BEYOND TEXT ANCHORS: DUAL-ANCHOR FLOW MATCHING FOR FEW-SHOT AND INCREMENTAL ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow-based adaptations improve few-shot recognition with vision–language models by transporting visual features toward their corresponding class textual prototype. However, they implicitly assume that a semantically correct textual prototype is a reliable and discriminative transport endpoint. This assumption may fail when textual class geometry is crowded or misaligned with the visual distinctions required by a downstream task. A support-derived visual prototype offers a natural alternative by capturing task-specific appearance cues, but can itself be unreliable under scarce supervision. This complementary reliability motivates **DualFlow**, a dual-anchor flow-matching framework that retains textual and visual prototypes as two endpoint hypotheses rather than committing solely to either one. DualFlow uses a shared flow backbone with two anchor-specific heads to learn textual and visual-guided velocities, which are adaptively coordinated by a support-derived reliability coefficient shared with the classifier anchors. Beyond conventional few-shot adaptation, this dual-anchor transport-based paradigm is naturally suited to few-shot class-incremental learning (FSCIL): newly arriving classes can be incorporated by appending their textual and visual anchors and adapting only the lightweight velocity fields, while frozen vision–language encoders preserve a shared representation geometry for previously learned classes. Experiments across conventional and incremental few-shot adaptation benchmarks show that DualFlow substantially improves over text-only flow adaptation and complements parameter-efficient VLM adaptation, highlighting endpoint reliability as a critical factor in discriminative feature transport.

1 INTRODUCTION

Pre-trained vision–language models (VLMs), such as CLIP (Radford et al., 2021), acquire broad recognition capabilities by associating visual content with natural-language supervision at scale. Despite their strong zero-shot transferability, the representations learned during pre-training are not always well suited to specialized downstream tasks, particularly when recognition depends on domain-specific or fine-grained visual distinctions. Few-shot adaptation therefore leverages a small set of labeled examples to specialize VLMs while retaining their transferable knowledge. Most existing approaches follow parameter-efficient fine-tuning paradigms, such as prompt learning (Zhou et al., 2022c;b) and low-rank adaptation (Zanella & Ben Ayed, 2024). More recently, flow-based methods have introduced a multi-step transport perspective on few-shot adaptation (Jiang et al., 2026; Li et al., 2026), using flow matching to learn a conditional velocity field that transports visual features toward their corresponding textual prototype constructed from class names and prompt templates through the text encoder. Despite improving few-shot recognition through this transport formulation, these methods implicitly treat the corresponding textual prototype as a reliable and discriminative destination for visual feature transport.

However, a semantically valid textual prototype does not necessarily constitute a reliable and discriminative transport endpoint, leading to two overlooked limitations (*cf.*, Fig. 1(a)). **1) Cross-Modal Misalignment.** Vision–language pre-training makes matched images and texts comparable, but does not guarantee that textual and visual representations induce the same class geometry. Textual prototypes may preserve general category semantics, while their inter-class geometry does

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

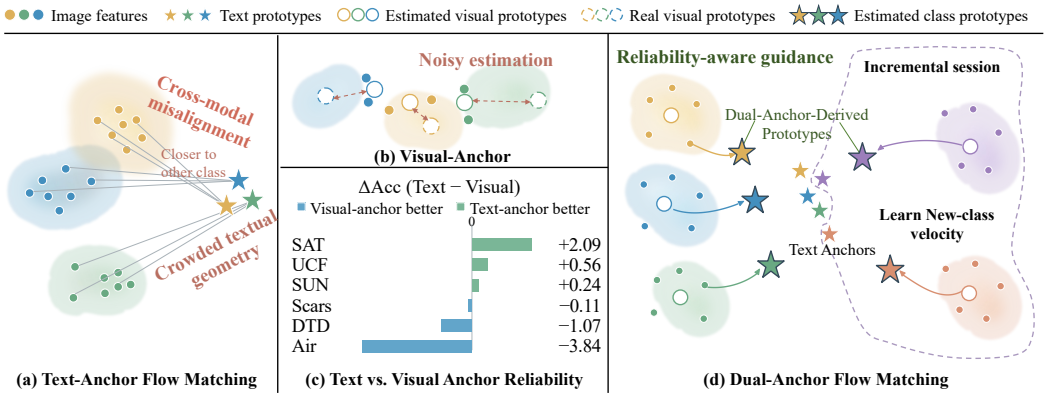


Figure 1: (a) Text-anchor flow matching may suffer from **cross-modal misalignment** and **crowded textual geometry**. (b) Visual anchors may be noisy under few-shot estimation. (c) The relative performance of textual and visual anchors varies across datasets, indicating dataset-dependent anchor reliability. (d) DualFlow combines both anchors with reliability-aware guidance for more stable prototype estimation and incremental learning.

not necessarily reflect the visual distinctions required by the downstream task (Qian et al., 2023; Liang et al., 2022). Transporting visual features toward these endpoints may improve cross-modal alignment, but this does not necessarily translate into sufficiently discriminative representations. **2) Crowded Textual Geometry.** For fine-grained categories, textual prototypes derived from semantically similar class names can become densely clustered, resulting in small inter-class margins (Ma et al., 2023). In flow-based alignment, textual prototypes serve as transport endpoints that directly shape the supervision direction. Nearby textual anchors may therefore induce weakly separated transport directions and reduce the discriminability of the transported features. Thus, relying exclusively on textual endpoints can constrain the discriminability of the learned transport.

These limitations motivate introducing a support-derived visual endpoint alongside the textual one. Following prototype-based few-shot learning (Snell et al., 2017), the visual prototype V_c can be computed as the normalized mean of the support embeddings for class c. Since it is derived from the current support set and remains in the visual representation space, it captures task-specific appearance cues that may be overlooked by class names and prompts. However, under scarce supervision, it remains a noisy finite-sample estimate whose reliability depends on the representativeness of the sampled support examples (Liu et al., 2020), as shown in Fig. 1(b). Textual and visual prototypes therefore offer different advantages: the former provides a stable but potentially misaligned semantic prior, whereas the latter contributes task-specific visual evidence. Our controlled comparison in Fig. 1(c) further shows that, under the same FMA framework (Jiang et al., 2026), replacing the textual endpoint with its visual counterpart improves some datasets but degrades others. This suggests that neither endpoint is universally preferable and that their relative reliability is task-dependent.

Motivated by this observation, we propose **DualFlow**, which retains textual and visual prototypes as two complementary endpoint hypotheses rather than collapsing them into a single predetermined destination. Given the textual prototype T_c and support-derived visual prototype V_c of class c, DualFlow employs a shared flow trunk to capture common transport structure and two modality-specific heads to predict text- and visual-guided velocities. A *reliability coefficient*, estimated directly from the current support set, then balances the two velocity fields and correspondingly calibrates the textual and visual prototypes used for classification. In this way, the same modality preference governs both feature transport and prototype construction, ensuring consistency between adaptation and prediction without introducing an additional learnable fusion module. DualFlow therefore differs from prototype-level fusion methods (Xing et al., 2019; Lyu & Wang, 2023), which train a neural module to predict an adaptive mixing weight and use the resulting fused prototype for metric classification. Directly transferring this strategy to flow matching would collapse T_c and V_c into a single endpoint, eliminating their distinct transport supervision. DualFlow instead preserves the two endpoint-induced velocities and combines them according to the estimated reliability, allowing the complementary information from both modalities to guide feature transport before being consistently incorporated into the final classifier.

DualFlow further carries its dual-anchor formulation into few-shot class-incremental learning (FSCIL), where new classes are introduced sequentially with only a few labeled examples. For each incoming class, the textual and support-derived visual prototypes serve as complementary transport supervision. The shared flow trunk and modality-specific heads are then updated with the new classes, while the corresponding dual-anchor prototypes are added to the classifier prototype set for inference over all observed classes. Importantly, the visual and textual encoders remain frozen during incremental sessions, so only the velocity networks are continually adapted over a stable representation space. This preserves previously established class prototypes, while the complementary supervision from both modalities helps alleviate forgetting across sessions. In this way, DualFlow extends the same dual-anchor transport formulation from conventional few-shot adaptation to FSCIL without requiring replay, knowledge distillation, or a continual-specific architecture.

Taken together, DualFlow establishes a unified dual-anchor transport formulation for both conventional and incremental few-shot adaptation. It preserves the complementary transport dynamics induced by textual and visual endpoints, while adaptively coordinating their contributions under different downstream tasks. Our main contributions are threefold:

1) We identify *endpoint adequacy* as an overlooked issue in flow-based few-shot adaptation and propose **DualFlow**, which learns complementary text- and visual-guided velocity fields and adaptively coordinates them according to the support-conditioned reliability.

2) We naturally extend DualFlow to FSCIL by preserving a fixed representation space while continually adapting the dual-anchor velocity fields to newly arriving categories. To the best of our knowledge, DualFlow is the first VLM adaptation framework unified across conventional and incremental few-shot classification under the same formulation.

3) Extensive experiments demonstrate that DualFlow achieves highly competitive performance on both conventional and incremental few-shot benchmarks, validating the effectiveness and generalizability of the proposed dual-anchor design.

2 Related Work

Few-shot Adaptation in VLMs. Few-shot learning seeks to generalize from limited labeled data (Wang et al., 2020; Yue et al., 2020). Vision–language models such as CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), SigLIP (Zhai et al., 2023), BLIP (Li et al., 2022), and RegionCLIP (Zhong et al., 2022) provide strong transferable representations, motivating parameter-efficient adaptation through prompts (Zhou et al., 2022c;b; Lee et al., 2023; Bulat & Tzimiropoulos, 2023; Khattak et al., 2023), lightweight adapters and classifiers (Houlsby et al., 2019; Gao et al., 2024; Zhang et al., 2022; Yu et al., 2023), and low-rank updates (Hu et al., 2022; Zanella & Ben Ayed, 2024). A related line combines textual priors with support-derived visual evidence, including Tip-Adapter (Zhang et al., 2022), Proto-CLIP (P et al., 2024), LP++ (Huang et al., 2024), and cross-modal prototype mixing (Goswami et al., 2026). However, these methods typically fuse the two prototypes statically, while flow-based adapters such as FMA (Jiang et al., 2026) and HFM (Li et al., 2026) still prescribe the text prototype as the transport destination. DualFlow instead learns separate text- and visual-directed flows and integrates them according to support-dependent reliability.

Few-shot Class-Incremental Learning (FSCIL). FSCIL requires learning novel classes from few examples without forgetting previously acquired ones (Tao et al., 2020). Early methods such as CEC (Zhang et al., 2021) and FACT (Zhou et al., 2022a) preserve discriminative space or adapt classifiers, while recent approaches exploit pretrained VLMs through prompting, knowledge distillation, prototype calibration, and feature refinement (Park et al., 2024; Chen et al., 2025; Xue et al., 2025; Sinha et al., 2026; Lee et al., 2026). In contrast, DualFlow is the first framework to directly transfer a few-shot adaptation method to the FSCIL setting, preserving its strong few-shot generalization while effectively mitigating catastrophic forgetting during incremental learning.

Flow Matching (FM). Flow Matching (FM) learns time-dependent vector fields to transport samples along prescribed probability paths (Lipman et al., 2023), while Rectified Flow simplifies this transport through straighter trajectories (Liu et al., 2023). Beyond generation (Esser et al., 2024; Le et al., 2023), FM has recently been adapted for discriminative representation transport: FMA and HFM align visual features toward textual prototypes (Jiang et al., 2026; Li et al., 2026), FlowComposer composes attribute- and object-directed flows (He et al., 2026b), FlowCIR performs reference-conditioned retrieval transport (He et al., 2026a), and Zero-Flow Encoders derives a simulation-free

representation objective (Wang et al., 2026). These works establish FM as a flexible perception operator, yet existing cross-modal methods still rely on a prescribed semantic destination. DualFlow instead jointly learns dual-modal flows and re-fuses them by support-dependent reliability.

3 PRELIMINARIES

3.1 PROBLEM FORMULATION

Few-shot Adaptation. Consider an N-way K-shot classification task with a support set $(\boldsymbol{x}_i, y_i)_{i=1}^{NK}$, where $y_i \in \mathcal{C}$ and each class provides only K labeled images, together with an unlabeled query set drawn from the same class set (Wang et al., 2020). Let E_I and E_T denote the image and text encoders of a pre-trained vision–language model, respectively. We use $\mathrm{norm}(\boldsymbol{z}) = \boldsymbol{z}/\|\boldsymbol{z}\|_2$ to denote ℓ_2 normalization. The textual anchor of class c is defined as $T_c = \mathrm{norm}\left(E_T(\boldsymbol{p}_c)\right)$, where $\boldsymbol{p}_c$ is a prompt containing the name of class c. In addition to the textual anchor T_c, the support images define an empirical visual anchor $V_c = \mathrm{norm}\left(\frac{1}{K}\sum_{i:y_i=c}\mathrm{norm}\left(E_I(\boldsymbol{x}_i)\right)\right)$ where the inner normalization operates on individual support embeddings, while the outer normalization ensures that the resulting class prototype has unit norm. This prototype construction follows the metric-learning principle that a class can be represented by the mean of its support embeddings (Snell et al., 2017).

Few-shot class-incremental learning (FSCIL). FSCIL presents a sequence of sessions $\{\mathcal{S}^{(s)}\}_{s=0}^{S}$ with mutually disjoint class sets $\{\mathcal{C}^{(s)}\}_{s=0}^{S}$. The base session $\mathcal{S}^{(0)}$ contains comparatively sufficient data, whereas each incremental session introduces only a few examples from new classes. After session s, the model aims to classify the union $\mathcal{C}^{(\le s)} = \bigcup_{j=0}^{s}\mathcal{C}^{(j)}$ without revisiting the full historical training set (Tao et al., 2020).

3.2 FLOW MATCHING FOR FEATURE TRANSPORT

Flow Matching learns a time-dependent velocity field by regressing the derivative of a prescribed path (Lipman et al., 2023). For a normalized source feature $\boldsymbol{z}$ and an anchor A, we construct the deterministic linear path $\boldsymbol{z}_t = (1-t)\boldsymbol{z} + tA$ with $t \sim \mathcal{U}(0,1)$ whose target velocity is $u(\boldsymbol{z}_t, t \mid \boldsymbol{z}, A) = \frac{\mathrm{d}\boldsymbol{z}_t}{\mathrm{d}t} = A - \boldsymbol{z}$. A velocity network v_θ is optimized with

$$\mathcal{L}_{\mathrm{FM}}(A) = \mathbb{E}_{\boldsymbol{z},A,t}\left[\|v_\theta(\boldsymbol{z}_t, t) - (A - \boldsymbol{z})\|_2^2\right]. \tag{1}$$

Previous cross-modal flow adaptations instantiate A with the corresponding textual prototype and transport visual features toward the semantic classifier (Jiang et al., 2026). The central question addressed here is what the velocity should be when that prescribed textual endpoint is not the most discriminative anchor for the downstream task.

4 METHOD

4.1 DUAL-ANCHOR FLOW MATCHING

Given a normalized visual feature $\boldsymbol{z}$ from class c, we construct two complementary transport targets: the textual prototype T_c and the support-derived visual prototype V_c. For each modality $M \in \{T, V\}$, we independently sample $t_M \sim \mathcal{U}(0,1)$ and define the linear interpolation $\boldsymbol{z}_{t_M}^M = (1-t_M)\boldsymbol{z} + t_M A_c^M$ where $A_c^T = T_c$ and $A_c^V = V_c$. The corresponding conditional target velocity is $\boldsymbol{u}_M = A_c^M - \boldsymbol{z}$. Rather than collapsing the two anchors into a single target before transport, DualFlow learns two endpoint-specific velocity fields. The two branches share a lightweight backbone h_θ while using modality-specific prediction heads ψ_M:

$$v_M(\boldsymbol{z}_t^M, t_M) = \psi_M\left(h_\theta(\boldsymbol{z}_t^M, t_M)\right). \tag{2}$$

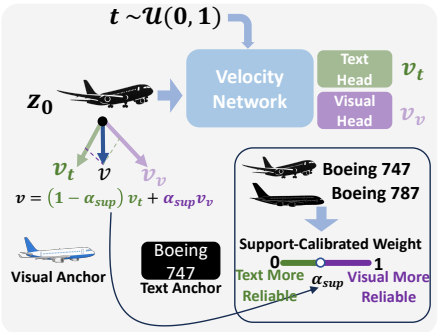


Figure 2: **Illustration of DualFlow.**

The shared backbone captures transport structure common to both branches, whereas the separate heads preserve the distinct textual- and visual-guided transport dynamics. The two branches are jointly optimized with the flow-matching objective

$$\mathcal{L}_{\text{DualFlow}} = \mathbb{E}_{\mathbf{z},c,t_T,t_V}\left[\sum_{M\in\{T,V\}} \left\|v^M(\mathbf{z}^M_{t_M}, t_M) - (A^M_c - \mathbf{z})\right\|_2^2\right]. \tag{3}$$

This formulation retains the complementary supervision provided by the two prototype modalities while introducing only a small number of additional parameters.

4.2 Theoretical Motivation for Dual-Anchor Fusion

The two branches above preserve endpoint-specific transport dynamics through separate prediction heads. We further justify that their complementary information can be effectively combined by composing the two transport directions from the perspective of endpoint estimation error. Let $\boldsymbol{\mu}_c \in \mathbb{S}^{d-1}$ denote the ideal class center. We view the textual and support-derived visual prototypes as *biased estimates*,

$$T_c \approx \boldsymbol{\mu}_c + \boldsymbol{e}_c^T, \qquad V_c \approx \boldsymbol{\mu}_c + \boldsymbol{e}_c^V, \tag{4}$$

where $\boldsymbol{e}_c^T$ and $\boldsymbol{e}_c^V$ denote their respective estimation errors. Since transporting $\boldsymbol{z}$ toward an anchor A uses the target velocity $A - \boldsymbol{z}$, its deviation from the ideal velocity $\boldsymbol{\mu}_c - \boldsymbol{z}$ is exactly $A - \boldsymbol{\mu}_c$. Hence, prototype estimation error is directly inherited by the corresponding transport direction.

Since the endpoint estimation error is directly inherited by its transport direction, a natural question is whether the complementary errors of the two anchors can be reduced through their joint use. To study this, we consider a convex composition of the two target transport directions $v_\alpha = (1-\alpha)(T_c - \boldsymbol{z}) + \alpha(V_c - \boldsymbol{z})$ where $\alpha \in [0,1]$. This composition is equivalent to transporting toward an interpolated endpoint $\widetilde{A}_c(\alpha) = (1-\alpha)T_c + \alpha V_c$. Let $\mathbb{E}$ average uniformly over classes and over support sampling. Writing $\mathcal{R}^M = \mathbb{E}\|\boldsymbol{e}_c^M\|_2^2$ and $C = \mathbb{E}[(\boldsymbol{e}_c^T)^\top \boldsymbol{e}_c^V]$, the expected endpoint error $\mathcal{R}(\alpha) = \mathbb{E}\|\widetilde{A}_c(\alpha) - \boldsymbol{\mu}_c\|_2^2$ is:

$$\mathcal{R}(\alpha) = (1-\alpha)^2\mathcal{R}^T + \alpha^2\mathcal{R}^V + 2\alpha(1-\alpha)C. \tag{5}$$

This risk is convex since $D = \mathcal{R}^T + \mathcal{R}^V - 2C = \mathbb{E}\|\boldsymbol{e}_c^T - \boldsymbol{e}_c^V\|_2^2 \geq 0$. For $D > 0$, its minimizer is $\alpha^\star = \text{clip}_{[0,1]}((\mathcal{R}^T - C)/D)$; for $D = 0$, the risk is constant. The optimum satisfies $\mathcal{R}(\alpha^\star) \leq \min\{\mathcal{R}^T, \mathcal{R}^V\}$, with strict improvement when $C < \min\{\mathcal{R}^T, \mathcal{R}^V\}$.

Therefore, when the two anchors provide complementary estimation errors, an appropriate fusion can yield a more accurate transport target than either anchor alone. Detailed derivations and sufficient conditions are provided in the appendix.

4.3 Support-Calibrated Dual-Anchor Fusion

The above analysis suggests that an appropriate fusion coefficient should reflect the relative reliability of the textual and visual anchors. However, the true risks $\mathcal{R}^T$, $\mathcal{R}^V$, and their cross-moment are unavailable in the few-shot setting. We therefore estimate their relative reliability from the discriminability of the support set. Intuitively, a reliable anchor modality should induce a geometry in which support samples are well separated from competing classes. Let $\boldsymbol{z}_i^c$ denote the i-th normalized visual feature of class c, and define $[a]_+ = \max(a, 0)$. For the textual prototypes, we measure the one-vs-rest classification margin and define the class-wise downside risk as

$$m_{c,i}^T = \langle \boldsymbol{z}_i^c, T_c\rangle - \max_{k\neq c}\langle \boldsymbol{z}_i^c, T_k\rangle, \quad \widehat{\mathcal{R}}^T_{\downarrow,c} = \frac{1}{K}\sum_{i=1}^{K}[-m_{c,i}^T]_+. \tag{6}$$

A positive margin indicates that the correct textual anchor is ranked above all competing anchors, whereas a negative margin reflects a classification-conflicting geometry. The downside risk therefore measures both the frequency and magnitude of such violations.

We estimate the visual-anchor reliability in the same manner. When $K > 1$, directly evaluating $\boldsymbol{z}_i^c$ against V_c is optimistically biased because the sample itself contributes to the prototype. We

therefore use the leave-one-out visual prototype and compute

$$V_{c,-i} = \mathrm{norm}\left(\frac{1}{K-1}\sum_{j\neq i} \boldsymbol{z}_j^c,\right), \quad m_{c,i}^{V,\mathrm{loo}} = \langle \boldsymbol{z}_i^c, V_{c,-i}\rangle - \max_{k\neq c}\langle \boldsymbol{z}_i^c, V_k\rangle. \tag{7}$$

The corresponding downside risk is $\widehat{\mathcal{R}}^V_{\downarrow,c} = \frac{1}{K}\sum_{i=1}^{K}[-m_{c,i}^{V,\mathrm{loo}}]_+$.

For the singleton case $K = 1$, leave-one-out estimation is unavailable. We therefore estimate the reliability of the visual anchor from the feature variation induced by stochastic augmentations of the single support image. In this case, a singleton visual anchor is treated as less reliable when its local uncertainty is large relative to its separation from nearby classes.

We macro-average the class-wise risks for each modality and define the support-calibrated visual weight following the form of the optimal coefficient in § 4.2:

$$\widehat{\mathcal{R}}^M_{\downarrow} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \widehat{\mathcal{R}}^M_{\downarrow,c}, \quad \alpha_{\mathrm{sup}} = \frac{\widehat{\mathcal{R}}^T_{\downarrow}}{\widehat{\mathcal{R}}^T_{\downarrow} + \widehat{\mathcal{R}}^V_{\downarrow}}. \tag{8}$$

Because α_{sup} weights the visual branch, a larger textual downside risk increases the contribution of the visual anchor, whereas a larger visual downside risk suppresses it. Macro averaging gives each class equal influence, and the coefficient is estimated entirely from the available support set.

Inference. The same coefficient combines the learned velocity fields and classifier anchors:

$$v_\alpha(\boldsymbol{z}_t, t) = (1-\alpha_{\mathrm{sup}})v_T(\boldsymbol{z}_t, t) + \alpha_{\mathrm{sup}} v_V(\boldsymbol{z}_t, t), \qquad \widetilde{A}_c = (1-\alpha_{\mathrm{sup}})T_c + \alpha_{\mathrm{sup}} V_c \tag{9}$$

At the conditional-target level, the same convex combination corresponds to the fused endpoint $\widetilde{A}_c$, since $(1-\alpha_{\mathrm{sup}})(T_c - \boldsymbol{z}) + \alpha_{\mathrm{sup}}(V_c - \boldsymbol{z}) = \widetilde{A}_c - \boldsymbol{z}$. This motivates using a consistent coefficient for velocity composition and gallery fusion. Starting from the query feature $\boldsymbol{z}_0 = \boldsymbol{z}$, we integrate $\mathrm{d}\boldsymbol{z}_t/\mathrm{d}t = v_\alpha(\boldsymbol{z}_t, t)$ to obtain $\boldsymbol{z}'$, and predict the class with the highest dot-product similarity between $\boldsymbol{z}'$ and its fused anchor $\widetilde{A}_c$.

4.4 Extension to Few-Shot Class-Incremental Learning

DualFlow naturally extends to few-shot class-incremental learning (FSCIL) by updating only the lightweight velocity networks as new classes arrive, while keeping the VLM representation space fixed. In the base session, we may directly use the frozen VLM encoders or optionally optimize them with PEFT. After fixing the representation space, DualFlow follows the training procedure in § 4.1, while constructing the textual and visual anchors of the base classes for classification.

At each incremental session, the frozen encoders extract features from the incoming few-shot data $\mathcal{S}^{(s)}$. For each new class c, we construct its textual anchor T_c and visual anchor V_c and add them to the existing class prototypes. The DualFlow parameters are initialized from the previous session and updated using only the current-session data, while all previously stored anchors remain fixed.

For support-calibrated fusion, we estimate the coefficient once from the base session, $\alpha_{\mathrm{FSCIL}} = \alpha_{\mathrm{sup}}(\mathcal{S}^{(0)})$, and reuse it in all subsequent sessions. Since the base session contains substantially more classes and labeled examples, it provides a more stable estimate of the relative reliability of the textual and visual anchors than the small incremental batches.

5 Experiments

Datasets. We evaluate few-shot classification on eleven benchmarks spanning fine-grained, remote-sensing, texture, scene, object, action, and generic recognition: FGVC-Aircraft (Maji et al., 2013), EuroSAT (Helber et al., 2019), DTD (Cimpoi et al., 2014), SUN397 (Xiao et al., 2010), Stanford Cars (Krause et al., 2013), Oxford Pets (Parkhi et al., 2012), UCF101 (Soomro et al., 2012), Flowers102 (Nilsback & Zisserman, 2008), Caltech101 (Fei-Fei et al., 2004), Food101 (Bossard et al., 2014), and ImageNet (Deng et al., 2009). Following FMA (Jiang et al., 2026), the first five form the *difficult* subset and the remaining six the *easy* subset. Using standard CLIP-adaptation splits, we sample $K = \{1, 4, 16\}$ images per class with a fixed seed and evaluate on the test set.

Table 1: **Comparison with state-of-the-art flow-matching methods for few-shot adaptation** using CLIP-LoRA and CLIP-Adapter backbones. † denotes a PEFT module trained with centripetal hyperbolic alignment (Desai et al., 2023). Δ is the average improvement. For CLIP-LoRA and CLIP-Adapter, we use the trained weights provided by FMA to ensure a fair comparison.

		Difficult Datasets							Easy Datasets							
Shot	Methods	Aircraft	SAT	DTD	SUN	Cars	AVG	Δ	Pets	UCF	Flowers	Caltech	Food	Net	AVG	Δ
1	CLIP-LoRA† (Desai et al., 2023)	29.2	73.1	56.4	69.9	68.7	59.5	-	92.1	75.8	81.9	94.5	86.8	69.7	83.5	-
	+HFM (Li et al., 2026)	30.5	76.8	57.4	71.2	69.8	61.1	+1.6	92.2	76.6	85.8	95.1	85.9	70.0	84.3	+0.8
	CLIP-LoRA (Zanella & Ben Ayed, 2024)	28.0	71.9	54.1	70.3	69.4	58.7	-	91.9	75.4	81.4	93.8	85.1	70.3	83.0	-
	+FMA (Jiang et al., 2026)	28.3	73.0	55.1	70.6	69.8	59.4	+0.7	92.1	75.9	84.9	94.5	85.2	70.2	83.8	+0.8
	+DualFlow	30.2	82.1	57.1	71.3	71.2	**62.4**	**+3.7**	93.2	76.6	89.2	94.4	84.7	69.8	**84.6**	**+1.6**
4	CLIP-LoRA† (Desai et al., 2023)	40.3	89.0	66.5	74.1	77.4	69.5	-	93.4	81.0	93.7	95.7	86.8	71.4	87.0	-
	+HFM (Li et al., 2026)	43.8	90.7	67.6	75.5	79.9	**71.5**	+2.0	93.7	83.4	95.5	96.1	86.8	72.1	**87.9**	+0.9
	CLIP-LoRA (Zanella & Ben Ayed, 2024)	38.8	83.5	64.0	72.8	77.4	67.3	-	90.6	81.1	92.9	95.0	82.6	71.4	85.6	-
	+FMA (Jiang et al., 2026)	40.3	85.0	67.0	73.7	78.9	69.0	+1.7	90.8	82.4	95.0	95.8	83.2	72.0	86.5	+0.9
	+DualFlow	41.8	89.2	69.4	74.4	79.6	70.9	**+3.6**	91.8	83.2	96.4	95.9	84.0	72.4	87.3	**+1.7**
16	CLIP-LoRA† (Desai et al., 2023)	57.6	91.3	73.7	77.1	86.5	77.2	-	93.6	87.3	98.7	96.4	87.3	73.6	89.5	-
	+HFM (Li et al., 2026)	61.4	93.1	75.2	77.6	88.6	79.2	+2.0	94.0	88.9	98.7	96.8	87.3	73.7	**89.9**	+0.4
	CLIP-LoRA (Zanella & Ben Ayed, 2024)	54.7	90.7	73.0	76.0	86.0	76.1	-	91.6	86.2	97.9	96.1	84.2	73.4	88.2	-
	+FMA (Jiang et al., 2026)	57.8	91.0	75.4	77.2	87.7	77.8	+1.7	91.6	87.1	99.1	96.5	85.1	73.5	88.8	+0.6
	+DualFlow	59.5	93.2	77.1	77.8	89.0	**79.3**	**+3.2**	92.8	87.9	99.1	96.5	86.0	74.6	89.5	**+1.3**
16	CLIP-Adapter† (Desai et al., 2023)	34.2	72.3	64.1	74.2	73.6	63.7	-	92.7	82.4	93.3	95.1	87.0	71.2	87.0	-
	+HFM (Li et al., 2026)	43.8	85.7	71.4	76.1	80.6	71.5	+7.8	93.3	84.6	97.2	96.2	87.1	72.1	88.4	+1.4
	CLIP-Adapter (Gao et al., 2024)	33.8	70.4	59.3	74.3	74.2	62.4	-	92.4	80.1	93.6	94.9	87.1	71.6	86.6	-
	+FMA (Jiang et al., 2026)	35.8	85.6	69.2	74.4	74.7	67.9	+5.5	92.9	81.5	95.6	96.0	87.2	71.3	87.4	+0.8
	+DualFlow	45.9	85.7	72.3	76.4	83.1	**72.7**	**+10.3**	92.4	85.1	98.1	96.1	87.4	73.1	**88.7**	**+2.1**

Baselines. We compare DualFlow with two groups of baselines. First, PEFT methods include prompt learning (Zhou et al., 2022c;b; Chen et al., 2023; Yao et al., 2023; Zhu et al., 2023), feature- and cache-based adaptation (Zhang et al., 2022; Gao et al., 2024), and low-rank adaptation (Zanella & Ben Ayed, 2024), with frozen CLIP (Radford et al., 2021) as the zero-shot reference. Second, flow-based adaptation methods include FMA (Jiang et al., 2026) and HFM (Li et al., 2026). For a fair comparison, the flow modules are attached to matched CLIP-LoRA or CLIP-Adapter backbones. Unless otherwise specified, all methods use CLIP ViT-B/16 and identical few-shot splits.

Implementation Details. We implement DualFlow as a lightweight add-on to pre-trained CLIP, CLIP-LoRA (Zanella & Ben Ayed, 2024), and CLIP-Adapter (Gao et al., 2024). Specifically, we first extract visual and textual features from the corresponding pre-trained models, and then employ a lightweight velocity network to learn two complementary velocity fields that transport image features toward the textual and visual prototypes, respectively. The two fields are subsequently combined using the support-conditioned coefficient for few-shot adaptation. We train DualFlow with AdamW using a learning rate of 2×10^{-4} and a weight decay of 0.01. All experiments are conducted on a single NVIDIA RTX 3090 GPU with 24GB memory. More details are in Appendix A2.

Performance against Flow-based Adaptations. As shown in Tab. 1, we evaluate DualFlow on top of established PEFT baselines, including CLIP-LoRA and CLIP-Adapter. DualFlow consistently delivers larger improvements over the corresponding baselines than competing flow-matching methods across different shot settings. For instance, under the 16-shot setting with CLIP-Adapter, DualFlow improves the average accuracy on the difficult datasets by $+10.3\%$, substantially surpassing the gains achieved by FMA (Jiang et al., 2026) ($+5.5\%$) and HFM (Li et al., 2026) ($+7.8\%$). Similar improvements are observed with CLIP-LoRA, where DualFlow achieves gains of $+3.7\%$, $+3.6\%$, and $+3.2\%$ under the 1-, 4-, and 16-shot settings, respectively. These results demonstrate that DualFlow can effectively complement different PEFT-based adaptations and provides particularly strong benefits on challenging datasets.

Comparison with state-of-the-art PEFT. As presented in Tab. 2, we compare DualFlow with a comprehensive set of established PEFT methods across different shot settings. Built upon CLIP-LoRA, DualFlow consistently achieves the best performance under the 1-, 4-, and 16-shot settings on both difficult and easy dataset groups. In particular, under the 16-shot setting, DualFlow achieves average accuracies of 79.3% and 89.5% on the difficult and easy datasets, respectively, outperforming the strongest competing method PLOT++ (74.1% and 88.7%) as well as the CLIP-LoRA baseline (76.1% and 88.2%). Similar improvements are observed in the lower-shot regimes, where DualFlow improves the difficult-dataset average over CLIP-LoRA from 58.7% to 62.4% in 1-shot and from 67.3% to 70.9% in 4-shot. These results demonstrate that explicitly leveraging complementary textual and visual anchors enables more effective few-shot adaptation, with particularly pronounced benefits on challenging datasets.

Table 2: **Comparison with other state-of-the-art PEFT methods.** Building upon CLIP-LoRA, adaptation performance can be further improved by our DualFlow.

		Difficult Datasets						Easy Datasets						
Shot	Method	Aircraft	SAT	DTD	SUN	Cars	**AVG**	Pets	UCF	Flowers	Caltech	Food	Net	**AVG**
0	CLIP (Radford et al., 2021)	24.8	47.8	43.8	62.6	65.5	48.9	89.1	66.8	71.2	92.9	86.1	66.7	78.8
1	CoOp (Zhou et al., 2022c)	20.8	56.4	50.1	67.0	67.5	52.4	90.2	71.2	78.3	92.5	84.3	65.7	80.4
	CoCoOp (Zhou et al., 2022b)	28.1	55.4	52.6	68.7	67.6	54.5	91.9	70.4	73.4	94.1	84.9	69.4	80.7
	TIP-Adapter (Zhang et al., 2022)	28.8	67.8	51.6	67.2	67.1	56.5	90.6	73.4	83.8	94.0	85.8	69.4	82.8
	CLIP-Adapter (Gao et al., 2024)	25.2	49.3	44.2	65.4	65.7	50.0	89.0	66.9	71.3	92.0	86.1	67.9	78.9
	PLOT++ (Chen et al., 2023)	28.6	65.4	54.6	66.8	68.8	56.8	91.9	74.3	80.5	94.3	86.2	66.5	82.3
	KgCoOp (Yao et al., 2023)	26.8	61.9	52.7	68.4	66.7	55.3	92.1	72.8	74.7	94.2	**86.4**	68.9	81.5
	ProGrad (Zhu et al., 2023)	28.9	57.0	52.8	67.0	68.2	54.8	91.4	73.3	80.9	93.5	84.9	67.0	81.8
	CLIP-LoRA (Zanella & Ben Ayed, 2024)	28.0	71.9	54.1	70.3	69.4	58.7	91.9	75.4	81.4	93.8	85.1	**70.3**	83.0
	+ DualFlow (ours)	**30.2**	**82.1**	**57.1**	**71.3**	**71.2**	**62.4**	**93.2**	**76.6**	**89.2**	**94.4**	84.7	69.8	**84.6**
4	CoOp (Zhou et al., 2022c)	30.9	69.7	59.5	69.7	74.4	60.8	92.5	77.6	92.2	94.5	84.5	68.8	85.0
	CoCoOp (Zhou et al., 2022b)	30.6	61.7	55.7	70.4	69.5	57.6	**92.7**	75.3	81.5	94.8	86.3	70.6	83.5
	TIP-Adapter (Zhang et al., 2022)	35.7	76.8	59.8	70.8	74.1	63.4	91.9	78.1	92.1	94.8	86.5	70.7	85.7
	CLIP-Adapter (Gao et al., 2024)	27.9	51.2	46.1	68.0	67.5	52.1	90.8	70.6	73.1	94.0	86.5	68.6	80.6
	PLOT++ (Chen et al., 2023)	35.3	83.2	62.4	71.7	76.3	65.8	**92.7**	79.8	92.9	95.1	86.5	70.4	86.2
	KgCoOp (Yao et al., 2023)	32.2	71.8	58.7	71.5	69.5	60.7	92.6	77.6	87.0	95.0	**86.9**	69.9	84.8
	ProGrad (Zhu et al., 2023)	34.1	69.6	59.7	71.7	75.0	62.0	92.1	77.9	91.1	94.4	85.4	70.2	85.2
	CLIP-LoRA (Zanella & Ben Ayed, 2024)	38.8	83.5	64.0	72.8	77.4	67.3	90.6	81.1	92.9	95.0	82.6	71.4	85.6
	+ DualFlow (ours)	**41.8**	**89.2**	**69.4**	**74.4**	**79.6**	**70.9**	91.8	**83.2**	**96.4**	**95.9**	84.0	**72.4**	**87.3**
16	CoOp (Zhou et al., 2022c)	43.3	86.0	70.0	74.9	83.1	71.5	91.1	83.1	97.2	95.5	84.4	71.4	87.1
	CoCoOp (Zhou et al., 2022b)	33.8	75.5	65.8	72.8	72.4	64.1	93.2	76.0	87.1	95.2	**87.4**	71.1	85.0
	TIP-Adapter (Zhang et al., 2022)	44.6	85.9	70.8	76.0	82.3	71.9	92.6	83.9	96.2	95.7	86.8	73.4	88.1
	CLIP-Adapter (Gao et al., 2024)	34.2	71.4	59.4	74.2	74.0	62.6	92.3	80.2	92.9	94.9	87.1	69.8	86.2
	PLOT++ (Chen et al., 2023)	46.7	92.0	71.4	76.0	84.6	74.1	**93.6**	85.3	97.6	96.0	87.1	72.6	88.7
	KgCoOp (Yao et al., 2023)	36.5	76.2	68.7	73.3	74.8	65.9	93.2	81.7	93.4	95.2	87.2	70.4	86.9
	ProGrad (Zhu et al., 2023)	43.0	83.6	68.8	75.1	82.9	70.7	92.8	82.7	96.6	95.9	85.8	72.1	87.7
	CLIP-LoRA (Zanella & Ben Ayed, 2024)	54.7	90.7	73.0	76.0	86.0	76.1	91.6	86.2	97.9	96.1	84.2	73.4	88.2
	+ DualFlow (ours)	**59.5**	**93.2**	**77.1**	**77.8**	**89.0**	**79.3**	92.8	**87.9**	**99.1**	**96.5**	86.0	**74.6**	**89.5**

Table 3: **Ablation study** on Double Branch Architecture (**D**), Text Endpoint (**T**), Visual Endpoint (**V**), and Fusion Weighting (**W**) under the 1-, 4-, and 16-shot settings based on CLIP.

Shot	Method	D	T	V	W	Aircraft	SAT	DTD	SUN	Cars	**AVG**	Pets	UCF	Flowers	Caltech	Food	Net	**AVG**
0	CLIP	×	×	×	×	24.8	47.8	43.8	62.6	65.5	48.9	89.1	66.8	71.2	92.9	86.1	66.7	78.8
1	(1)	×	✓	×	×	26.9	66.4	53.4	66.5	67.1	56.1	**90.1**	72.6	80.6	94.2	**86.1**	**68.0**	81.9
	(2)	×	×	✓	×	19.6	55.8	35.8	38.8	27.1	35.4	40.6	50.3	67.9	76.2	43.8	29.3	51.4
	(3)	×	✓	✓	0.5	24.7	60.4	44.7	53.9	51.0	46.9	78.0	64.2	82.2	84.9	69.4	50.2	71.5
	(4)	✓	✓	✓	0.5	25.3	60.4	44.8	54.1	51.2	47.2	78.3	63.8	**82.6**	87.1	70.0	50.7	72.1
	Ours	✓	✓	✓	Adaptive	**28.3**	**66.5**	**53.8**	**67.0**	**67.4**	**56.6**	90.0	**72.9**	82.2	**94.4**	85.4	67.4	**82.1**
4	(1)	×	✓	×	×	31.1	76.9	60.7	70.1	71.2	62.0	92.3	77.7	92.3	95.5	86.4	68.9	85.5
	(2)	×	×	✓	×	32.2	71.1	58.2	63.9	59.1	56.9	76.6	72.1	92.1	91.3	72.1	52.6	76.1
	(3)	×	✓	✓	0.5	35.0	75.3	62.5	69.8	70.7	62.7	90.4	78.2	94.2	94.5	82.9	64.3	84.1
	(4)	✓	✓	✓	0.5	**35.8**	77.1	61.9	69.9	71.8	63.3	90.1	78.4	94.6	95.0	82.7	65.0	84.3
	Ours	✓	✓	✓	Adaptive	35.2	**77.2**	**63.9**	**72.0**	**73.1**	**64.3**	**92.6**	**79.1**	**94.8**	**95.8**	**86.6**	**69.7**	**86.4**
16	(1)	×	✓	×	×	38.1	84.3	69.5	73.5	78.6	68.8	**92.8**	82.0	97.5	**95.9**	87.0	70.4	87.6
	(2)	×	×	✓	×	45.4	84.2	70.2	73.9	80.9	70.9	87.6	81.4	97.7	95.0	82.5	66.7	85.2
	(3)	×	✓	✓	0.5	45.0	85.6	71.9	75.3	**82.8**	72.1	91.7	**84.5**	**98.1**	**95.9**	86.0	71.2	87.9
	(4)	✓	✓	✓	0.5	**45.7**	86.2	**72.6**	**75.8**	**82.8**	72.6	92.2	84.4	**98.1**	95.7	86.4	71.6	88.1
	Ours	✓	✓	✓	Adaptive	**45.7**	**86.9**	**72.6**	**75.8**	82.4	**72.7**	92.5	**84.5**	**98.1**	**95.9**	**87.2**	**72.1**	**88.4**

Component Analysis. We conduct ablations to quantify the contributions of key designs in DualFlow: the double-branch architecture (D), textual endpoint (T), visual endpoint (V), and fusion weighting (W). As shown in Tab. 3, we evaluate different combinations under all shot settings. First, single-branch variants using only the textual endpoint (Row(1)) or visual endpoint (Row(2)) show varying performance across datasets and shot settings, indicating that neither modality is uniformly reliable. Second, when both textual and visual endpoints are fused within a single branch using a fixed weight of 0.5 (Row(3)), the performance does not consistently improve, especially in the low-shot regime, suggesting that simply merging the two endpoint signals is insufficient. Third, the double-branch architecture with a fixed weight of 0.5 (Row(4)) benefits some higher-shot cases but remains suboptimal in the low-shot regime. Finally, our full model adaptively balances the two branches based on their reliability, yielding more robust performance across settings. These results demonstrate the complementarity of textual and visual supervision and the benefit of adaptive fusion.

Performance under FSCIL. Following prior FSCIL works, we evaluate DualFlow on three standard benchmarks, including CIFAR100 (Krizhevsky, 2009), miniImageNet (Ravi & Larochelle, 2017), and CUB200 (Wah et al., 2011). We compare against representative methods specifically designed for FSCIL, including SAVC (Song et al., 2023), OrCo (Ahmed et al., 2024), CLOSER (Oh

Table 4: **Comparison with state-of-the-art methods on three FSCIL benchmarks.** $\mathrm{Acc}_{\mathrm{base}}$ is the base-session accuracy, $\mathrm{Acc}_{\mathrm{navg}}$ is the average accuracy for incremental sessions, and $\mathrm{Acc}_{\mathrm{avg}}$ is the accuracy averaged over all sessions.

	Methods	miniImageNet			CUB200			CIFAR100		
		$\mathrm{Acc}_{\mathrm{base}}$	$\mathrm{Acc}_{\mathrm{navg}}$	$\mathrm{Acc}_{\mathrm{avg}}$	$\mathrm{Acc}_{\mathrm{base}}$	$\mathrm{Acc}_{\mathrm{navg}}$	$\mathrm{Acc}_{\mathrm{avg}}$	$\mathrm{Acc}_{\mathrm{base}}$	$\mathrm{Acc}_{\mathrm{navg}}$	$\mathrm{Acc}_{\mathrm{avg}}$
FSCIL-specific	SAVC [CVPR 2023]	81.12	65.29	67.05	81.85	68.10	69.35	78.94	62.27	64.13
	OrCo [CVPR 2024]	83.30	65.09	67.12	81.86	68.17	69.42	80.18	60.09	62.33
	CLOSER [ECCV 2024]	76.02	61.12	62.78	79.40	65.83	67.07	75.72	61.52	63.10
	YourSelf [ECCV 2024]	84.00	66.90	68.80	83.40	68.49	69.85	82.90	65.03	67.02
	Tri-WE [CVPR 2025]	84.13	68.93	70.62	81.26	70.01	71.03	81.78	67.39	68.99
	TEEN [NeurIPS 2023]	92.58	88.46	88.92	81.86	72.96	73.77	77.09	67.46	68.53
	Approxima [ECCV 2024]	95.21	92.28	92.61	86.46	78.38	79.12	89.09	80.41	81.38
	BiMC [CVPR 2025]	95.47	93.36	93.60	83.79	75.86	76.58	84.14	77.79	78.50
	IVFL [ACM MM 2025]	96.37	93.25	93.62	85.68	77.02	77.81	86.47	77.94	78.89
	FDR [ICCV 2025]	95.30	88.43	89.20	84.40	77.03	77.70	87.70	78.25	79.30
Few-Shot Adaptation	CLIP-LoRA [CVPRW 2024]	96.20	57.27	61.59	81.01	46.80	49.91	88.48	78.13	79.28
	FMA [ICLR 2026]	88.58	24.84	31.93	82.93	60.47	62.51	84.65	73.15	74.42
	DualFlow	96.67	89.48	90.28	85.54	75.96	76.83	86.65	76.67	77.78
	DualFlow+	96.67	90.32	91.02	85.47	73.34	74.44	89.75	82.70	83.48

et al., 2024), YourSelf (Tang et al., 2024), Tri-WE (Lee et al., 2025), TEEN (Wang et al., 2023), Approxima (Wang et al., 2024), BiMC (Chen et al., 2025), IVFL (Li et al., 2025), and FDR (Xue et al., 2025), as well as the few-shot adaptation methods FMA (Jiang et al., 2026) and CLIP-LoRA (Zanella & Ben Ayed, 2024). We report two variants of our method. **DualFlow** directly operates on the frozen embedding space of pre-trained CLIP. **DualFlow+** instead builds on a CLIP-LoRA adapted embedding space: we first train CLIP-LoRA using the base-session data, then freeze the adapted encoder throughout all subsequent sessions and train DualFlow on the resulting fixed embeddings.

As shown in Tab. 4, DualFlow achieves competitive performance across all three FSCIL benchmarks despite not being specifically designed for this setting. DualFlow+ reaches 83.48% average accuracy on CIFAR100, outperforming all compared FSCIL-specific methods, and also surpasses most of them on miniImageNet and CUB200. Compared with the flow-based few-shot method FMA (Jiang et al., 2026), both DualFlow variants achieve substantially higher average accuracy on all three datasets, indicating better robustness to performance degradation across incremental sessions. Compared with CLIP-LoRA, DualFlow with frozen CLIP embeddings performs markedly better on miniImageNet and CUB200, while CLIP-LoRA is stronger on CIFAR100. We attribute this to the relatively consistent distribution shift introduced by upsampling CIFAR100 images from 32×32 to the CLIP input resolution, together with its semantically well-separated categories, which allows session-wise LoRA updates to learn more transferable corrections with limited interference across classes. Importantly, DualFlow is complementary to such PEFT-based adaptation. By building on CLIP-LoRA features learned in the base session, DualFlow+ achieves 91.02%, 74.44%, and 83.48% on miniImageNet, CUB200, and CIFAR100, respectively, consistently improving over CLIP-LoRA and demonstrating its flexibility as an add-on for continual adaptation.

6 Conclusion

We present **DualFlow**, a dual-anchor flow-matching framework for few-shot and incremental VLM adaptation. Motivated by the overlooked limitations of treating a single textual prototype as the transport endpoint, DualFlow preserves both textual and support-derived visual prototypes and learns their complementary transport dynamics through a shared flow trunk with modality-specific velocity heads. A support-conditioned reliability coefficient coordinates the two flows and constructs a consistent classifier gallery, allowing their contributions to be adaptively balanced according to the estimated reliability of the two endpoints. We further extend the same dual-anchor formulation to FSCIL, where newly arriving classes are incorporated through their textual and visual anchors while the shared velocity fields are continually adapted over a fixed VLM feature space. Extensive experiments across conventional few-shot and FSCIL benchmarks demonstrate the effectiveness of DualFlow under both adaptation settings. Overall, our results show that, beyond learning how to transport visual features, the adequacy of the transport endpoints and their adaptive coordination are critical to effective flow-based VLM adaptation.

AI Use Statement. Generative AI tools were used to polish the manuscript and improve the clarity and readability of the writing. Their use was strictly limited to linguistic refinement and did not contribute to research ideation, data collection or generation, mathematical derivations or proofs, experimental design or implementation, data analysis, or the formulation of scientific claims and conclusions. All AI-assisted edits were carefully reviewed and, where necessary, revised by the authors to ensure that they faithfully preserved the intended scientific content. The authors take full responsibility for the accuracy, integrity, and entirety of the final manuscript, including all methods, results, analyses, and conclusions.

REFERENCES

Noor Ahmed, Anna Kukleva, and Bernt Schiele. Orco: Towards better generalization via orthogonality and contrast for few-shot class-incremental learning. In *CVPR*, 2024.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101: Mining discriminative components with random forests. In *ECCV*, 2014.

Adrian Bulat and Georgios Tzimiropoulos. LASP: Text-to-text optimization for language-aware soft prompting of vision & language models. In *CVPR*, 2023.

Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Prompt learning with optimal transport for vision-language models. In *ICLR*, 2023.

Yiyang Chen, Tianyu Ding, Lei Wang, Jing Huo, Yang Gao, and Wenbin Li. Enhancing few-shot class-incremental learning via training-free bi-level modality calibration. In *CVPR*, 2025.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *ICML*, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2004.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better vision-language models with feature adapters. *IJCV*, 2024.

Dipam Goswami, Simone Magistri, Gido M. van de Ven, Bartlomiej Twardowski, Andrew D. Bagdanov, Tinne Tuytelaars, and Joost van de Weijer. Cross-modal prototype alignment and mixing for training-free few-shot classification. *arXiv preprint arXiv:2603.24528*, 2026.

Zhenqi He, Ziqi Jiang, Yuanpei Liu, Yanghao Wang, Teng Wang, and Long Chen. FlowCIR: Semantic transport via flow matching for zero-shot composed image retrieval. In *ECCV*, 2026a.

Zhenqi He, Lin Li, and Long Chen. FlowComposer: Composable flows for compositional zero-shot learning. In *CVPR*, 2026b.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *ICML*, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. LP++: A surprisingly strong linear probe for few-shot CLIP. In *CVPR*, 2024.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

Ziqi Jiang, Yanghao Wang, and Long Chen. Exploring cross-modal flows for few-shot learning. In *ICLR*, 2026.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *CVPR*, 2023.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops*, 2013.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale. In *NeurIPS*, 2023.

Dongjun Lee, Seokwon Song, Jihee Suh, Joonmyeong Choi, Sanghyeok Lee, and Hyunwoo J. Kim. Read-only prompt optimization for vision-language few-shot learning. In *ICCV*, 2023.

Eunju Lee, MiHyeon Kim, JuneHyoung Kwon, Yoonji Lee, JiHyun Kim, Soojin Jang, and Young-Bin Kim. HyCal: A training-free prototype calibration method for cross-discipline few-shot class-incremental learning. In *CVPR*, 2026.

Juntae Lee, Munawar Hayat, and Sungrack Yun. Tripartite weight-space ensemble for few-shot class-incremental learning. In *CVPR*, 2025.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

Lin Li, Ziqi Jiang, Gefan Ye, Zhenqi He, Jiahui Li, Jun Xiao, Kwang-Ting Cheng, and Long Chen. Path-decoupled hyperbolic flow matching for few-shot adaptation. In *ICML*, 2026.

Shuo Li, Xingchen Liu, Fang Liu, Licheng Jiao, Jiahao Wang, Xinyan Huang, Yanbiao Ma, Puhua Chen, Lingling Li, Xu Liu, et al. Imagining vision from language for few-shot class-incremental learning. In *ACM MM*, 2025.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *NeurIPS*, 2024.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2022.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Jinlu Liu, Liang Song, and Yongqiang Qin. Prototype rectification for few-shot learning. In *ECCV*, 2020.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Qiang Lyu and Weiqiang Wang. Compositional prototypical networks for few-shot classification. In *AAAI*, 2023.

Wenxuan Ma, Shuang Li, JinMing Zhang, Chi Harold Liu, Jingxuan Kang, Yulin Wang, and Gao Huang. Borrowing knowledge from pre-trained language model: A new data-efficient visual learning paradigm. In *ICCV*, 2023.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

Junghun Oh, Sungyong Baik, and Kyoung Mu Lee. Closer: Towards better representation learning for few-shot class-incremental learning. In *ECCV*, 2024.

Jishnu Jaykumar P, Kamalesh Palanisamy, Yu-Wei Chao, Xinya Du, and Yu Xiang. Proto-CLIP: Vision-language prototypical network for few-shot learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2024.

Keon-Hee Park, Kyungwoo Song, and Gyeong-Moon Park. Pre-trained vision and language transformers are few-shot incremental learners. In *CVPR*, 2024.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012.

Qi Qian, Yuanhong Xu, and Juhua Hu. Intra-modal proxy learning for zero-shot visual categorization with clip. In *NeurIPS*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017.

Abhishek Kumar Sinha, Nitant Dube, and Soma Biswas. Quantized residuals to continuous prompts for few-shot class incremental learning in vision-language models. In *CVPR*, 2026.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

Zeyin Song, Yifan Zhao, Yujun Shi, Peixi Peng, Li Yuan, and Yonghong Tian. Learning with fantasy: Semantic-aware virtual contrastive constraint for few-shot class-incremental learning. In *CVPR*, 2023.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Yu-Ming Tang, Yi-Xing Peng, Jingke Meng, and Wei-Shi Zheng. Rethinking few-shot class-incremental learning: Learning from yourself. In *ECCV*, 2024.

Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *CVPR*, 2020.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Qi-Wei Wang, Da-Wei Zhou, Yi-Kai Zhang, De-Chuan Zhan, and Han-Jia Ye. Few-shot class-incremental learning via training-free prototype calibration. In *NeurIPS*, 2023.

Xuan Wang, Zhong Ji, Xiyao Liu, Yanwei Pang, and Jungong Han. On the approximation risk of few-shot class-incremental learning. In *ECCV*, 2024.

Yakun Wang, Leyang Wang, Song Liu, and Taiji Suzuki. Zero-flow encoders. In *ICML*, 2026.

Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, 2020.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

Chen Xing, Negar Rostamzadeh, Boris N. Oreshkin, and Pedro O. Pinheiro. Adaptive cross-modal few-shot learning. In *NeurIPS*, 2019.

Zongyao Xue, Meina Kan, Shiguang Shan, and Xilin Chen. Feature decomposition-recomposition in large vision-language model for few-shot class-incremental learning. In *ICCV*, 2025.

Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*, 2023.

Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *CVPR*, 2023.

Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. In *NeurIPS*, 2020.

Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *CVPR*, 2021.

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 *ECCV*, 2022.

Li-Jun Zhao, Zhen-Duo Chen, Yongxin Wang, Xin Luo, and Xin-Shun Xu. Attraction diminishing and distributing for few-shot class-incremental learning. In *CVPR*, 2025.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-based language-image pretraining. In *CVPR*, 2022.

Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, Liang Ma, Shiliang Pu, and De-Chuan Zhan. Forward compatible few-shot class-incremental learning. In *CVPR*, 2022a.

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

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022c.

Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *ICCV*, 2023.

APPENDIX

A1 Theoretical Analysis **15**

A1.1 Endpoint Error and Conditional Velocity Error 15

A1.2 Optimal Shared Fusion Coefficient 15

A1.3 Strict Improvement from Complementary Errors 16

A1.4 Class-wise Derivation 16

A2 Reproducibility and Implementation Details **16**

A2.1 Dataset Statistics 16

A2.2 Details in Support-Calibrated Dual-Anchor Fusion 17

A2.3 Additional Implementation Details 18

A3 More Experimental Results **18**

A3.1 Stability Analysis 18

A3.2 Additional Ablation Analysis in Few-Shot Adaptation 19

A3.3 Additional Experimental Results in FSCIL 20

A1 THEORETICAL ANALYSIS

We expand the analysis in § 4.2 by deriving the optimal fusion coefficient and its improvement over either anchor, followed by the corresponding class-wise formulation.

A1.1 ENDPOINT ERROR AND CONDITIONAL VELOCITY ERROR

Let $\boldsymbol{\mu}_c \in \mathbb{S}^{d-1}$ be the ideal center of class c, and let T_c and V_c be its normalized textual and visual anchors. Following the main paper, we view the textual and support-derived visual prototypes as *biased estimates* of the ideal class center, and hence estimate their errors by

$$\boldsymbol{e}_c^T \approx T_c - \boldsymbol{\mu}_c, \qquad \boldsymbol{e}_c^V \approx V_c - \boldsymbol{\mu}_c. \tag{A1}$$

For a source feature $\boldsymbol{z}$, the conditional target velocity toward anchor A_c^M is $u_M = A_c^M - \boldsymbol{z}$, where $M \in \{T, V\}$. Its deviation from the ideal target velocity $u_\star = \boldsymbol{\mu}_c - \boldsymbol{z}$ satisfies

$$u_M - u_\star = A_c^M - \boldsymbol{\mu}_c = \boldsymbol{e}_c^M, \qquad \|u_M - u_\star\|_2^2 = \|\boldsymbol{e}_c^M\|_2^2. \tag{A2}$$

Thus, the source feature cancels, and the quality of the transport target is determined by the anchor error. Textual anchors capture semantic information from the VLM, while visual anchors capture the appearance of the sampled support examples. These two sources of information can produce complementary estimation errors.

For a shared coefficient $\alpha \in [0, 1]$, combining the conditional target velocities gives

$$\begin{aligned} u_\alpha &= (1-\alpha)(T_c - \boldsymbol{z}) + \alpha(V_c - \boldsymbol{z}) \\ &= \widetilde{A}_c(\alpha) - \boldsymbol{z}, \qquad \widetilde{A}_c(\alpha) = (1-\alpha)T_c + \alpha V_c. \end{aligned} \tag{A3}$$

The corresponding error is $(1-\alpha)\boldsymbol{e}_c^T + \alpha \boldsymbol{e}_c^V$. We therefore study the expected squared error of this unnormalized fused endpoint.

A1.2 OPTIMAL SHARED FUSION COEFFICIENT

Let $\mathbb{E}$ average uniformly over classes and over support sampling. Define

$$\mathcal{R}^T = \mathbb{E}\|\boldsymbol{e}_c^T\|_2^2, \qquad \mathcal{R}^V = \mathbb{E}\|\boldsymbol{e}_c^V\|_2^2, \qquad C = \mathbb{E}[(\boldsymbol{e}_c^T)^\top \boldsymbol{e}_c^V]. \tag{A4}$$

Expanding the squared norm yields

$$\begin{aligned} \mathcal{R}(\alpha) &= \mathbb{E}\big\|(1-\alpha)\boldsymbol{e}_c^T + \alpha \boldsymbol{e}_c^V\big\|_2^2 \\ &= (1-\alpha)^2 \mathcal{R}^T + \alpha^2 \mathcal{R}^V + 2\alpha(1-\alpha)C \\ &= \mathcal{R}^T - 2\alpha(\mathcal{R}^T - C) + \alpha^2 D, \end{aligned} \tag{A5}$$

where

$$D = \mathcal{R}^T + \mathcal{R}^V - 2C = \mathbb{E}\|\boldsymbol{e}_c^T - \boldsymbol{e}_c^V\|_2^2 = \mathbb{E}\|T_c - V_c\|_2^2 \geq 0. \tag{A6}$$

The cross-moment C describes how closely the two errors align. For fixed individual risks, a smaller C reduces the risk of every interior mixture through the term $2\alpha(1-\alpha)C$.

Differentiating Eq. A5 gives

$$\mathcal{R}'(\alpha) = -2(\mathcal{R}^T - C) + 2\alpha D, \qquad \mathcal{R}''(\alpha) = 2D \geq 0. \tag{A7}$$

Hence the risk is convex. For $D > 0$, its unconstrained minimizer is $\bar{\alpha} = (\mathcal{R}^T - C)/D$, and completing the square gives

$$\mathcal{R}(\alpha) = \mathcal{R}^T - \frac{(\mathcal{R}^T - C)^2}{D} + D(\alpha - \bar{\alpha})^2. \tag{A8}$$

Minimizing over $[0, 1]$ therefore gives

$$\alpha^\star = \mathrm{clip}_{[0,1]}\left(\frac{\mathcal{R}^T - C}{\mathcal{R}^T + \mathcal{R}^V - 2C}\right). \tag{A9}$$

When $D = 0$, the two errors coincide almost surely and the risk is constant. Since both single-anchor choices belong to the feasible interval, the optimal fusion always satisfies

$$\mathcal{R}(\alpha^\star) \leq \min\{\mathcal{R}(0), \mathcal{R}(1)\} = \min\{\mathcal{R}^T, \mathcal{R}^V\}. \tag{A10}$$

A1.3 Strict Improvement from Complementary Errors

The optimum lies in $(0, 1)$ exactly when

$$C < \mathcal{R}^T \quad \text{and} \quad C < \mathcal{R}^V. \tag{A11}$$

Under this condition, substituting $\alpha^\star = \bar{\alpha}$ into Eq. A8 yields

$$\mathcal{R}(\alpha^\star) = \frac{\mathcal{R}^T\mathcal{R}^V - C^2}{D}. \tag{A12}$$

The reductions relative to the individual anchors are explicitly

$$\mathcal{R}^T - \mathcal{R}(\alpha^\star) = \frac{(\mathcal{R}^T - C)^2}{D} > 0, \qquad \mathcal{R}^V - \mathcal{R}(\alpha^\star) = \frac{(\mathcal{R}^V - C)^2}{D} > 0. \tag{A13}$$

These expressions quantify the benefit of complementary anchor errors: fusion improves on both anchors when their shared error is smaller than either individual risk.

When the cross-moment is zero, the coefficient and optimal risk simplify to

$$\alpha^\star = \frac{\mathcal{R}^T}{\mathcal{R}^T + \mathcal{R}^V}, \qquad \mathcal{R}(\alpha^\star) = \frac{\mathcal{R}^T\mathcal{R}^V}{\mathcal{R}^T + \mathcal{R}^V}. \tag{A14}$$

A larger textual risk therefore assigns more weight to the visual anchor, and vice versa. For equal positive risks $\mathcal{R}^T = \mathcal{R}^V = r$, equal fusion gives risk $r/2$. More generally, Eq. A8 shows that, for an interior optimum, using a coefficient α incurs excess risk $D(\alpha - \alpha^\star)^2$.

A1.4 Class-wise Derivation

For a fixed class c, let $\mathbb{E}_c$ denote expectation over support sampling and define $\mathcal{R}_c^M = \mathbb{E}_c\|\boldsymbol{e}_c^M\|_2^2$ and $C_c = \mathbb{E}_c[(\boldsymbol{e}_c^T)^\top \boldsymbol{e}_c^V]$. The same expansion gives

$$\mathcal{R}_c(\alpha) = (1-\alpha)^2\mathcal{R}_c^T + \alpha^2\mathcal{R}_c^V + 2\alpha(1-\alpha)C_c. \tag{A15}$$

Writing $D_c = \mathcal{R}_c^T + \mathcal{R}_c^V - 2C_c$, the class-wise optimum for $D_c > 0$ is

$$\alpha_c^\star = \mathrm{clip}_{[0,1]}\left(\frac{\mathcal{R}_c^T - C_c}{D_c}\right). \tag{A16}$$

It satisfies $\mathcal{R}_c(\alpha_c^\star) \le \min\{\mathcal{R}_c^T, \mathcal{R}_c^V\}$, with strict improvement when $C_c < \min\{\mathcal{R}_c^T, \mathcal{R}_c^V\}$. Averaging the class-wise risks recovers the shared-coefficient objective:

$$\mathcal{R}(\alpha) = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\mathcal{R}_c(\alpha), \qquad (\mathcal{R}^T, \mathcal{R}^V, C) = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}(\mathcal{R}_c^T, \mathcal{R}_c^V, C_c). \tag{A17}$$

Thus, the shared optimum is obtained from the aggregate risks and cross-moment. This relative-risk structure motivates the support-calibrated coefficient in Sec. 4.3, whose computation is detailed in Sec. A2.2.

A2 Reproducibility and Implementation Details

A2.1 Dataset Statistics

Few-shot learning. We evaluate on eleven datasets covering diverse recognition tasks. FGVC-Aircraft (Maji et al., 2013) distinguishes aircraft variants, while Stanford Cars (Krause et al., 2013) focuses on fine-grained car categories. Oxford Pets (Parkhi et al., 2012) contains cat and dog breeds, and Flowers102 (Nilsback & Zisserman, 2008) contains flower categories. Food101 (Bossard et al., 2014) covers different types of food. EuroSAT (Helber et al., 2019) provides satellite imagery for land-use and land-cover classification, and DTD (Cimpoi et al., 2014) groups textures by visual attributes. SUN397 (Xiao et al., 2010) covers indoor and outdoor scenes. UCF101 (Soomro et al., 2012) is a human-action video dataset, from which the image-based evaluation uses extracted frames. Caltech101 (Fei-Fei et al., 2004) contains common object categories, while ImageNet (Deng et al., 2009) provides large-scale recognition across a broad range of objects and natural categories.

We use the standard CLIP-adaptation splits and sample $K \in \{1, 4, 16\}$ labeled training images per class, with evaluation on the corresponding test split.

Few-shot class-incremental learning. We use CIFAR100 (Krizhevsky, 2009), miniImageNet (Ravi & Larochelle, 2017), and CUB200 (Wah et al., 2011). CIFAR100 contains low-resolution natural images from 100 object categories. miniImageNet is a 100-class subset of ImageNet used for few-shot recognition. For both datasets, 60 classes form the base session and the remaining 40 classes are introduced over eight incremental sessions. Each incremental session follows a 5-way 5-shot setting, with five new classes and five labeled examples per class. CUB200 contains 200 bird species for fine-grained recognition. Its base session contains 100 classes, followed by ten incremental sessions, each introducing ten new classes with five labeled examples per class, i.e., a 10-way 5-shot setting. Base sessions use the available training examples of their assigned classes. After each session, evaluation covers all classes observed so far, giving nine sessions for CIFAR100 and miniImageNet and eleven for CUB200.

A2.2 Details in Support-Calibrated Dual-Anchor Fusion

Here, we provide additional details on the Support-Calibrated Dual-Anchor Fusion described in § 4.3. We compute a single visual-branch weight from the labeled support set. Let z_i^c be a normalized support feature and K_c the number of support examples in class c. The visual anchor is $V_c = \mathrm{norm}(K_c^{-1} \sum_i z_i^c)$, where $\mathrm{norm}(x) = x/\|x\|_2$, and T_c is the normalized textual anchor. We measure each modality's reliability through its negative one-vs-rest margins, using $[a]_+ = \max(a, 0)$.

Text Branch. For textual anchors, the margin compares the correct class with its strongest competitor:

$$m_{c,i}^T = \langle z_i^c, T_c \rangle - \max_{k \neq c} \langle z_i^c, T_k \rangle, \qquad \widehat{\mathcal{R}}_{\downarrow,c}^T = \frac{1}{K_c} \sum_{i=1}^{K_c} [-m_{c,i}^T]_+. \tag{A18}$$

A negative margin contributes in proportion to the amount by which a competing anchor outranks the correct anchor.

Visual Branch. For visual anchors with $K_c > 1$, we exclude the evaluated example from its own prototype:

$$V_{c,-i} = \mathrm{norm}\left(\frac{1}{K_c - 1} \sum_{j \neq i} z_j^c \right), \qquad m_{c,i}^{V,\mathrm{loo}} = \langle z_i^c, V_{c,-i} \rangle - \max_{k \neq c} \langle z_i^c, V_k \rangle. \tag{A19}$$

The resulting risk is

$$\widehat{\mathcal{R}}_{\downarrow,c}^V = \frac{1}{K_c} \sum_{i=1}^{K_c} [-m_{c,i}^{V,\mathrm{loo}}]_+. \tag{A20}$$

This leave-one-out computation removes the direct self-contribution that would otherwise inflate the correct-class similarity. Competing prototypes retain all of their support examples.

For $K_c = 1$, we extract $N_a = 8$ stochastic views of the support image, including the view used to construct V_c. Let their normalized features be $\{z_{c,r}^{\mathrm{aug}}\}_{r=1}^{N_a}$. We define the observed uncertainty radius and pairwise anchor separation as

$$\eta_c = \max_r \|z_{c,r}^{\mathrm{aug}} - V_c\|_2, \qquad d_{ck} = [1 - \langle V_c, V_k \rangle]_+. \tag{A21}$$

To obtain a margin estimate, consider a feature $z = V_c + \delta$ with $\|\delta\|_2 \leq \eta_c$. Since the anchors are normalized, $\|V_c - V_k\|_2 = \sqrt{2d_{ck}}$, and the Cauchy–Schwarz inequality gives

$$\begin{aligned} \langle z, V_c \rangle - \langle z, V_k \rangle &= d_{ck} + \langle \delta, V_c - V_k \rangle \\ &\geq d_{ck} - \eta_c \sqrt{2d_{ck}}. \end{aligned} \tag{A22}$$

Taking the least favorable competing class yields

$$\underline{m}_c^V = \min_{k \neq c} \{ d_{ck} - \eta_c \sqrt{2d_{ck}} \}, \qquad \widehat{\mathcal{R}}_{\downarrow,c}^V = [-\underline{m}_c^V]_+. \tag{A23}$$

A singleton visual anchor therefore receives higher risk when its augmentation-induced variation is large relative to its separation from nearby classes.

Finally, we average the risks uniformly across classes and compute

$$\widehat{\mathcal{R}}^{M}_{\downarrow} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \widehat{\mathcal{R}}^{M}_{\downarrow,c}, \qquad \alpha_{\text{sup}} = \frac{\widehat{\mathcal{R}}^{T}_{\downarrow}}{\widehat{\mathcal{R}}^{T}_{\downarrow} + \widehat{\mathcal{R}}^{V}_{\downarrow}}. \tag{A24}$$

If the summed risk is numerically zero, we set $\alpha_{\text{sup}} = 0.5$. Macro averaging gives each class equal weight even when support counts differ. We compute this coefficient once and use it for both velocity composition and anchor fusion. In FSCIL, it is estimated from the base-session support data and reused throughout the incremental sessions.

A2.3 Additional Implementation Details

Similar to FMA (Jiang et al., 2026), we adopt the lightweight architecture from MAR (Li et al., 2024) as our flow-matching network, implemented as a deep residual MLP with timestep conditioning. The network contains a shared input projection and 12 residual blocks, followed by separate textual and visual prediction heads. Its hidden width matches the CLIP embedding dimension, which is 512 for ViT-B/16. Each residual block uses LayerNorm, two linear layers with a SiLU activation, and timestep-conditioned shift, scale, and residual gating. Timesteps are scaled by 1000, encoded with a 256-dimensional sinusoidal embedding, and passed through a two-layer MLP. Each output head applies a timestep-modulated normalization and a linear projection back to the feature dimension. The residual modulation layers and output projections are initialized to zero.

The default flow-training configuration uses AdamW with a learning rate of 2×10^{-4}, weight decay of 0.01, and batch size of 32 for 200 epochs. We use one epoch of linear warmup followed by cosine learning-rate decay. The textual and visual flow-matching losses have equal weights, with independent timesteps sampled uniformly from $[0, 1]$ for the two branches. Training uses automatic mixed precision. The VLM feature extractor remains frozen during flow training; for CLIP-LoRA and CLIP-Adapter, the flow operates on the corresponding adapted embeddings.

Training images undergo random resized cropping to 224×224 with crop scale in $[0.08, 1]$, followed by horizontal flipping with probability 0.5 and CLIP normalization. The same stochastic transform provides the views for one-shot uncertainty estimation. Test images are resized and center-cropped to 224×224. Support sampling uses the configured random seed, which defaults to 1.

Table A1: **Stability on difficult datasets across five random seeds.** Results are mean test accuracy (%) $\pm$ standard deviation. AVG denotes the average over the five difficult datasets.

		Difficult Datasets					
Shot	Backbone	Aircraft	SAT	DTD	SUN	Cars	**AVG**
1	CLIP	27.98 ± 0.33	65.45 ± 1.07	54.20 ± 0.48	67.00 ± 0.21	67.02 ± 0.52	56.33 ± 0.16
	CLIP-LoRA	30.40 ± 0.34	81.18 ± 1.00	57.14 ± 0.58	71.02 ± 0.16	70.51 ± 0.55	62.05 ± 0.31
4	CLIP	34.89 ± 0.53	76.73 ± 0.83	63.86 ± 0.23	72.01 ± 0.07	72.39 ± 0.50	63.98 ± 0.14
	CLIP-LoRA	42.02 ± 0.20	89.49 ± 0.26	69.32 ± 0.11	74.48 ± 0.11	80.04 ± 0.28	71.07 ± 0.14
16	CLIP	46.92 ± 0.98	86.88 ± 0.63	73.10 ± 0.49	75.79 ± 0.09	82.48 ± 0.15	73.03 ± 0.27
	CLIP-LoRA	59.24 ± 0.20	93.14 ± 0.09	76.91 ± 0.20	77.83 ± 0.08	88.80 ± 0.10	79.19 ± 0.07
	CLIP-Adapter	46.19 ± 0.31	86.48 ± 0.73	72.73 ± 0.40	76.46 ± 0.09	82.86 ± 0.28	72.94 ± 0.18

A3 More Experimental Results

A3.1 Stability Analysis

We assess the robustness of DualFlow to few-shot sampling by repeating each configuration with five random seeds. Tabs A1 and A2 report the mean test accuracy and standard deviation across these runs on the difficult and easy dataset groups, respectively. Specifically, we evaluate DualFlow on the feature spaces provided by CLIP, CLIP-LoRA, and CLIP-Adapter. CLIP and CLIP-LoRA

Table A2: **Stability on easy datasets across five random seeds.** Results are mean test accuracy (%) ± standard deviation. AVG denotes the average over the six easy datasets.

		Easy Datasets						
Shot	Backbone	Pets	UCF	Flowers	Caltech	Food	Net	**AVG**
1	CLIP	89.83 ± 0.31	72.44 ± 0.54	81.32 ± 0.77	93.97 ± 0.35	85.62 ± 0.14	67.48 ± 0.12	81.78 ± 0.12
	CLIP-LoRA	92.82 ± 0.23	75.78 ± 0.80	88.96 ± 0.25	94.30 ± 0.26	85.03 ± 0.23	69.95 ± 0.11	84.47 ± 0.19
4	CLIP	92.39 ± 0.13	79.04 ± 0.34	94.70 ± 0.08	95.55 ± 0.22	86.55 ± 0.04	69.59 ± 0.04	86.30 ± 0.09
	CLIP-LoRA	91.81 ± 0.17	83.20 ± 0.27	96.39 ± 0.24	95.82 ± 0.07	84.13 ± 0.08	72.37 ± 0.08	87.29 ± 0.06
16	CLIP	92.96 ± 0.29	84.40 ± 0.18	98.12 ± 0.21	96.04 ± 0.20	87.19 ± 0.06	72.13 ± 0.05	88.47 ± 0.07
	CLIP-LoRA	92.69 ± 0.11	87.91 ± 0.22	99.00 ± 0.11	96.50 ± 0.11	86.09 ± 0.06	74.59 ± 0.06	89.46 ± 0.03
	CLIP-Adapter	92.84 ± 0.25	84.78 ± 0.25	98.19 ± 0.24	96.03 ± 0.09	87.35 ± 0.07	72.93 ± 0.11	88.69 ± 0.04

are evaluated under all three shot settings, while CLIP-Adapter is evaluated under the 16-shot setting. Across these configurations, the standard deviations of the group averages remain below 0.32 percentage points, and all per-dataset standard deviations are at most 1.07 points. These small variations indicate that DualFlow remains broadly stable across support-set sampling, feature extractors, and shot regimes.

Table A3: **Ablation study** on Double Branch (**D**), Text Endpoint (**T**), Visual Endpoint (**V**), and Fusion Weighting (**W**) under the 1-, 4-, and 16-shot settings based on CLIP.

Shot	Method	D	T	V	W	Aircraft	SAT	DTD	SUN	Cars	**AVG**	Pets	UCF	Flowers	Caltech	Food	Net	**AVG**
0	CLIP	×	×	×	×	24.8	47.8	43.8	62.6	65.5	48.9	89.1	66.8	71.2	92.9	86.1	66.7	78.8
1	(1)	×	✓	×	×	26.9	66.4	53.4	66.5	67.1	56.1	**90.1**	72.6	80.6	94.2	**86.1**	**68.0**	81.9
	(2)	×	×	✓	×	19.6	55.8	35.8	38.8	27.1	35.4	40.6	50.3	67.9	76.2	43.8	29.3	51.4
	(3)	×	✓	✓	0.5	24.7	60.4	44.7	53.9	51.0	46.9	78.0	64.2	82.2	84.9	69.4	50.2	71.5
	(4)	✓	✓	✓	0.5	25.3	60.4	44.8	54.1	51.2	47.2	78.3	63.8	**82.6**	87.1	70.0	50.7	72.1
	(5)	×	×	×	Adaptive	26.7	58.8	50.2	65.0	66.1	53.4	89.0	69.8	73.9	93.6	86.0	67.6	80.0
	(6)	×	✓	✓	Adaptive	27.7	**66.6**	53.6	**67.0**	66.9	56.4	90.0	**72.9**	81.7	94.2	86.0	**68.0**	**82.1**
	Ours	✓	✓	✓	Adaptive	**28.3**	66.5	**53.8**	**67.0**	**67.4**	**56.6**	90.0	**72.9**	82.2	**94.4**	85.4	67.4	**82.1**
4	(1)	×	✓	×	×	31.1	76.9	60.7	70.1	71.2	62.0	92.3	77.7	92.3	95.5	86.4	68.9	85.5
	(2)	×	×	✓	×	32.2	71.1	58.2	63.9	59.1	56.9	76.6	72.1	92.1	91.3	72.1	52.6	76.1
	(3)	×	✓	✓	0.5	35.0	75.3	62.5	69.8	70.7	62.7	90.4	78.2	94.2	94.5	82.9	64.3	84.1
	(4)	✓	✓	✓	0.5	**35.8**	77.1	61.9	69.9	71.8	63.3	90.1	78.4	94.6	95.0	82.7	65.0	84.3
	(5)	×	×	×	Adaptive	29.9	68.4	58.8	70.3	69.8	59.4	91.2	74.2	90.5	94.8	86.2	68.7	84.3
	(6)	×	✓	✓	Adaptive	33.5	77.0	**63.9**	71.8	72.0	63.6	92.3	79.0	94.3	**95.8**	86.1	69.3	86.1
	Ours	✓	✓	✓	Adaptive	35.2	**77.2**	**63.9**	**72.0**	**73.1**	**64.3**	**92.6**	**79.1**	**94.8**	**95.8**	**86.6**	**69.7**	**86.4**
16	(1)	×	✓	×	×	38.1	84.3	69.5	73.5	78.6	68.8	**92.8**	82.0	97.5	**95.9**	87.0	70.4	87.6
	(2)	×	×	✓	×	45.4	84.2	70.2	73.9	80.9	70.9	87.6	81.4	97.7	95.0	82.5	66.7	85.2
	(3)	×	✓	✓	0.5	45.0	85.6	71.9	75.3	**82.8**	72.1	91.7	84.5	**98.1**	**95.9**	86.0	71.2	87.9
	(4)	✓	✓	✓	0.5	**45.7**	86.2	**72.6**	**75.8**	**82.8**	72.6	92.2	84.4	**98.1**	95.7	86.4	71.6	88.1
	(5)	×	×	×	Adaptive	35.3	71.6	64.5	73.3	75.0	63.9	91.9	79.0	95.7	95.4	86.5	70.2	86.5
	(6)	×	✓	✓	Adaptive	45.5	85.7	71.9	75.6	82.3	72.2	92.5	**84.6**	97.6	**95.9**	86.7	**72.1**	88.2
	Ours	✓	✓	✓	Adaptive	**45.7**	**86.9**	**72.6**	**75.8**	82.4	**72.7**	92.5	84.5	**98.1**	**95.9**	**87.2**	**72.1**	**88.4**

A3.2 Additional Ablation Analysis in Few-Shot Adaptation

We provide a detailed component analysis in Tab. A3. We conduct ablations to quantify the contributions of the key designs in DualFlow, including the double-branch architecture (D), textual endpoint (T), visual endpoint (V), and fusion weighting (W). As shown in Tab. A3, we evaluate different variants under all shot settings. First, Rows (1) and (2) employ a single flow branch with only the textual or visual endpoint, respectively. Their relative performance varies substantially across datasets and shot settings, indicating that the two endpoints provide complementary but differently reliable supervision. In particular, the textual endpoint is generally stronger in the low-shot regime, while the visual endpoint becomes increasingly competitive as more support samples are available. Row (3) further combines the two sources within a *single branch*: the textual and visual endpoints are first blended with a fixed weight of 0.5, and the flow is then learned toward this blended endpoint. Although this improves over the weaker single-endpoint variant in many cases, it does not consistently outperform the stronger endpoint, especially under 1-shot, suggesting that naively averaging the two endpoints may dilute the more reliable supervision. Row (4) instead adopts two separate flow branches for the textual and visual endpoints, while still using a fixed 0.5 weight to fuse their predictions. Compared with Row (3), this variant consistently improves or maintains performance across shot settings, showing that separately modeling the two endpoint-induced flows

is preferable to forcing them into a single flow toward a blended endpoint. To isolate the effect of adaptive weighting, Row (5) removes flow learning entirely and only uses the support set to estimate the adaptive fusion weight, which is then used to directly combine the textual and visual anchors. Its performance shows that support-conditioned weighting itself provides useful calibration, but cannot account for the gains of DualFlow without explicitly learning the transport dynamics. Row (6) reintroduces flow learning but remains a single-branch design: the textual and visual endpoints are first blended using the adaptively estimated weight, and one flow is trained toward the resulting endpoint. This substantially improves over the corresponding fixed-weight single-branch variant in Row (3), demonstrating the importance of adapting the endpoint composition according to support-set reliability. Finally, our full model combines both ingredients: separate flows are learned toward the textual and visual endpoints, and their contributions are adaptively fused using the support-conditioned reliability weights. Its consistently strong performance across 1-, 4-, and 16-shot settings demonstrates that the gains arise from the combination of endpoint-specific flow modeling and adaptive fusion, rather than either component alone.

A3.3 Additional Experimental Results in FSCIL

Fig. A1 compares DualFlow with FMA across all FSCIL sessions on CIFAR100, CUB200, and miniImageNet. Besides the conventional overall accuracy, we further evaluate the performance on incremental classes following Zhao et al. (2025). In FSCIL, overall accuracy is evaluated over all classes observed so far, including the base classes. Since the base session typically contains substantially more classes and abundant training samples, the resulting overall accuracy can be dominated by the performance on base classes and may therefore obscure how well the model adapts to newly introduced categories. To provide a more direct evaluation of incremental learning ability, we additionally report the cumulative incremental-class accuracy at session t:

$$\mathrm{IncAcc}_t = \frac{\sum_{(\mathbf{x},y)\in\mathcal{D}_{\mathrm{test}}^{1:t}} \mathbb{I}[\hat{y}(\mathbf{x}) = y]}{|\mathcal{D}_{\mathrm{test}}^{1:t}|}, \tag{A25}$$

where $\mathcal{D}_{\mathrm{test}}^{1:t}$ contains the test samples from all incremental classes introduced from Sessions 1 to t, excluding the base classes in Session 0.

As shown in Fig. A1, DualFlow exhibits a clear advantage over FMA in incremental-class recognition. On CIFAR100, DualFlow improves the average incremental-class accuracy from 63.63% to 69.38%. The improvement becomes more pronounced on CUB200, increasing from 51.05% to 63.71%. On miniImageNet, although FMA performs slightly better in the first incremental session, its incremental accuracy rapidly degrades as more classes are introduced, whereas DualFlow remains consistently around 73–81% in the subsequent sessions, resulting in an average improvement of 37.12 percentage points. Meanwhile, DualFlow also consistently achieves higher overall accuracy, indicating that the gain on incremental categories does not come at the expense of previously learned knowledge. These results further support our main design motivation: by exploiting support-conditioned dual-anchor information, DualFlow provides a more reliable adaptation target for few-shot novel classes, enabling effective acquisition of new-category knowledge while preserving the transferable knowledge learned from previous sessions.

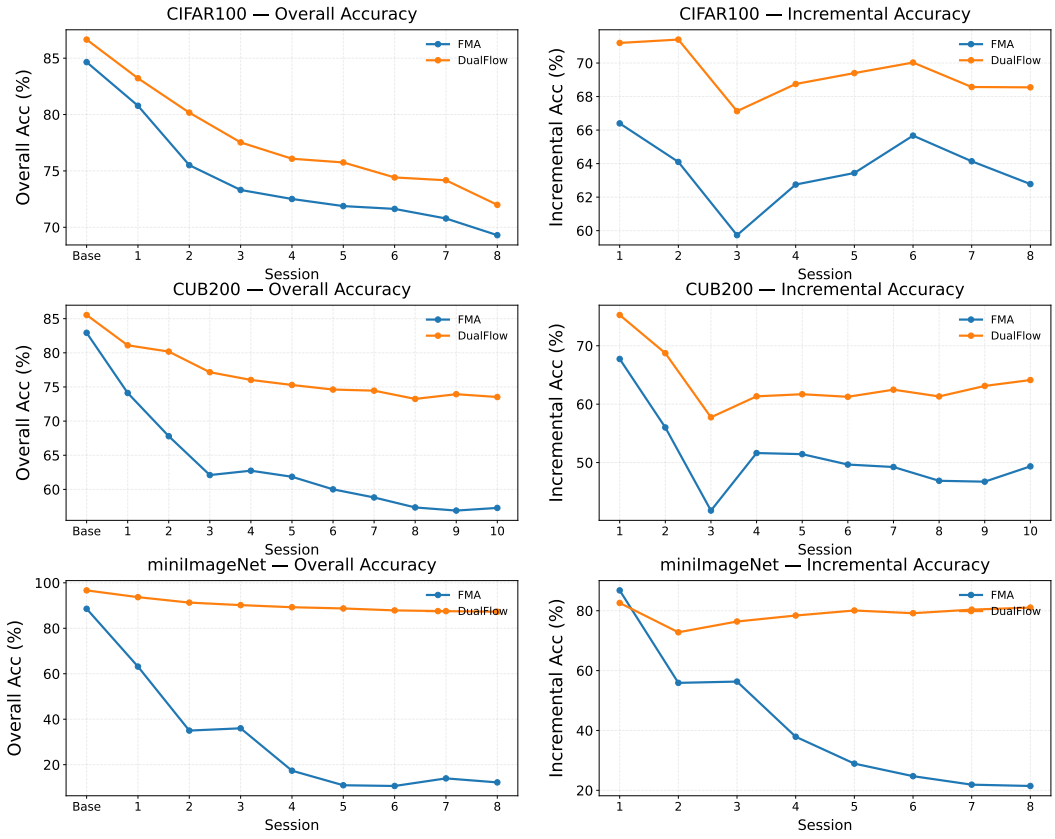


Figure A1: Comparison between FMA and DualFlow on CIFAR100, CUB200, and miniImageNet throughout the FSCIL sessions.