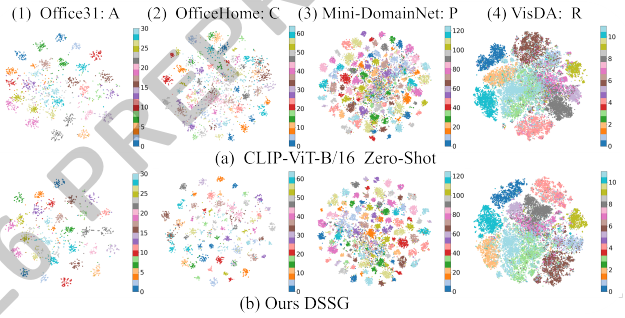


Figure 3: t-SNE visualization of the unlabeled target domain across four benchmarks, with sample counts indicated in the top-left.

backbones like ViT-L/14. As ImCapDA [Xiang *et al.*, 2025] has already extensively validated these basic mechanisms, we do not repeat these analyses and refer readers to the original work for further discussion.

Instead, we focus on caption-based SFDA and compare DSSG with LLaVA-v1.6 [Liu *et al.*, 2023], InstructBLIP [Dai *et al.*, 2023], and ShareGPT4V [Chen *et al.*, 2024b]. These methods employ large VLMs for caption generation and an STS module [Chen *et al.*, 2024a] for classification (see Table 6). DSSG integrates visual features with captions via smaller VLMs (e.g., CLIP 0.12B), surpassing large-scale VLMs (up to 34B) in both accuracy and efficiency, even when accounting for BLIP-3 (4.1B) [Xue *et al.*, 2024]. Moreover, DSSG enables flexible deployment by supporting lightweight models, API-based generation, or pre-computed captions.

Method	Generator (Param.)	Office-Home					MiniDomainNet					VisDA
		A	C	P	R	Avg.	C	P	R	S	Avg.	Val
LLaVA-34B†	Self (∼34B)	87.0	78.3	93.7	89.5	87.1	85.5	**84.4**	91.0	83.7	86.2	**92.1**
InstBLIP-XXL†	Self (∼12B)	82.2	82.0	91.6	88.8	86.2	86.7	82.5	89.0	83.0	85.3	86.7
ShrGPT4V-13B†	Self (∼13B)	83.2	66.7	85.8	84.8	80.1	79.9	79.7	87.9	79.2	81.7	90.4
DSSG-0.12B	BLIP-3 (∼4.1B)	**90.1**	**89.4**	**97.4**	**94.6**	**92.9**	**91.1**	83.7	**94.9**	**87.6**	**89.3**	**92.1**

Table 6: Comparison with caption-based SFDA methods. †: Uses large VLMs for captioning and an STS module (*all-MiniLM-L6-v2*, ∼0.02B) for classification.

5 Conclusion

In this paper, we investigate the challenging Source-Fully-Free Domain Adaptation (SFF-DA) task, identifying Dual Semantic Drift—the conflict between static embedding rigidity and caption divergence—as the primary obstacle to effective adaptation. To address this, we propose the DSSG framework, which reconciles the stability-plasticity dilemma by integrating complementary semantic streams. Specifically, the DSG module harnesses domain-specific knowledge from generated captions while anchoring global categorical consistency via class anchors. Furthermore, the Dynamic CMKD module calibrates the evolving teacher distribution via self-distillation to enforce teacher–student consistency. Extensive experiments on multiple benchmarks demonstrate that DSSG significantly outperforms existing methods, establishing a new state-of-the-art. Our work provides a novel perspective on balancing universal knowledge preservation and domain-specific knowledge acquisition in vision-language foundation models, paving the way for more robust unsupervised deployment in privacy-sensitive and source-constrained scenarios.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62372167 and 62272156, and in part by the Natural Science Foundation of Hunan Province of China under Grants 2025JJ70455 and 2024JJ5097.

References

[Ali *et al.*, 2025] Eman Ali, Sathira Silva, and Muhammad Haris Khan. Dpa: Dual prototypes alignment for unsupervised adaptation of vision-language models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6083–6093. IEEE, 2025.

[Chen *et al.*, 2024a] Dongjie Chen, Kartik Patwari, Zhengfeng Lai, Xiaoguang Zhu, Sen-ching Cheung, and Chen-Nee Chuah. Empowering source-free domain adaptation via mllm-guided reliability-based curriculum learning. *arXiv preprint arXiv:2405.18376*, 2024.

[Chen *et al.*, 2024b] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024.

[Chen *et al.*, 2025] Lijuan Chen, Yunxiang Bai, Ying Hu, Qiong Wang, and Xiaozhi Qi. Source-free domain adaptation via multimodal space-guided alignment. *Pattern Recognition*, page 112827, 2025.

[Dai *et al.*, 2023] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

[Hu *et al.*, 2024] Xuefeng Hu, Ke Zhang, Lu Xia, Albert Chen, Jiajia Luo, Yuyin Sun, Ken Wang, Nan Qiao, Xiao Zeng, Min Sun, et al. Reclip: Refine contrastive language image pre-training with source free domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2994–3003, 2024.

[Huang *et al.*, 2022] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022.

[Hwang *et al.*, 2024] Uiwon Hwang, Jonghyun Lee, Juhyeon Shin, and Sungroh Yoon. Sf(da): Source-free domain adaptation through the lens of data augmentation. 2024.

[Li *et al.*, 2025] Jindong Li, Yongguang Li, Yali Fu, Jiahong Liu, Yixin Liu, Menglin Yang, and Irwin King. Clip-powered domain generalization and domain adaptation: A comprehensive survey. *arXiv preprint arXiv:2504.14280*, 2025.

[Liang *et al.*, 2020] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[Long *et al.*, 2024] Sifan Long, Linbin Wang, Zhen Zhao, Zichang Tan, Yiming Wu, Shengsheng Wang, and Jingdong Wang. Training-free unsupervised prompt for vision-language models. *arXiv preprint arXiv:2404.16339*, 2024.

[Mirza *et al.*, 2023] Muhammad Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Horst Possegger, Mateusz Kozinski, Rogerio Feris, and Horst Bischof. Lafter: Label-free tuning of zero-shot classifier using language and unlabeled image collections. *Advances in Neural Information Processing Systems*, 36:5765–5777, 2023.

[Radford *et al.*, 2021] 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.

[Sanyal *et al.*, 2023] Sunandini Sanyal, Ashish Ramayee Asokan, Suvaansh Bhambri, Akshay Kulkarni, Jogendra Nath Kundu, and R Venkatesh Babu. Domain-specificity inducing transformers for source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18928–18937, 2023.

[Tang *et al.*, 2024a] Song Tang, An Chang, Fabian Zhang, Xiatian Zhu, Mao Ye, and Changshui Zhang. Source-free domain adaptation via target prediction distribution searching. *International journal of computer vision*, 132(3):654–672, 2024.

[Tang *et al.*, 2024b] Song Tang, Wenxin Su, Mao Ye, and Xiatian Zhu. Source-free domain adaptation with frozen multimodal foundation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23711–23720, 2024.

[Tang *et al.*, 2024c] Song Tang, Wenxin Su, Mao Ye, and Xiatian Zhu. Source-free domain adaptation with frozen multimodal foundation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23711–23720, 2024.

[Tang *et al.*, 2025] Song Tang, Wenxin Su, Yan Gan, Mao Ye, Jianwei Zhang, and Xiatian Zhu. Proxy denoising for source-free domain adaptation. 2025.

[Tanwisuth *et al.*, 2023] Korawat Tanwisuth, Shujian Zhang, Huangjie Zheng, Pengcheng He, and Mingyuan Zhou. Pouf: Prompt-oriented unsupervised fine-tuning for large pre-trained models. In *International conference on machine learning*, pages 33816–33832. PMLR, 2023.

[Wang *et al.*, 2022] Fan Wang, Zhongyi Han, Yongshun Gong, and Yilong Yin. Exploring domain-invariant parameters for source free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2022.

[Xiang *et al.*, 2025] Weiwei Xiang, Guangyi Xiao, Shun Peng, Hao Chen, and Li-Ming Ding. Imcapda: Fine-tuning clip via image captions for unsupervised domain adaptation. *Expert Systems with Applications*, 2025.

[Xue *et al.*, 2024] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024.

[Yang *et al.*, 2022] Shiqi Yang, Shangling Jui, Joost Van De Weijer, et al. Attracting and dispersing: A simple approach for source-free domain adaptation. *Advances in Neural Information Processing Systems*, 35:5802–5815, 2022.

[Zhan *et al.*, 2024] Mengmeng Zhan, Zongqian Wu, Rongyao Hu, Ping Hu, Heng Tao Shen, and Xiaofeng Zhu. Towards dynamic-prompting collaboration for source-free domain adaptation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 182–188, 2024.

[Zhan *et al.*, 2026] Mengmeng Zhan, Zongqian Wu, Jiaying Yang, Lin Peng, Jialie Shen, and Xiaofeng Zhu. Dual transferable knowledge interaction for source-free domain adaptation. *Information Processing & Management*, 63(1):104302, 2026.

[Zhang *et al.*, 2022a] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer, 2022.

[Zhang *et al.*, 2022b] Ziyang Zhang, Houwen Chen, Xiaoxuan Jiang, and Yunhe Lin. Domain adaptation via prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16601–16611, 2022.

[Zhang *et al.*, 2025] Wenyu Zhang, Li Shen, and Chuan-Sheng Foo. Source-free domain adaptation guided by vision and vision-language pre-training. *International Journal of Computer Vision*, 133(2):844–866, 2025.

[Zhou and Zhou, 2024] Wenlve Zhou and Zhiheng Zhou. Unsupervised domain adaption harnessing vision-language pre-training. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[Zhu *et al.*, 2025] Ronghang Zhu, Mengxuan Hu, Weiming Zhuang, Lingjuan Lyu, Xiang Yu, and Sheng Li. Revisiting source-free domain adaptation: Insights into representativeness, generalization, and variety. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25688–25697, 2025.