Direct Product Flow Matching: Decoupling Radial and Angular Dynamics for Few-Shot Adaptation

Hongxu Chen[1], Yanghao Wang[1], Bowei Zhu[2], Hongxiang Li[1], Zhen Wang[1], Ziqi Jiang[1], Lin Li[1], Rui Liu[3], Long Chen[1]*
[1] HKUST, [2] USTC, [3] Huawei Research

Abstract

Recent flow matching (FM) methods improve the few-shot adaptation of vision-language models, by modeling cross-modal alignment as a continuous multi-step flow. In this paper, we argue that existing FM methods are inherently constrained by incompatible geometric priors on pre-trained cross-modal features, resulting in suboptimal adaptation performance. We first analyze these methods from a polar decomposition perspective (*i.e.*, radial and angular sub-manifolds). Under this new geometric view, we identify three overlooked limitations in them: *1) Angular dynamics distortion*: The radial-angular coupling induces non-uniform speed on the angular sub-manifold, leading to regression training difficulty and extra truncation errors. *2) Radial dynamics neglect*: Feature normalization discards modality confidence, failing to distinguish out-of-distribution and in-distribution data, and abandoning crucial radial dynamics. *3) Context-agnostic unconditional flow*: Dataset-specific information loss during pre-trained cross-modal feature extraction remains unrecovered. To resolve these issues, we propose **warped product flow matching (WP-FM)**, a unified Riemannian framework that reformulates alignment on a warped product manifold. Within this framework, we derive **direct product flow matching (DP-FM)** by introducing a constant-warping metric, which yields a decoupled cylindrical manifold (*i.e.*, direct product manifold). DP-FM enables independent radial evolution and constant-speed angular geodesic transport, effectively eliminating angular dynamics distortion while preserving radial consistency. Meanwhile, we incorporate classifier-free guidance by conditioning the flow on the pre-trained VLMs' hidden states to inject missing dataset-specific information. Extensive results across 11 benchmarks have demonstrated that DP-FM achieves a new state-of-the-art for multi-step few-shot adaptation.

1 Introduction

The success of vision-language models (VLMs) [1, 2, 3, 4, 5], such as CLIP [4], fundamentally relies on cross-modal alignment. This capability embeds semantically corresponding visual and textual signals into a shared, expressive latent space, where their alignment is primarily characterized by *angular similarity*. This precise alignment serves as the core mechanism enabling robust generalization and semantic understanding across diverse downstream tasks. However, when transferring these pre-trained VLMs to specific target domains with limited data, the pre-trained alignment often suffers from domain shift, requiring domain-specific fine-tuning (*i.e.*, few-shot adaptation).

Traditionally, such adaptation relies on parameter-efficient fine-tuning (PEFT) methods [6, 7, 8, 9, 10], which can enhance the cross-modal alignment capability of the pretrained VLMs. However, PEFT methods typically perform alignment in a single-step adjustment, where cross-modal features align utilizing only one forward pass of the PEFT module (*cf.*, Figure 1(a)), lacking an explicit mechanism to model *continuous* and *multi-step* alignment dynamics. As a result, they often struggle

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

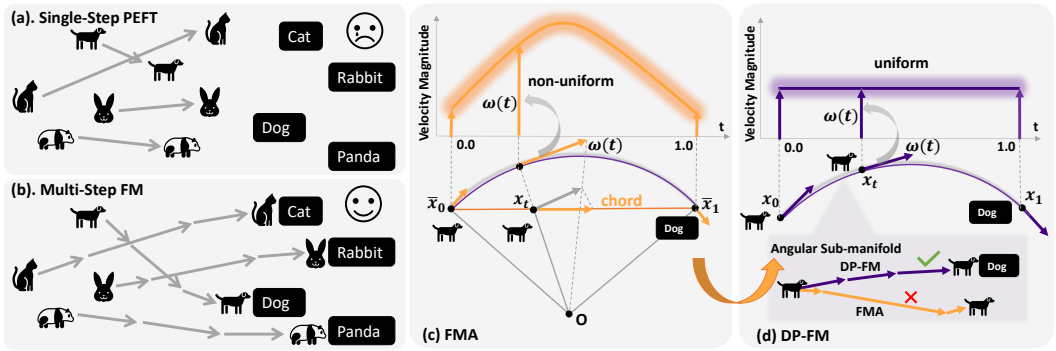

Figure 1: **(a).** Single-step parameter-efficient fine-tuning (PEFT) mostly performs cross-modal alignment in a single-step manner. **(b).** Multi-step flow matching (FM) methods model continuous and multi-step alignment dynamics. During the training stage, **(c).** FMA undergoes a non-uniform angular speed induced by radial–angular coupling. However, **(d).** DP-FM follows a constant-speed angular geodesic due to decoupled radial and angular dynamics.

to resolve the alignment of highly entangled cross-modal distributions (*i.e.*, image features and text features) in challenging datasets [11]. To address this limitation, flow matching for few-shot adaptation (FMA) [11] introduces a paradigm shift by conceptualizing cross-modal alignment as an *unconditional*[2] continuous probability flow.

Specifically, FMA first normalizes the pre-trained cross-modal features (*e.g.*, CLIP image and text features $\boldsymbol{x}_0, \boldsymbol{x}_1$) into unit-length pairs $(\bar{\boldsymbol{x}}_0, \bar{\boldsymbol{x}}_1)$, where $\bar{\boldsymbol{x}}_i = \boldsymbol{x}_i / \|\boldsymbol{x}_i\|_2$ for $i \in \{0, 1\}$. It then trains a velocity network to regress the target velocity $\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_0$ induced by the *linear interpolation* trajectory $\boldsymbol{x}_t = (1-t)\bar{\boldsymbol{x}}_0 + t\bar{\boldsymbol{x}}_1$ on a flat manifold (*i.e.*, the chord as shown in Figure 1(c)). By multi-step integration over the predicted velocity, FMA progressively transports the image features to the text features (Figure 1(b)), thereby achieving improved alignment compared to previous PEFT methods.

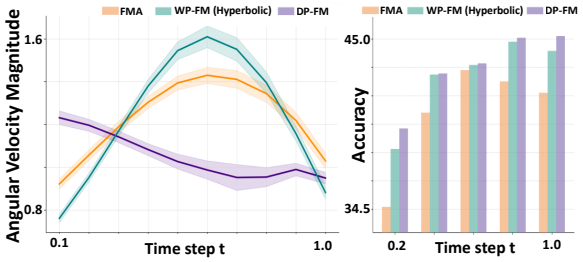

Figure 2: **Comparison between FMA [11], WP-FM (Hyperbolic) [12], and DP-FM** on Aircraft dataset at 100 epochs. DP-FM shows enhanced accuracy and more uniform angular speed across time step t at inference.

To analyze the inherent mechanics of cross-modal alignment, it is natural to apply a **polar decomposition** [13] to the interpolated features $\boldsymbol{x}_t$, disentangling them into a radial component ($r_t = \|\boldsymbol{x}_t\|_2$, *i.e.*, radial sub-manifold $\mathcal{M}_r$, reflecting modality confidence) and an angular component ($\boldsymbol{\theta}_t = \boldsymbol{x}_t / \|\boldsymbol{x}_t\|_2$, *i.e.*, angular sub-manifold $\mathcal{M}_{\boldsymbol{\theta}}$, encoding semantic direction). Under this geometric and probabilistic view, we argue that existing continuous flow matching (FM) methods [11, 12] for few-shot adaptation remain bottlenecked, failing to exploit the underlying capabilities of FM with suboptimal adaptation performance mostly due to **three main limitations**:

1) **Angular Dynamics Distortion.** They impose an incompatible geometric prior that distorts the intrinsic alignment dynamics. In FMA, the linear interpolation trajectory on a flat manifold induces severe *radial–angular coupling*. As shown in Figure 1(c), when the interpolated feature $\boldsymbol{x}_t$ traverses this chord, the speed $\boldsymbol{\omega}(t)$[3] on the angular sub-manifold (*i.e.*, purple arc) is projected from velocity $\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_0$. It undergoes angular dynamics distortion, characterized by *non-uniform* speed with non-zero acceleration. This induces extra curvature in the target velocity field, which increases the difficulty of regression training and leads to additional truncation errors, resulting in suboptimal performance during multi-step alignment. The following work HFM [12] implicitly bypasses this distortion by forcefully adjusting pre-trained features into a hyperbolic space [14] before applying FM. However, it introduces unnecessary geometric complexity and destroys the inherent angular sub-manifold inherent to these pre-trained features, leading to incompatibility with standard, off-the-shelf adaptation pipelines widely used in current cross-modal tasks.

[2] An *unconditional* flow indicates that the learned velocity field $v(\boldsymbol{x}_t, t)$ depends solely on the current state $\boldsymbol{x}_t$ and time t. It models the marginal probability path between two distributions without relying on any external conditions.

[3] We use *speed* as a scalar to denote the velocity magnitude, and *velocity* represents both magnitude and direction.

2) **Radial Dynamics Neglect.** Radial information intrinsically reflects modality confidence, acting as an explicit signal to distinguish out-of-distribution (OOD) and in-distribution (ID) data [15] (Refer to Appendix B). This confidence is helpful for the velocity network in few-shot adaptation, as it enables the alignment process to remain sensitive to uncertain OOD samples and apply focused corrections. However, due to the inherent radial–angular coupling in existing methods, explicitly retaining this radial information exacerbates the distortion in angular dynamics, leading to degraded performance.
3) **Context-Agnostic Unconditional Flow.** In few-shot adaptation scenarios, cross-modal features extracted by pre-trained VLMs are highly generalized and lack dataset-specific details for the target domain. Existing FM methods model cross-modal alignment as an unconditional probability flow, relying solely on these generalized features. The problem is that an unconditional flow purely aligns pair-wise features, making it inherently *context-agnostic*. Since it is blind to the whole dataset context, it cannot well adapt these generalized features into domain-specific ones. Therefore, there is no additional conditional guidance to introduce the target domain context into the alignment process, *i.e.*, the adaptation performance inevitably degrades.

Unified WP-FM Framework. To systematically resolve the geometric Limitation 1) and 2), we propose **warped product flow matching (WP-FM)** framework, which elevates the aforementioned polar decomposition into the Riemannian framework [16]. Rather than describing cross-modal alignment on a flat manifold, WP-FM structures the features as a warped product manifold $\mathcal{M} = \mathcal{M}_r \times_{g_\phi} \mathcal{M}_{\boldsymbol{\theta}}$. However, merely decoupling the features topologically does not resolve radial–angular coupling, as coupling behavior is dictated by the underlying Riemannian metric g_ϕ. To capture this, we equip $\mathcal{M}$ with a unified metric tensor $g_\phi : ds^2 = dr^2 + \phi(r)^2 d\boldsymbol{\theta}^2$. Here, ds, dr, and $d\boldsymbol{\theta}$ denote the infinitesimal distance elements on the full manifold $\mathcal{M}$, the radial sub-manifold $\mathcal{M}_r$, and the angular sub-manifold $\mathcal{M}_{\boldsymbol{\theta}}$, respectively, while $\phi(r)$ acts as a *radial warping function* that controls the coupling behavior. Under this general framework, we can define different metrics to instantiate different variants covering previous FM methods for adaptation:

- When $\phi(r) = r$, it degrades to FM on the flat manifold, as in FMA [11].
- When $\phi(r) = \sinh(r)$, it degrades to FM on the Hyperbolic manifold, as in HFM [12].

It is worth noting that these non-constant warping functions (in above-mentioned variants) intrinsically couple the radial and angular dynamics, which are "warped", directly causing the distortion discussed previously. To ensure cross-modal features naturally evolve along dual, independent geodesics $\gamma(t) = (\gamma_r(t), \gamma_{\boldsymbol{\theta}}(t))$ with a constant-speed angular geodesic, we introduce **direct product flow matching (DP-FM)** under the unified WP-FM framework: by deriving the metric for unraveling the radial–angular coupling, we acquire the constant warping function (*i.e.*, $\phi(r) = \text{CONSTANT}$), yielding FM on a decoupled cylindrical manifold (*i.e.*, direct product manifold). In DP-FM, both Limitation 1) and 2) are resolved: the radial trajectory $\gamma_r(t)$ preserves the dynamics of modality confidence, while the angular trajectory $\gamma_{\boldsymbol{\theta}}(t)$ independently traverses a *constant-speed* geodesic (*cf.*, Figure 1(d)), eliminating angular dynamics distortion for enhanced performance (*cf.*, Figure 2).

Furthermore, to resolve the context-agnostic Limitation 3) caused by information loss, DP-FM can incorporate classifier-free guidance (CFG) [17]. Instead of depending exclusively on the pre-trained generalized cross-modal features, we condition the flow on the pre-trained model's hidden states [18]. This explicitly injects the missing dataset-specific information into the alignment, allowing DP-FM to recover fine-grained details in the target domain and guarantee enhanced few-shot adaptation.

Extensive experiments across 11 diverse few-shot benchmarks demonstrate that DP-FM achieves enhanced adaptation accuracy compared to previous state-of-the-art FM methods for few-shot adaptation. In summary, our main contributions are threefold: **1). Theoretical Insight:** We identify fundamental geometric limitations in existing continuous FM methods for few-shot adaptation, namely angular dynamics distortion and radial dynamics neglect. **2). Unified WP-FM Framework:** We propose warped product FM to reformulate cross-modal alignment on a unified Riemannian manifold. **3). DP-FM with CFG:** We derive DP-FM to strictly decouple radial and angular dynamics. Additionally, we integrate CFG to inject dataset-specific information back into the alignment process.

2 Related Work

Few-shot Adaptation in VLMs. Few-shot adaptation of vision-language models (VLMs) [4, 19, 20, 21, 22] aims to transfer pre-trained cross-modal alignment to downstream domains using limited data. To achieve this, prior work predominantly adopts parameter-efficient fine-tuning (PEFT)

methods [23], which adapt models without updating the entire model. Among these approaches, prompt tuning methods [7, 6] optimize learnable context vectors, typically within the text encoder, while adapter-based methods [24, 25, 9, 26] introduce lightweight trainable modules into the model. These methods have demonstrated effective empirical performance in few-shot adaptation. Recent works [11, 12, 27, 28] explore continuous formulations based on flow matching, where cross-modal alignment is modeled as a multi-step unconditional probability flow. While they provide an expressive multi-step framework for cross-modal alignment, they fail to exploit the underlying capabilities of FM with suboptimal adaptation performance. As analyzed in Limitation **1)** and **2)**, these issues are closely related to the underlying geometric limitations inherent in existing methods.

Flow Matching (FM) and Riemannian FM. FM [29, 30] has emerged as an efficient framework for generative modeling [31, 32, 33, 34, 35, 36]. By directly regressing target velocity fields induced by simple and tractable probability paths, FM enables a streamlined training pipeline and improved numerical stability compared to prior approaches (*e.g.*, traditional diffusion models [37, 38]). To accommodate non-Euclidean data, Riemannian FM (RFM) [16] constructs optimal transport trajectories under the Riemannian metric. While theoretically principled, RFM lacks a task-specific metric design. Recent FM method for few-shot adaptation [11] defaults to incompatible geometries, ignoring the intrinsic radial and angular structure of pre-trained cross-modal features. To bridge this gap, our proposed warped product FM (WP-FM) tailors RFM for cross-modal alignment by introducing a unified metric with a radial warping function. Crucially, by adopting an optimal constant-warping metric, we derive direct product FM (DP-FM), which decouples radial and angular dynamics. This resolves Limitations **1)** and **2)**, establishing a robust geometric foundation for multi-step adaptation.

3 Methodology

3.1 Unified Warped Product Flow Matching (WP-FM) Framework

Preliminaries: Warped Product Manifolds. In differential geometry, a *warped product manifold* [39] provides a generalized structure for the product of two Riemannian manifolds. Given a base manifold $(\mathcal{B}, g_{\mathcal{B}})$ and a fiber manifold $(\mathcal{F}, g_{\mathcal{F}})$, their warped product manifold $\mathcal{M} = \mathcal{B} \times_f \mathcal{F}$ is constructed by equipping the product topological space $\mathcal{B} \times \mathcal{F}$ with a metric tensor characterized by the following line element: $ds^2 = ds^2_{\mathcal{B}} + f(x)^2 ds^2_{\mathcal{F}}$, where $f : \mathcal{B} \to \mathbb{R}_{>0}$ is a smooth scalar function called the *warping function*. Under this structure, the intrinsic geometry of the fiber manifold $\mathcal{F}$ is modified (*i.e.*, "warped") by $f(x)$, depending solely on the point $x \in \mathcal{B}$. This framework provides a foundation for pre-trained cross-modal feature representation, allowing us to map the modality confidence to the base manifold and the semantic direction to the fiber manifold.

Riemannian FM on Warped Product Manifolds. We begin by formalizing cross-modal alignment as a continuous probability flow on a Riemannian manifold. Given features $\boldsymbol{x} \in \mathbb{R}^d$ (*i.e.*, any possible interpolated features between cross-modal features) in the cross-modal alignment, we apply a polar decomposition to disentangle the feature: $\boldsymbol{x} = r \cdot \boldsymbol{\theta}, r = \|\boldsymbol{x}\|_2, \boldsymbol{\theta} \in \mathbb{S}^{d-1}$. This induces a warped product manifold $\mathcal{M} = \mathcal{M}_r \times_{g_\phi} \mathcal{M}_{\boldsymbol{\theta}}$, where $\mathcal{M}_r = \mathbb{R}_{>0}$ can be interpreted as reflecting modality confidence (*i.e.*, base radial sub-manifold), and $\mathcal{M}_{\boldsymbol{\theta}} = \mathbb{S}^{d-1}$ encodes the semantic direction (*i.e.*, fiber angular sub-manifold). To characterize the dynamics and radial-angular coupling behavior, we establish a unified Riemannian manifold $(\mathcal{M}, g_\phi)$, where the metric tensor g_ϕ is defined by a family of rotationally symmetric line elements: $ds^2 = dr^2 + \phi(r)^2 d\boldsymbol{\theta}^2$, where ds, dr, and $d\boldsymbol{\theta}$ denote the infinitesimal distance elements on the full manifold $\mathcal{M}$, the radial sub-manifold $\mathcal{M}_r$, and the angular sub-manifold $\mathcal{M}_{\boldsymbol{\theta}}$, respectively, $\phi(r)$ is a radial warping function.

Warped Product Flow Matching. This formulation provides a unified geometric framework that subsumes a broad class of FM methods based on different choices of g_ϕ (*e.g.*, $\phi(r) = r$ degrades to the flat manifold in FMA [11], $\phi(r) = \sinh(r)$ degrades to the Hyperbolic manifold in HFM [12]). We define Riemannian FM on this warped product manifold as **warped product flow matching (WP-FM)**. Given paired pretrained cross-modal features $(\boldsymbol{x}_0, \boldsymbol{x}_1)$ (*i.e.*, $\boldsymbol{x}_0$ is image feature, and $\boldsymbol{x}_1$ is text feature), the optimal transport trajectory corresponds to a geodesic $\gamma(t)$ under the chosen metric g_ϕ. The target velocity fields regressed by the network are then defined as $\boldsymbol{v}^*(\boldsymbol{x}_t, t) = \frac{d}{dt}\gamma(t)$.

3.2 Analyses: Angular Dynamics Distortion

FMA [11] acts as a suboptimal instantiation of WP-FM that explicitly discards radial dynamics via L_2 normalization. By implicitly setting the warping function to $\phi(r) = r$, FMA degrades the metric to a flat space. Under this incompatible geometric prior, the optimal transport geodesic defaults

to a linear interpolation: $\gamma(t) = (1-t)\bar{\boldsymbol{x}}_0 + t\bar{\boldsymbol{x}}_1$, where $\bar{\boldsymbol{x}}_i = \boldsymbol{x}_i / \|\boldsymbol{x}_i\|_2$. While this trajectory maintains a constant velocity in the ambient space, it inherently introduces *radial–angular coupling* and induces severe *angular dynamics distortion*. Specifically, let $\gamma_r(t) = \|\gamma(t)\|_2$ denote the radial magnitude and $\omega(t) = \|\dot{\gamma}_{\boldsymbol{\theta}}(t)\|_2$ denote the angular speed, where $\gamma_{\boldsymbol{\theta}}(t) := \gamma(t)/\|\gamma(t)\|_2$. Under this interpolation, the trajectory behaves as a linear chord, causing the magnitude $\gamma_r(t)$ to sag ($\dot{\gamma}_r(t) \neq 0$). This geometric sagging amplifies the angular speed near the midpoint ($t = 0.5$). Taking the time derivative of the angular speed yields a non-zero variation in angular velocity:

$$\dot{\omega}(t) \propto -\dot{\gamma}_r(t)/\gamma_r(t)^3 \neq 0. \tag{1}$$

Eq. (1) shows that the angular dynamics are actually affected by radial dimension in the flat space (*i.e.*, radial–angular coupling). This incompatible geometry forces the trajectory to exhibit *non-uniform angular speed* with non-zero angular acceleration (*cf.*, Figure 1(c)). It inherently introduces additional curvature into the effective velocity field on the angular sub-manifold, exacerbating the difficulty of regression training and leading to extra truncation errors[4], which causes suboptimal alignment performance during multi-step alignment for adaptation. If FMA retains the radial information without any normalization, the angular speed will be even more non-uniform.

The following work HFM [12] forcefully adjusts the pre-trained cross-modal features into a hyperbolic space [14] before applying FM via centripetal hyperbolic alignment, which can implicitly bypass this distortion. Although adjusting features toward a hyperbolic boundary can artificially boost class separability to achieve competitive few-shot performance, this extra mapping step introduces an inefficient two-stage training paradigm and unnecessary geometric complexity, which disrupts the inherent angular sub-manifold of pre-trained cross-modal features, leading to incompatibility with standard, off-the-shelf adaptation pipelines widely used in current cross-modal tasks.

3.3 Direct Product Flow Matching (DP-FM)

To eliminate this distortion while preserving the inherent manifold of pre-trained cross-modal representations, the angular trajectory should independently follow a *constant-speed geodesic* on $\mathbb{S}^{d-1}$. In Riemannian geometry, this requires the covariant derivative of the angular velocity to strictly vanish: $\frac{D}{dt}\dot{\gamma}_{\boldsymbol{\theta}}(t) = \mathbf{0}$. Under the generalized metric, expanding the angular acceleration via Christoffel symbols [39] ($\Gamma^{\theta}_{r\theta} = \frac{\phi'(r)}{\phi(r)}$) reveal the coupling:

$$\frac{D}{dt}\dot{\gamma}_{\boldsymbol{\theta}}(t) = -2\frac{\phi'(\gamma_r(t))}{\phi(\gamma_r(t))}\dot{\gamma}_r(t)\dot{\gamma}_{\boldsymbol{\theta}}(t). \tag{2}$$

This mandates $\phi'(\gamma_r(t)) = 0$, leading to the metric: $\phi(r) = \text{CONSTANT}$. This constant-warping metric establishes a theoretically optimal product manifold, *i.e.*, a decoupled cylindrical manifold $\mathbb{R}_{>0} \times \mathbb{S}^{d-1}$. Driven by this finding, we propose **direct product flow matching (DP-FM)**. By constructing the continuous flow exclusively under this optimal metric, the target velocity field naturally decomposes into two independent dynamic objectives: $\boldsymbol{v}^*(\boldsymbol{x}_t, t) = \frac{d}{dt}\gamma(t) = (\dot{\gamma}_r(t), \dot{\gamma}_{\boldsymbol{\theta}}(t))$. Specifically, given an image feature $\boldsymbol{x}_0 = r_0\boldsymbol{\theta}_0$ and a text feature $\boldsymbol{x}_1 = r_1\boldsymbol{\theta}_1$, the decoupled ground-truth geodesics are formulated as:

$$\gamma_r(t) = (1-t)r_0 + tr_1, \quad \gamma_{\boldsymbol{\theta}}(t) = \frac{\sin((1-t)\alpha)}{\sin\alpha}\boldsymbol{\theta}_0 + \frac{\sin(t\alpha)}{\sin\alpha}\boldsymbol{\theta}_1, \tag{3}$$

where $\alpha = \arccos(\langle\boldsymbol{\theta}_0, \boldsymbol{\theta}_1\rangle) \in [0, \pi]$. This explicit dual-geodesic dynamics cleanly resolves the two geometric limitations of previous methods: **1). Mitigating Angular Dynamics Distortion:** As formulated in Eq. (3), the angular geodesic $\gamma_{\boldsymbol{\theta}}(t)$ naturally evolves via exact spherical linear interpolation (slerp) on $\mathbb{S}^{d-1}$. This guarantees a constant-speed geodesic ($\frac{D}{dt}\dot{\gamma}_{\boldsymbol{\theta}}(t) = \mathbf{0}$), fundamentally eliminating the angular dynamics distortion, thereby preventing extra curvature in the target velocity field. **2). Addressing Radial Dynamics Neglect:** Rather than discarding radial information, the radial geodesic $\gamma_r(t)$ is explicitly retained and modeled as an independent radial trajectory, which preserves modality confidence to distinguish OOD from ID samples for the velocity network.

3.4 Algorithmic Design & Implementation of WP-FM

We introduce several critical designs for both the training and inference stages for WP-FM. The detailed algorithm is shown in Algorithm 1.

[4]Derivations are in the Appendix A, showing that any non-constant warping metric induces an extra truncation error.

Algorithm 1: Warped Product Flow Matching (WP-FM)

Phase 1: Training

Require: Image and text features $\mathcal{P}_0, \mathcal{P}_1$, velocity field $\boldsymbol{v}_\psi$, steps N, dropout p_{drop}.

while *not converged* **do**

- Sample $\boldsymbol{x}_0 \sim \mathcal{P}_0, \boldsymbol{x}_1 \sim \mathcal{P}_1, t \sim \mathcal{U}(0,1)$, extract hidden states condition $\boldsymbol{c}$;
- Time-shift schedule $t \leftarrow \frac{s \cdot t}{1+(s-1)t}$, set $\boldsymbol{c} \leftarrow \emptyset$ with probability p_{drop};
- Polar decompose $r_i \leftarrow \|\boldsymbol{x}_i\|_2, \boldsymbol{\theta}_i \leftarrow \boldsymbol{x}_i / r_i$ $(i \in \{0,1\})$;
- Compute coupled geodesics $(\gamma_r(t), \gamma_{\boldsymbol{\theta}}(t))$ and target velocities $(\dot{\gamma}_r(t), \dot{\gamma}_{\boldsymbol{\theta}}(t))$;
- Assemble feature $\boldsymbol{x}_t \leftarrow \gamma_r(t) \cdot \gamma_{\boldsymbol{\theta}}(t)$, predict velocity $\boldsymbol{v}_\psi \leftarrow \boldsymbol{v}_\psi(\boldsymbol{x}_t, t|\boldsymbol{c})$;
- Project to tangent spaces: $v_{rad} \leftarrow \langle \boldsymbol{v}_\psi, \gamma_{\boldsymbol{\theta}}(t) \rangle$, $\boldsymbol{v}_{ang} \leftarrow (\boldsymbol{v}_\psi - v_{rad}\gamma_{\boldsymbol{\theta}}(t))/\gamma_r(t)$;
- Compute metric-aware loss: $\mathcal{L} \leftarrow \|v_{rad} - \dot{\gamma}_r(t)\|_2^2 + \phi(\gamma_r(t))^2 \|\boldsymbol{v}_{ang} - \dot{\gamma}_{\boldsymbol{\theta}}(t)\|_2^2$;
- Update network parameters ψ via $\nabla_\psi \mathcal{L}$;

Phase 2: Inference

Require: Image feature $\boldsymbol{x}_{test} \sim \mathcal{P}_0$, hidden states condition $\boldsymbol{c}$, CFG scale w.

Initialize $r_0 \leftarrow \|\boldsymbol{x}_{test}\|_2, \boldsymbol{\theta}_0 \leftarrow \boldsymbol{x}_{test}/r_0$, set step size $\Delta t \leftarrow 1/N$;

for $k = 0$ **to** $N-1$ **do**

- State $\boldsymbol{x}_k \leftarrow r_k \cdot \boldsymbol{\theta}_k$ at time $t \leftarrow k\Delta t$;
- Predict condition-guided velocity: $\tilde{\boldsymbol{v}}_\psi \leftarrow \boldsymbol{v}_\psi(\boldsymbol{x}_k, t|\emptyset) + w \cdot \big(\boldsymbol{v}_\psi(\boldsymbol{x}_k, t|\boldsymbol{c}) - \boldsymbol{v}_\psi(\boldsymbol{x}_k, t|\emptyset)\big)$;
- Project to tangent spaces: $v_{rad} \leftarrow \langle \tilde{\boldsymbol{v}}_\psi, \boldsymbol{\theta}_k \rangle$, $\boldsymbol{v}_{ang} \leftarrow (\tilde{\boldsymbol{v}}_\psi - v_{rad}\boldsymbol{\theta}_k)/r_k$;
- Integration on the manifold: $(r_{k+1}, \boldsymbol{\theta}_{k+1}) \leftarrow \text{Exp}^{\mathcal{M}}_{(r_k, \boldsymbol{\theta}_k)}([v_{rad}, \boldsymbol{v}_{ang}]\Delta t)$;

Output: $\boldsymbol{x}_N \leftarrow r_N \cdot \boldsymbol{\theta}_N$;

Classifier-Free Guidance in Adaptation. As identified in Limitation 3), the pre-trained VLMs often yield generalized features that discard dataset-specific information in few-shot scenarios. To explicitly inject the missing information back into the alignment process, we equip WP-FM with classifier-free guidance (CFG) [17], where the conditioning signal $\boldsymbol{c}$ is derived from the pre-trained model's hidden states given the input $\boldsymbol{x}_0$. During training, we parameterize conditional vector fields $v_\theta(\boldsymbol{x}_t, t|\boldsymbol{c})$. We jointly train an unconditional field by randomly replacing $\boldsymbol{c}$ with a learnable null token $\emptyset$ at a fixed dropout probability p_{drop}. During inference, we construct a semantic-guided velocity field $\tilde{v}_\psi$:

$$\tilde{v}_\psi(\boldsymbol{x}_t, t|\boldsymbol{c}) = v_\psi(\boldsymbol{x}_t, t|\emptyset) + w \cdot \Big(v_\psi(\boldsymbol{x}_t, t|\boldsymbol{c}) - v_\psi(\boldsymbol{x}_t, t|\emptyset)\Big), \tag{4}$$

where $w \geq 1$ is the guidance scale. By integrating over $\tilde{v}_\psi$, DP-FM amplifies the target-domain information. This effectively steers the flow away from context-agnostic trajectories, recovering the fine-grained information discarded during pre-trained feature extraction for enhanced adaptation.

Velocity Decomposition. Standard neural networks typically output velocity predictions in the ambient Euclidean space, denoted as $\boldsymbol{v}_\psi(\boldsymbol{x}_t, t) \in \mathbb{R}^d$. To adhere to our warped product manifold geometry, this output is explicitly projected onto the tangent spaces $\mathcal{TM}_r$ and $\mathcal{TM}_{\boldsymbol{\theta}}$. Specifically, as shown in Figure 3, the radial velocity is extracted via the inner product: $v_{rad} = \langle \boldsymbol{v}_\psi, \gamma_{\boldsymbol{\theta}}(t) \rangle$. For the angular component, we isolate the orthogonal projection and fundamentally divide it by the radius r_t:

$$\boldsymbol{v}_{ang} = \frac{1}{\gamma_r(t)} \left(\boldsymbol{v}_\psi - v_{rad}\gamma_{\boldsymbol{\theta}}(t)\right), \tag{5}$$

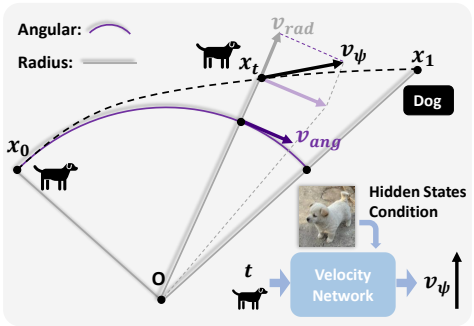

Figure 3: **Illustration of DP-FM.**

which converts the predicted velocity in ambient Euclidean space into the angular velocity measured in radians, ensuring strict dimensionality alignment with the ground-truth geodesic.

Metric-Aware Loss Objective. Unlike standard FM, which minimizes mean squared error (MSE) uniformly across Euclidean dimensions, WP-FM utilizes a specific optimization objective function corresponding to the Riemannian manifold. Guided by the metric tensor $ds^2 = dr^2 + \phi(r)^2 d\boldsymbol{\theta}^2$, the

Table 1: **Comparison with Pure FM methods for few-shot adaptation.**

Shots	Methods	Aircraft	SAT	DTD	SUN	Cars	**AVG**	Pets	UCF	Flowers	Caltech	Food	Net	**AVG**
0	CLIP [4]	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	WP-FM (Euclidean) [11]	26.8	66.7	53.5	66.3	67.0	56.1$_{+7.2}$	90.2	71.2	82.2	93.9	86.1	67.5	81.9$_{+3.1}$
	WP-FM (Hyperbolic) [12]	26.3	66.5	53.5	65.7	67.4	55.9$_{+7.0}$	89.8	71.3	81.7	94.1	86.1	67.5	81.8$_{+3.0}$
	DP-FM	26.3	68.6	53.3	66.3	67.2	**56.3**$_{+7.4}$	90.5	72.2	83.2	94.0	86.1	67.4	**82.2**$_{+3.4}$
4	WP-FM (Euclidean) [11]	30.7	77.9	61.4	69.9	71.0	62.2$_{+13.3}$	91.8	77.8	92.3	95.2	86.4	68.9	85.4$_{+6.6}$
	WP-FM (Hyperbolic) [12]	32.4	78.5	61.1	70.1	72.6	62.9$_{+14.0}$	91.9	78.5	94.3	94.4	86.5	68.8	85.7$_{+6.9}$
	DP-FM	32.4	79.9	61.9	70.6	71.9	**63.3**$_{+14.4}$	92.1	78.1	93.6	95.5	86.5	69.0	**85.8**$_{+7.0}$
16	WP-FM (Euclidean) [11]	44.6	87.0	70.6	74.4	82.7	71.9$_{+23.0}$	92.9	82.7	98.3	95.7	86.8	71.1	87.9$_{+9.1}$
	WP-FM (Hyperbolic) [12]	44.2	86.9	71.8	73.5	83.2	71.9$_{+23.0}$	93.1	83.9	97.8	95.8	86.9	71.0	**88.1**$_{+9.3}$
	DP-FM	46.0	87.8	71.2	73.8	83.6	**72.5**$_{+23.6}$	93.2	83.4	97.8	96.2	87.0	71.1	**88.1**$_{+9.3}$

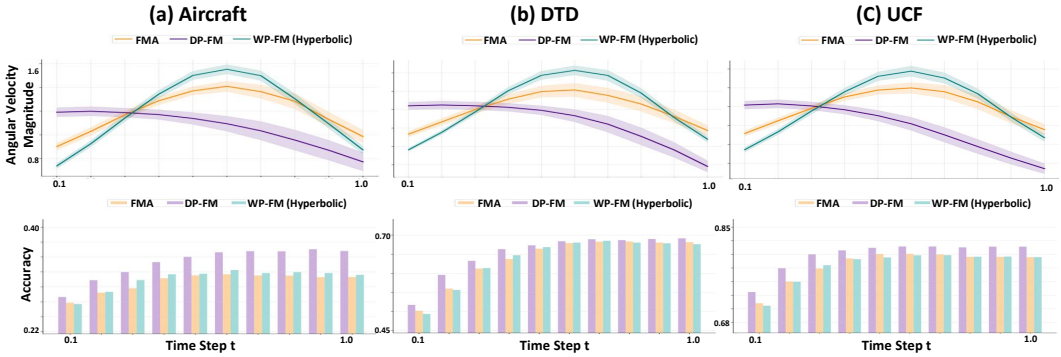

Figure 4: **Comparison between WP-FM (Euclidean [11], Hyperbolic [12]) and DP-FM** on (a). Aircraft, (b). DTD, and (c). UCF dataset at 20 epochs.

regression loss for WP-FM is formulated as a metric-aware objective:

$$\mathcal{L}_{\text{WP-FM}} = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{x}_1}\left[\|v_{rad} - \dot{\gamma}_r(t)\|_2^2 + \phi(\gamma_r(t))^2\|\boldsymbol{v}_{ang} - \dot{\gamma}_{\boldsymbol{\theta}}(t)\|_2^2\right]. \tag{6}$$

This ensures that the gradient updates strictly follow the geometric structure of the manifold.

Inference via Exponential Map. During the inference phase[5] of WP-FM, discrete multi-step integration (*e.g.*, Euler method) in ambient space easily drifts off the manifold, leading to severe error accumulation. To guarantee mathematical closure, our method utilizes the formal Riemannian exponential map $\text{Exp}_{\boldsymbol{x}}^{\mathcal{M}}$ [16] for the ordinary differential equation (ODE) integration step Δt:

$$(r_{t+\Delta t}, \boldsymbol{\theta}_{t+\Delta t}) \leftarrow \text{Exp}^{\mathcal{M}}_{(r_t,\boldsymbol{\theta}_t)}([v_{rad}, \boldsymbol{v}_{ang}]\Delta t). \tag{7}$$

Time-shift Schedule. Standard uniform time sampling allocates insufficient capacity to the complex initial phase of cross-modal alignment. To address this, we apply a non-linear time-shift schedule, which is widely used in high-resolution generative modeling [40, 41, 42]: $t' = \frac{s \cdot t}{1+(s-1)t}$, where s is a time-shift parameter. Setting $s < 1$ monotonically biases the sampling density toward $t \to 0$, forcing the network to prioritize early velocity estimation and accelerating overall convergence.

4 Experiments

Datasets and Baselines. We evaluated DP-FM on few-shot image classification tasks. Specifically, we conducted experiments on 11 benchmarks, including Aircraft [43], EuroSAT [44], DTD [45], SUN397 [46], StanfordCars [47], OxfordPets [48], UCF101 [49], Flowers102 [50], Caltech101 [51], Food101 [52], and ImageNet [53]. Following the FMA setting, we partitioned them into two subsets: five datasets form the difficult group, and the remaining six constitute the easy group. For each dataset, we adopted the standard train/validation/test splits. Under the K-shot setting, we constructed the training set by randomly sampling K labeled images per class, with the rest used for validation and testing. For PEFT methods, we adopted CoOp [7], CoCoOp [6], TIP-Adapter [54], CLIP-Adapter [25], PLOT++ [55], KgCoOp [56], ProGrad [57], and CLIP-LoRA [8] as our baselines. For multi-step FM methods, we adopted FMA [11] and HFM [12] as our baselines.

Settings. We implemented DP-FM based on the pre-trained CLIP, CLIP-LoRA [8], and CLIP-Adapter [25] model. Specifically, we first extract cross-modal features from the pre-trained models,

[5] In the inference stage, we utilize r_t and $\boldsymbol{\theta}_t$ to replace $\gamma_r(t)$ and $\gamma_{\boldsymbol{\theta}}(t)$.

Table 2: **Comparison with state-of-the-art FM methods for adaptation** building upon CLIP-LoRA and CLIP-Adapter. † denotes that the PEFT module is trained using centripetal hyperbolic alignment [14]. Δ represents the average performance improvement by FM over the baseline.

		Difficult Datasets							Easy Datasets							
Shot	Methods	Aircraft	SAT	DTD	SUN	Cars	**AVG**	Δ	Pets	UCF	Flowers	Caltech	Food	Net	**AVG**	Δ
1	CLIP-LoRA† [14]	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 [12]	30.5	76.8	57.4	71.2	69.8	61.1	+1.7	92.2	76.6	85.8	95.1	85.9	70.0	84.3	+0.8
	CLIP-LoRA [8]	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 [11]	28.3	73.0	55.1	70.6	69.8	59.4	+0.6	92.1	75.9	84.9	94.5	85.2	70.2	83.8	+0.8
	+DP-FM	29.6	81.7	56.9	71.0	70.8	**62.0**	**+3.3**	93.1	76.8	89.0	94.4	85.0	70.3	**84.8**	**+1.8**
4	CLIP-LoRA† [14]	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 [12]	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 [8]	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 [11]	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
	+DP-FM	41.7	89.3	69.2	73.4	80.5	70.8	**+3.5**	92.2	82.3	96.8	96.0	83.7	72.1	87.2	**+1.6**
16	CLIP-LoRA† [14]	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 [12]	61.4	93.1	75.2	77.6	88.6	**79.2**	+1.9	94.0	88.9	98.7	96.8	87.3	73.7	**89.9**	+0.4
	CLIP-LoRA [8]	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 [11]	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
	+DP-FM	59.2	93.1	77.1	77.6	88.8	**79.2**	**+3.1**	92.8	88.0	99.0	96.6	85.8	74.4	89.4	**+1.2**
16	CLIP-Adapter† [14]	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 [12]	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	+1.5
	CLIP-Adapter [25]	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 [11]	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
	+DP-FM	45.1	86.4	71.6	75.8	82.7	**72.3**	**+9.9**	92.9	84.6	97.9	96.0	87.1	72.4	**88.5**	**+1.9**

then utilize a velocity network to regress velocity fields that transport image features to text features. Following the FMA [11] setting, we utilize a lightweight MAR [58] network as our velocity network. In the DP-FM algorithm, we set CONSTANT $= 25$ (see Appendix D for derivation and sensitivity analysis), time-shift parameter $s = 0.1$(see Appendix E for sensitivity analysis), dropout probability $p_{drop} = 0.1$, and CFG scale $w = 5.0$ across all the experiments. We trained DP-FM with AdamW [59] using a learning rate of 2×10^{-4} and a weight decay of 0.01.

Performance against FM Baselines. We evaluate DP-FM against recent FM methods for few-shot adaptation, including WP-FM (in Euclidean[6] [11], Hyperbolic[7] [12] space). For a fair comparison, all the methods operate on the same pre-trained CLIP model and are equipped with the time-shift schedule and CFG. As shown in Table 1, DP-FM consistently achieves higher average accuracy across the 1, 4, and 16-shot settings. By isolating the geometric design on the pre-trained cross-modal features, these results explicitly confirm that the decoupled cylindrical manifold provides an enhanced foundation for FM compared to flat Euclidean or Hyperbolic spaces. Furthermore, as illustrated in Figure 4, DP-FM exhibits a more uniform angular speed and enhanced adaptation performance.

Furthermore, we extend our evaluation to compare DP-FM against state-of-the-art FM methods when integrated with established PEFT baselines, specifically CLIP-LoRA and CLIP-Adapter, as shown in Table 2. DP-FM consistently delivers the most substantial performance improvements (Δ) over the baselines across different datasets. For instance, under the 16-shot setting with CLIP-Adapter, DP-FM achieves a remarkable average improvement of $+9.9\%$ on the difficult datasets, which significantly surpasses the gains provided by FMA ($+5.5\%$) and HFM ($+7.8\%$). Similarly, when built upon the CLIP-LoRA architecture, DP-FM outperforms competing FM methods in most cases, demonstrating its enhanced capability to refine feature alignment effectively.

Comparison with state-of-the-art PEFT. As presented in Table 3, we evaluate the performance of our DP-FM against a comprehensive set of single-step PEFT methods. Built upon CLIP-LoRA, DP-FM consistently establishes new state-of-the-art results across 1, 4, and 16-shot settings for both difficult and easy dataset groups. Most notably, under the 16-shot setting, DP-FM achieves an average accuracy of 79.2% on the difficult datasets and 89.4% on the easy datasets. This performance significantly exceeds strong single-step baselines such as PLOT++ (74.1% and 88.7%) and the standalone CLIP-LoRA (76.1% and 88.2%). These results confirm that modeling the adaptation process as a continuous flow on a decoupled cylindrical manifold yields superior generalization compared to conventional static, single-step prompting or adaptation strategies.

[6]WP-FM (Euclidean) is the same as FMA [11], operating on pre-trained cross-modal features

[7]To isolate the impact of different geometries, we omit HFM's centripetal hyperbolic alignment and diameter-based stopping phases, ensuring FMA, HFM, and DP-FM operate on identical pre-trained cross-modal features in 1-stage training.

Table 3: **Comparison with other state-of-the-art PEFT methods.** Building upon CLIP-LoRA, adaptation performance can be further improved by our DP-FM.

		Difficult Datasets						Easy Datasets						
Shot	Method	Aircraft	SAT	DTD	SUN	Cars	**AVG**	Pets	UCF	Flowers	Caltech	Food	Net	**AVG**
0	CLIP [4]	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 [7]	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 [6]	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 [54]	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 [25]	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++ [55]	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 [56]	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 [57]	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 [8]	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
	+DP-FM	29.6	81.7	56.9	71.0	70.8	**62.0**$_{+3.3}$	93.1	76.8	89.0	94.4	85.0	70.3	**84.8**$_{+1.8}$
4	CoOp [7]	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 [6]	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 [54]	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 [25]	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++ [55]	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 [56]	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 [57]	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 [8]	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
	+DP-FM	41.7	89.3	69.2	73.4	80.5	**70.8**$_{+3.5}$	92.2	82.3	96.8	96.0	83.7	72.1	**87.2**$_{+1.6}$
16	CoOp [7]	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 [6]	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 [54]	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 [25]	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++ [55]	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 [56]	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 [57]	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 [8]	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
	+DP-FM	59.2	93.1	77.1	77.6	88.8	**79.2**$_{+3.1}$	92.8	88.0	99.0	96.6	85.8	74.4	**89.4**$_{+1.2}$

Table 4: **Ablation study** on Time-shift Schedule (**T**), Radius component (**R**), and Classifier-free guidance (**C**) under the 16-shot setting. The numbers in parentheses denote CFG scale w.

Method	T	R	C	Aircraft	SAT	DTD	SUN	Cars	**AVG**	Pets	UCF	Flowers	Caltech	Food	Net	**AVG**
CLIP [4]	×	×	×	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
+DP-FM	×	×	×	42.7	85.5	70.8	73.7	82.9	71.1$_{+22.2}$	92.9	82.3	97.3	95.9	86.8	70.8	87.7$_{+8.9}$
	×	✓	✓(5)	41.3	86.7	69.8	72.4	79.8	70.0$_{+21.1}$	92.1	82.6	97.3	95.2	86.4	71.0	87.4$_{+8.6}$
	✓	×	✓(5)	44.2	86.8	69.6	73.6	83.6	71.6$_{+22.7}$	93.0	82.6	97.8	96.1	86.9	71.0	87.9$_{+9.1}$
	✓	✓	×	44.6	87.0	71.1	73.8	82.6	71.8$_{+22.9}$	92.7	82.6	97.7	95.9	87.1	71.1	87.9$_{+9.1}$
	✓	✓	✓(1)	45.8	87.8	71.2	73.1	82.6	72.1$_{+23.2}$	93.1	82.6	97.7	96.2	87.0	71.1	**88.0**$_{+9.2}$
	✓	✓	✓(5)	46.0	87.8	71.2	73.8	83.6	**72.5**$_{+23.6}$	93.2	83.4	97.8	96.2	87.0	71.1	**88.1**$_{+9.3}$

Ablation Study and Component Analysis. To analyze the contribution of each proposed component, we conduct ablation experiments under the 16-shot setting (Table 4). First, we examine the effect of the Time-shift Schedule (T) on training convergence. Notably, adding the Radius component (R) and classifier-free guidance (C) without T (Row 3) yields a lower performance (70.0%) than the baseline lacking all three components (Row 2, 71.1%). This is likely because introducing extra radial and conditional information increases optimization complexity, which delays the training process. The time-shift schedule addresses this by prioritizing training of early-stage velocity estimation. With T stabilizing the optimization, the empirical benefits of R and C become clearly evident, as integrating the radial component and CFG both lead to consistent performance improvements. Ultimately, combining all three components achieves the highest average accuracy of 72.5% on difficult datasets, validating the integrated design of our method.

5 Conclusion

In this paper, we investigated the geometric limitations of existing FM methods for few-shot VLM adaptation, demonstrating that they inherently suffer from angular dynamics distortion, radial dynamics neglect, and context-agnostic flows. To address this, we introduced warped product flow matching (WP-FM), a unified Riemannian framework. By adopting a constant-warping metric, we derived DP-FM to strictly decouple radial and angular dynamics, ensuring constant-speed geodesic transport and preserving modality confidence. Additionally, we integrated hidden-state conditioned classifier-free guidance to recover dataset-specific context. Extensive experiments across 11 vision-language benchmarks confirm our method achieves enhanced adaptation accuracy during multi-step alignment.

Acknowledgment

This work was supported by National Natural Science Foundation of China (NSFC) Young Scientists Fund Category B (62522216), Hong Kong SAR Research Grants Council (RGC) General Research Fund (16219025), National Natural Science Foundation of China (NSFC) Young Scientists Fund Category C (62402408), and Hong Kong SAR Research Grants Council (RGC) Early Career Scheme (26208924).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 35:23716–23736, 2022.

[2] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023.

[3] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36:34892–34916, 2023.

[4] 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 *ICML*, pages 8748–8763. PMLR, 2021.

[5] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900. PMLR, 2022.

[6] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022.

[7] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022.

[8] Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1593–1603, 2024.

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

[10] Hongxu Chen, Zhen Wang, Runshi Li, Bowei Zhu, and Long Chen. Iteris: Iterative inference-solving alignment for lora merging. In *CVPR*, pages 4829–4838. IEEE, 2025.

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

[12] 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. *arXiv preprint*, 2026.

[13] Minsoo Jo, Dongyoon Yang, and Taesup Kim. Angular gradient sign method: Uncovering vulnerabilities in hyperbolic networks. In *AAAI*, volume 40, pages 5566–5574, 2026.

[14] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *ICML*, pages 7694–7731. PMLR, 2023.

[15] Jaewoo Park, Jacky Chen Long Chai, Jaeho Yoon, and Andrew Beng Jin Teoh. Understanding the feature norm for out-of-distribution detection. In *ICCV*, pages 1557–1567, 2023.

[16] Ricky TQ Chen and Yaron Lipman. Flow matching on general geometries. *arXiv preprint*, 2023.

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint*, 2022.

[18] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, pages 18082–18091, 2022.

[19] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023.

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021.

[21] Zhen Wang, Lin Li, and Long Chen. Recent advances in finetuning multimodal large language models. *AI Magazine*, 46(3):e70025, 2025.

[22] Beier Zhu and Hanwang Zhang. Debiasing vision-language models for vision tasks: a survey. *Frontiers of Computer Science*, 19(1):191321, 2025.

[23] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint*, 2024.

[24] 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*, pages 2790–2799. PMLR, 2019.

[25] 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*, 132(2):581–595, 2024.

[26] Zhen Wang, Jun Xiao, Yueting Zhuang, Fei Gao, Jian Shao, and Long Chen. Learning combinatorial prompts for universal controllable image captioning. *IJCV*, 133(1):129–150, 2025.

[27] 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*, 2026.

[28] Zhenqi He, Lin Li, and Long Chen. Flowcomposer: Composable flows for compositional zero-shot learning. In *CVPR*, 2026.

[29] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint*, 2022.

[30] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint*, 2022.

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[32] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *ECCV*, pages 23–40. Springer, 2024.

[33] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023.

[34] Hongxu Chen, Hongxiang Li, Zhen Wang, and Long Chen. Bi-anchor interpolation solver for accelerating generative modeling. *arXiv preprint*, 2026.

[35] Hongxiang Li, Hongxu Chen, Chenyang Zhu, Xiaoshuang Huang, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Long Chen. Spar: Semantic-pixel self-alignment and adaptive routing for unified multimodal models. In *ECCV*, pages 324–343. Springer, 2026.

[36] Zhen Wang, Youcan Xu, Jun Xiao, and Long Chen. Flowmotion: training-free flow guidance for video motion transfer. In *CVPR*, pages 38143–38153, 2026.

[37] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020.

[38] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint*, 2020.

[39] Barrett O'neill. *Semi-Riemannian geometry with applications to relativity*, volume 103. Academic press, 1983.

[40] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

[41] Jiayan Teng, Wendi Zheng, Ming Ding, Wenyi Hong, Jianqiao Wangni, Zhuoyi Yang, and Jie Tang. Relay diffusion: Unifying diffusion process across resolutions for image synthesis. *arXiv preprint*, 2023.

[42] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *ICML*, pages 13213–13232. PMLR, 2023.

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

[44] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *J-STARS*, 12(7):2217–2226, 2019.

[45] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014.

[46] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492. IEEE, 2010.

[47] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV*, pages 554–561, 2013.

[48] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505. IEEE, 2012.

[49] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint*, 2012.

[50] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pages 722–729. IEEE, 2008.

[51] 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 *CVPR*, pages 178–178. IEEE, 2004.

[52] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461. Springer, 2014.

[53] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[54] 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*, pages 493–510. Springer, 2022.

[55] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. *arXiv preprint*, 2022.

[56] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*, pages 6757–6767, 2023.

[57] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *ICCV*, pages 15659–15669, 2023.

[58] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *NeurIPS*, 37:56424–56445, 2024.

[59] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint*, 2017.

A Angular Truncation Error of WP-FM

In this section, we provide a detailed derivation of the angular truncation error for first-order Riemannian ODE solvers under the warped product geometry introduced in Sec. 3.

Proposition 1 (Angular Truncation Error). Let $(\mathcal{M}, g_\phi)$ be the warped product manifold defined in Sec. 3.1:

$$\mathcal{M} = \mathcal{M}_r \times_{g_\phi} \mathcal{M}_{\boldsymbol{\theta}}, \quad g_\phi = dr^2 + \phi(r)^2 d\boldsymbol{\theta}^2, \tag{8}$$

where $\mathcal{M}_r = \mathbb{R}_{>0}$, $\mathcal{M}_{\boldsymbol{\theta}} = \mathbb{S}^{d-1}$, $\phi \in C^2(\mathbb{R}_{>0})$ and $\phi(r) > 0$. Let $\gamma(t) = (\gamma_r(t), \gamma_{\boldsymbol{\theta}}(t))$ be a unit-speed geodesic defined on $t \in [0, T]$. Denote the angular velocity magnitude (speed) as $\omega(t) := \|\dot{\gamma}_{\boldsymbol{\theta}}(t)\|_2$. Consider the first-order Riemannian integrator applied to the angular component:

$$\hat{\gamma}_{\boldsymbol{\theta}}(t + \Delta t) = \mathrm{Exp}^{\mathcal{M}_{\boldsymbol{\theta}}}_{\gamma_{\boldsymbol{\theta}}(t)} \big(\Delta t\, \dot{\gamma}_{\boldsymbol{\theta}}(t)\big). \tag{9}$$

Then there exists a constant $C > 0$, depending only on bounded trajectory quantities, such that the angular truncation error $\epsilon_{\boldsymbol{\theta}}$ satisfies:

$$\epsilon_{\boldsymbol{\theta}} := d_{\mathcal{M}_{\boldsymbol{\theta}}}\big(\hat{\gamma}_{\boldsymbol{\theta}}(t + \Delta t), \gamma_{\boldsymbol{\theta}}(t + \Delta t)\big) \le C\, \Delta t^2 \left| \frac{\phi'(\gamma_r(t))}{\phi(\gamma_r(t))} \dot{\gamma}_r(t) \right| \|\dot{\gamma}_{\boldsymbol{\theta}}(t)\|_2 + O(\Delta t^3). \tag{10}$$

Moreover, under the non-degeneracy condition $\dot{\gamma}_r(t) \neq 0$ and $\dot{\gamma}_{\boldsymbol{\theta}}(t) \neq \mathbf{0}$, the leading-order term vanishes if and only if:

$$\phi'(r) \equiv 0. \tag{11}$$

Proof.

Local truncation error on Riemannian manifolds. *Lemma 1 (Exponential Euler local error).* Let $(\mathcal{N}, h)$ be a Riemannian manifold and $x(t)$ a C^2 curve. Then:

$$d_{\mathcal{N}}\big(\mathrm{Exp}_{x(t)}(\Delta t\, \dot{x}(t)),\, x(t + \Delta t)\big) = \frac{1}{2}\Delta t^2 \|\nabla_{\dot{x}} \dot{x}\|_h + O(\Delta t^3), \tag{12}$$

where the remainder is uniform on compact intervals.

Levi-Civita connection under warped product geometry. For $\gamma(t) = (\gamma_r(t), \gamma_{\boldsymbol{\theta}}(t))$, the Levi-Civita connection under metric g_ϕ satisfies:

$$\nabla_{\dot{\gamma}} \dot{\gamma} = \Big(\ddot{\gamma}_r - \phi(\gamma_r)\phi'(\gamma_r) \|\dot{\gamma}_{\boldsymbol{\theta}}\|_2^2,\ \frac{D}{dt}\dot{\gamma}_{\boldsymbol{\theta}} + 2\frac{\phi'(\gamma_r)}{\phi(\gamma_r)} \dot{\gamma}_r \dot{\gamma}_{\boldsymbol{\theta}} \Big). \tag{13}$$

Since γ is a geodesic on $\mathcal{M}$, we have:

$$\nabla_{\dot{\gamma}} \dot{\gamma} = \mathbf{0}, \tag{14}$$

which directly implies the angular acceleration component is strictly zero:

$$\frac{D}{dt}\dot{\gamma}_{\boldsymbol{\theta}} = -2\frac{\phi'(\gamma_r)}{\phi(\gamma_r)} \dot{\gamma}_r \dot{\gamma}_{\boldsymbol{\theta}}. \tag{2}$$

Angular truncation error. Applying Lemma 1 on the angular sub-manifold $\mathcal{M}_{\boldsymbol{\theta}}$ gives:

$$\epsilon_{\boldsymbol{\theta}} = \frac{1}{2}\Delta t^2 \left\| \frac{D}{dt}\dot{\gamma}_{\boldsymbol{\theta}} \right\|_2 + O(\Delta t^3). \tag{15}$$

Substituting Eq. (2), we obtain the magnitude of the angular acceleration:

$$\left\| \frac{D}{dt}\dot{\gamma}_{\boldsymbol{\theta}} \right\|_2 = 2 \left| \frac{\phi'(\gamma_r)}{\phi(\gamma_r)} \dot{\gamma}_r \right| \|\dot{\gamma}_{\boldsymbol{\theta}}\|_2. \tag{16}$$

Therefore, the truncation error is:

$$\epsilon_{\boldsymbol{\theta}} = \Delta t^2 \left| \frac{\phi'(\gamma_r)}{\phi(\gamma_r)} \dot{\gamma}_r \right| \|\dot{\gamma}_{\boldsymbol{\theta}}\|_2 + O(\Delta t^3). \tag{17}$$

Characterization of vanishing error. The leading-order term vanishes iff:

$$\left| \frac{\phi'(\gamma_r)}{\phi(\gamma_r)} \dot{\gamma}_r \right| \|\dot{\gamma}_{\boldsymbol{\theta}}\|_2 = 0. \tag{18}$$

Under the non-degeneracy condition $\dot{\gamma}_r \neq 0$ and $\dot{\gamma}_{\boldsymbol{\theta}} \neq \mathbf{0}$, this holds iff:

$$\phi'(r) \equiv 0, \tag{19}$$

i.e., the warping function ϕ is constant.

This result characterizes the source of angular dynamics distortion in WP-FM. Any non-constant warping function introduces a coupling term proportional to $\frac{\phi'(r)}{\phi(r)}\dot{\gamma}_r$, which induces angular acceleration (Eq. (2)) and leads to a second-order truncation error on the angular sub-manifold of pre-trained cross-modal features. In the Euclidean case $\phi(r) = r$, this coefficient reduces to $\frac{\dot{\gamma}_r}{\gamma_r}$, which becomes large when $\gamma_r \to 0$, explaining the instability of angular updates observed in Sec. 3.

B Radial Information Reflects Modality Confidence

As illustrated in Figure 5, we visualize the statistical distribution of the radial magnitudes (*i.e.*, the L_2 norm) of image features extracted by the pre-trained CLIP image encoder across three different datasets. The overall magnitude distribution exhibits a clear hierarchical trend, ranking as OxfordPets > StanfordCars > FGVCAircraft. Notably, this trend aligns perfectly with the zero-shot classification accuracy of the pre-trained model on these respective datasets: 89.1% for OxfordPets, 65.5% for StanfordCars, and 24.8% for FGVCAircraft.

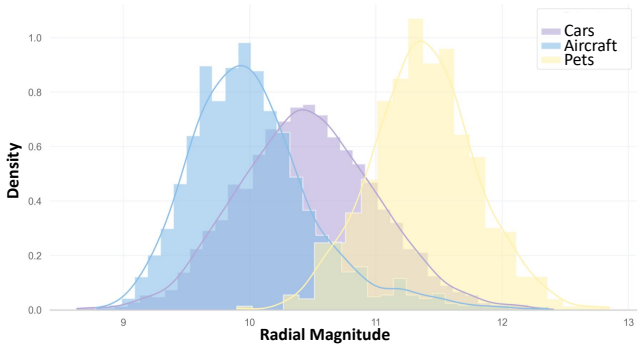

Figure 5: **Radial Magnitude Distribution across Datasets.**

This positive correlation provides compelling empirical evidence that the radial dimension intrinsically encodes the model's modality confidence. Specifically, when the model processes images from a domain it is highly confident in (such as OxfordPets), it produces features with larger radial magnitudes. Conversely, for more challenging or out-of-distribution domains (such as FGVCAircraft), the resulting feature magnitudes are significantly smaller. Consequently, these findings emphasize that radial information serves as a crucial explicit signal for distinguishing in-distribution data from uncertain samples, and discarding it inherently blinds the velocity network, degrading the performance observed in the ablation study.

C Controlled Geometry-Only Ablation and Multi-Seed Evaluation

Isolating the contribution of the underlying geometric design from auxiliary algorithmic components—namely Classifier-Free Guidance (**C**) and the Time-shift Schedule (**T**)—is essential for a rigorous evaluation. Specifically, comparing DP-FM (without **C**) against FMA (with **C**) would create an asymmetric and unfair comparison: because CFG independently improves adaptation performance, such a setup introduces a major confounding variable that obscures the intrinsic effect of the manifold geometry itself. Instead, a principled protocol to isolate the geometry's contribution is to equip all baseline geometries with the exact same enhancements (both **T** and **C**) and solely swap the underlying manifold metric. This controlled "geometry-only" comparison is presented in Table 1 of the main text. All methods in Table 1—including WP-FM (Euclidean), which represents the flat FMA geometry—operate on the identical pre-trained CLIP backbone and are equipped with the same **T** and **C** to ensure head-to-head fairness. Under this controlled geometry-only setting on the difficult datasets (16-shot) in Table 1:

- **WP-FM (Euclidean) [with T & C]** achieves **71.9%** (*i.e.*, FMA equipped with **T** and **C**).

- **DP-FM [with T & C]** achieves **72.5%**.

Because all other components are identical, this $+0.6\%$ gain strictly isolates the pure contribution of our decoupled direct-product geometry. To verify this improvement statistically, we re-evaluated these fully equipped methods across 7 representative datasets (4 difficult datasets and 3 easy datasets) using 3 additional random seeds, ensuring **identical data splits across all geometries for head-to-head fairness**. Table 5 demonstrates that DP-FM (**Constant (Ours)**) consistently outperforms competing geometries across random seeds and achieves the highest accuracy in 13 out of the 21 individual settings, as well as the highest average accuracy on difficult datasets across all 1-shot (54.07%), 4-shot (61.41%), and 16-shot (71.48%) regimes, further corroborating the effectiveness of the proposed decoupled geometry.

Table 5: **Multi-seed extended evaluation of the controlled geometry-only comparison (Table 1).** All methods operate on the identical CLIP features and are equipped with the same Time-shift Schedule (**T**) and Classifier-Free Guidance (**C**) across 3 random seeds with identical data splits.

		Difficult Datasets					Easy Datasets			
Shots	Methods	Aircraft	Cars	SAT	DTD	**AVG (Diff)**	Pets	Caltech	Flowers	**AVG (Easy)**
1	Euclidean	**26.99 ± 0.30**	67.00 ± 0.03	67.72 ± 0.96	54.02 ± 0.47	53.93	91.02 ± 0.48	94.09 ± 0.11	82.06 ± 0.29	89.06
	Hyperbolic	26.86 ± 0.42	67.11 ± 0.19	67.66 ± 1.03	53.98 ± 0.42	53.90	90.91 ± 0.65	**94.12 ± 0.03**	81.92 ± 0.30	88.98
	Constant (Ours)	26.77 ± 0.28	**67.28 ± 0.20**	**68.08 ± 1.05**	**54.14 ± 0.54**	**54.07**	**91.27 ± 0.46**	94.06 ± 0.06	**83.28 ± 0.45**	**89.54**
4	Euclidean	31.42 ± 0.48	72.05 ± 0.62	78.46 ± 0.39	61.05 ± 0.45	60.75	91.88 ± 0.13	95.52 ± 0.18	93.45 ± 0.68	93.62
	Hyperbolic	31.85 ± 0.42	**72.43 ± 0.17**	78.58 ± 0.21	60.99 ± 0.38	60.96	**91.90 ± 0.13**	95.31 ± 0.52	**93.95 ± 0.24**	93.72
	Constant (Ours)	**32.66 ± 0.39**	71.69 ± 0.22	**79.37 ± 0.43**	**61.93 ± 0.29**	**61.41**	91.88 ± 0.17	**95.67 ± 0.11**	93.84 ± 0.19	**93.80**
16	Euclidean	43.55 ± 0.94	82.14 ± 0.33	87.05 ± 0.15	71.24 ± 0.41	71.00	93.12 ± 0.15	96.06 ± 0.21	**98.31 ± 0.16**	**95.83**
	Hyperbolic	43.50 ± 0.84	82.29 ± 0.53	87.06 ± 0.14	**71.60 ± 0.21**	71.11	**93.15 ± 0.07**	96.07 ± 0.16	98.18 ± 0.27	95.80
	Constant (Ours)	**44.90 ± 0.64**	**82.49 ± 0.64**	**87.42 ± 0.24**	71.12 ± 0.19	**71.48**	93.04 ± 0.15	**96.12 ± 0.18**	98.07 ± 0.18	95.74

D Principled Justification and Sensitivity Analysis of the Warping Constant

In this section, we clarify the theoretical origin of setting CONSTANT $= 25$ (denoted as C below) in DP-FM and evaluate its empirical sensitivity across diverse datasets. Specifically, CONSTANT $= 25$ is not a grid-searched hyperparameter, but a **principled balancing factor** derived directly from the dimensionality (d) of the pre-trained CLIP features:

- **The Dimensional Derivation:** Under polar decomposition, the target velocity decomposes into a 1D radial scalar $\dot{\gamma}_r$ and a d-dimensional angular vector $\dot{\gamma}_{\boldsymbol{\theta}}$. While the 1D radial target is $\mathcal{O}(1)$, the angular signal is diluted across d dimensions, making the per-dimension regression target substantially smaller, on the order of $\mathcal{O}(1/\sqrt{d})$. For the CLIP features ($d = 512$) used in our experiments, this dilution factor $\sqrt{d}$ is approximately 22.6. We therefore select 25 as a principled proxy for $\sqrt{d}$.
- **Organic Loss Balancing:** In the metric-aware loss objective (Eq. (6)), the angular regression error is weighted by $\phi(r)^2$. By setting the constant warping function $\phi(r) \equiv 25 \approx \sqrt{d}$, the angular loss is naturally scaled by d. This compensates for the $1/\sqrt{d}$ per-dimension signal dilution, organically balancing the angular gradient updates with the 1D radial updates.

To empirically validate this derivation and examine dataset sensitivity, we evaluated DP-FM across a broad spectrum of CONSTANT ($C \in \{5, 15, 25, 35, 45, 100\}$) on 6 diverse datasets under the 16-shot setting at 100 epochs.

From the empirical results in Table 6, we observe three key findings:

- **Degradation at the Lower Bound ($C = 5$):** When the constant is significantly lower than the dimensional proxy $\sqrt{d}$, the average performance drops to 87.06%. A small constant fails to compensate for the $1/\sqrt{d}$ signal dilution, thereby under-optimizing the angular alignment.
- **Dataset-Agnostic Robustness:** While the exact peak varies slightly across datasets, the overall performance remains highly stable. Across the broad range of $C \in [25, 100]$, the average accuracy fluctuates by merely $\sim 0.2\%$ (87.44% to 87.65%). This confirms that as long as C sufficiently compensates for the $\sqrt{d}$ signal dilution, the optimization remains well-balanced across domains.
- **Choice of $C = 25$:** While $C = 35$ yields a marginally higher average accuracy (87.65% *vs.* 87.44%), we retain $C = 25$ across all experiments to align strictly with our theoretical dimensional derivation, as the empirical results confirm its strong robustness without per-dataset tuning.

Table 6: **Sensitivity analysis of the constant warping factor CONSTANT (C)** across 6 diverse datasets under the 16-shot setting at 100 epochs.

CONSTANT (C)	**5**	**15**	**25**	**35**	**45**	**100**
Flowers	98.0	98.1	**98.4**	98.1	98.0	98.1
Pets	92.6	92.6	92.6	**92.7**	92.6	92.3
Caltech	95.9	95.6	95.7	95.6	**95.9**	95.6
DTD	69.5	70.9	69.9	70.9	70.0	**71.1**
SAT	86.7	87.0	87.1	87.4	**87.5**	86.3
Cars	79.7	79.8	81.0	81.2	81.3	**81.7**
AVG	87.06	87.36	87.44	**87.65**	87.55	87.52

E Sensitivity Analysis of the Time-Shift Schedule

As introduced in Sec. 3.4, standard uniform time sampling allocates insufficient capacity to the complex initial phase of cross-modal alignment. To address this, we employ a non-linear time-shift schedule (**T**): $t' = \frac{s \cdot t}{1+(s-1)t}$, where s denotes the time-shift parameter. Setting $s < 1$ monotonically biases the sampling density toward $t \to 0$, prioritizing early-stage velocity estimation and stabilizing training convergence. To evaluate the sensitivity of DP-FM to this schedule, we vary the time-shift parameter $s \in \{1.0, 0.5, 0.1, 0.05\}$ across 6 diverse datasets under the 16-shot setting at 100 epochs, where $s = 1.0$ degenerates to standard uniform sampling ($t' = t$) and $s = 0.1$ is our default setting.

Table 7: **Sensitivity analysis of the Time-shift Schedule T (s)** across 6 diverse datasets under the 16-shot setting at 100 epochs. Note that $s = 1.0$ corresponds to standard uniform time sampling.

Time-Shift T (s)	**1.0 (Uniform)**	**0.5**	**0.1 (Default)**	**0.05**
Flowers	97.3	97.6	**97.9**	**97.9**
Pets	91.9	92.2	**92.3**	92.2
Caltech	95.8	**96.2**	96.1	96.1
DTD	71.2	**71.3**	70.3	69.7
SAT	83.8	85.0	86.6	**86.8**
Cars	80.3	81.1	81.4	**81.5**
AVG	86.72	87.22	**87.44**	87.36

From the empirical results in Table 7, we highlight two key observations:

- **Effectiveness of Non-Uniform Sampling ($s \leq 0.5$):** Non-uniform time-shift schedules ($s \leq 0.5$) consistently outperform standard uniform sampling ($s = 1.0$, 86.72%) by a clear margin. In particular, fine-grained and challenging domains such as SAT ($+2.8\%$ at $s = 0.1$) and Cars ($+1.1\%$ at $s = 0.1$) benefit substantially from concentrating training capacity near $t \to 0$, confirming that stabilizing early-phase velocity estimation is critical when adapting coupled radial and conditional dynamics.
- **Robustness Around Default $s = 0.1$:** Across the non-uniform regime $s \in [0.05, 0.5]$, the overall performance remains stable, fluctuating within only $\sim 0.22\%$ (87.22% to 87.44%). Our default choice of $s = 0.1$ achieves the highest average accuracy (87.44%) without requiring dataset-specific tuning.

F Limitations

To resolve the alignment of highly entangled cross-modal distributions, this method (DP-FM) models the alignment process as a continuous multi-step probability flow. Therefore, during the inference phase, the model relies on multi-step integration to progressively transport features, which differs from traditional parameter-efficient fine-tuning (PEFT) methods that only require a single-step adjustment. Objectively speaking, this does mean that multiple steps must be executed during runtime. However, this multi-step mechanism is precisely the core foundation that enables the Flow Matching framework to improve alignment and achieve enhanced performance.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the geometric limitations of existing methods and precisely describe the proposed WP-FM and DP-FM frameworks, matching our theoretical analysis and empirical achievements.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our proposed method in Appendix C. Specifically, we address the inherent multi-step nature of the inference process, noting that while it requires multiple integration steps compared to traditional single-step PEFT methods, this mechanism is a necessary and reasonable trade-off for achieving enhanced cross-modal alignment.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The formal proposition regarding the angular truncation error and its complete mathematical proof are provided in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 thoroughly details the network architecture (MAR), hyperparameter settings (*e.g.*, time-shift $s = 0.1$, CFG scale $w = 5.0$, learning rate), and the dataset configurations used.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The source code and running scripts cannot be publicly released at this time due to institutional proprietary restrictions. However, all 11 benchmark datasets used for evaluation are publicly available academic datasets.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The optimizer (AdamW), learning rate, weight decay, specific DP-FM hyperparameters, and standard K-shot few-shot sampling protocols are explicitly specified in Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: Due to the extensive number of benchmarks (11 datasets) and the large number of comparative baselines across different prompt settings, running multiple random seeds to compute error bars is computationally prohibitive. We follow standard practices in VLM few-shot adaptation literature by reporting the average performance.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Detailed compute resources, including specific GPU hardware specifications and exact execution times, are not reported because the experiments were conducted across various heterogeneous internal clusters over an extended period, making precise resource tracking difficult.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our research focuses on improving algorithmic cross-modal alignment methodologies and fully conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: This is foundational algorithmic research focusing on optimizing alignment geometries for VLMs. We do not foresee direct negative societal impacts beyond the broader capabilities inherited from generic foundation models.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper proposes a fine-tuning algorithm for existing models and does not release new high-risk foundation models or scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All pre-trained VLMs (e.g., CLIP) and the 11 established datasets used in our evaluations are properly cited in the references.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: This work introduces a novel algorithmic methodology rather than releasing newly curated datasets or foundation models.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The research does not involve any human subjects or crowdsourcing.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The research does not involve human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the core algorithmic methodology in this work.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.