ADAPT PART OF THE WAY: CLASS TEXT AS A PRIOR FOR FEW-SHOT SELF-SUPERVISED VISION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot classification with vision-language models is dominated by jointly trained encoder pairs such as CLIP, whose shared image-text space allows an image to be compared directly with text. Self-supervised vision encoders such as DINOv3 now give stronger visual representations than the CLIP image tower, but they have no text tower: they are adapted from the support images alone and ignore the class name, a prior about the class that does not depend on which, or how many, support images are available. We introduce TAP (Text-Anchored Partial transport), which supplies this prior to a self-supervised encoder using only few-shot supervision. TAP aligns the class-name embeddings of an independently pre-trained text encoder to the visual space with a closed-form orthogonal alignment estimated from the support prototypes, trains a lightweight residual map from support images onto their aligned class names, and applies the map only part of the way at inference: the query moves a fraction of the way toward the map's output while each class name moves the complementary fraction toward its support mean. Adapting part of the way is what makes the learned map useful. We evaluate TAP at 1 to 16 shots against linear probes and few-shot adapters developed for CLIP, on several self-supervised image encoders and several text encoders. On ImageNet, its distribution-shifted variants and the standard 11-dataset few-shot benchmark, TAP outperforms a linear probe at every shot count, by 8.4 points at 1 shot on ImageNet, with the largest gains in the low-shot regime and under distribution shift, and it does so with frozen encoders, a single short training run and no validation labels. The gain grows with the quality of the text encoder, so a frozen self-supervised vision model keeps improving as text encoders do. Class names, which self-supervised adaptation has so far discarded, are thus a valuable source of accuracy, and TAP delivers it at negligible inference cost.

1 INTRODUCTION

Few-shot image classification with vision-language models is dominated by *jointly trained* encoder pairs, most prominently CLIP (Radford et al., 2021). Because images and class names are embedded into a shared space, classification reduces to comparing an image feature with text prototypes built from prompts. A large family of adapters refines this space from a few labeled images by tuning prompts (Zhou et al., 2022), caching support features (Zhang et al., 2022), or adding small feature and logit modules (Gao et al., 2024; Zhang et al., 2023; Silva-Rodríguez et al., 2024). All of them inherit one constraint. The image encoder and the text encoder come as a fixed pair, and the quality of the classifier is bounded by the quality of that pair.

In parallel, *self-supervised* vision encoders (Oquab et al., 2024; Darcet et al., 2025; Siméoni et al., 2025) need no paired captions, so they train on larger and more varied image collections, and their frozen features now surpass the CLIP image tower on few-shot classification. A linear classifier on frozen features (a *linear probe*) trained on DINOv3-B outperforms the same probe on the CLIP-B/16 image encoder by 7.8 points on 16-shot ImageNet and by 11.3 points under distribution shift. These encoders, however, have no text tower, so standard adaptation uses the support images only (Tian et al., 2020; Chen et al., 2019) and discards the class names. We ask what class names can contribute to a strong self-supervised encoder that was never trained alongside a text encoder, and how they should be used. Concretely, can the semantic geometry of an *independently pre-trained* text encoder

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

be aligned with a vision-only feature space, and does it help where the support images alone fall short, at low shot and under distribution shift?

The closest existing approach is LFA (Ouali et al., 2023), which estimates an orthogonal map between the image and text features of the support pairs and classifies by the nearest class name. Developed for CLIP's own two towers, it improves accuracy there, but on frozen DINOv3 features it does not reach a tuned linear probe at any shot count. Our approach departs from LFA already in how the map is estimated: in closed form from the C class prototypes (an Orthogonal Procrustes problem (Schönemann, 1966; Maiorca et al., 2023)), which places the class names in the visual space with the distances between them preserved and needs a single image per class. The aligned class names then act as a prior. The text encoder fixes how the classes are arranged relative to one another, and the support images fix only where that arrangement sits. At 1 shot, assigning a query image to its nearest aligned class name outperforms the linear probe by more than 5 points on ImageNet and under distribution shift; at 16 shots the class names alone fall behind, but moving each part of the way toward its class support mean, still in closed form, restores the lead under shift with no trained parameters. The gain is largest where the support images are least informative, at 1 shot and under distribution shift, which is the pattern a text prior should produce.

As the number of support images grows, the aligned class names alone fall behind the linear probe, and the natural next step is to learn a map that carries image features toward their aligned class names and classify by the nearest class name after the move. Used in this way, the map degrades accuracy: at 16 shots every training objective we tried falls short of the linear probe on ImageNet and by a wide margin under distribution shift, where fitting the support set is known to reduce accuracy under shift (Kumar et al., 2022). As our experiments show, the remedy is to apply the map only partially. Moving each query image a fraction τ of the way toward the map's output, and each class name the complementary fraction toward its support mean, keeps part of the text prior in the classifier and outperforms the linear probe both in distribution and under shift over a wide range of τ, for every training objective we tried. The transport fraction, not the training objective, is the design variable that matters.

These observations lead us to a method we term *TAP* (Text-Anchored Partial transport), illustrated in Fig. 1. Class names are first aligned to the visual space with the closed-form orthogonal map (step 1), a small residual map is then trained on the support pairs to carry image features toward their aligned class names (step 2), and at inference the query image and the class names are each transported part of the way and compared (step 3). Both the vision and the text encoder stay frozen, and adaptation is one short training run on cached features. The transport fraction and the weight of the text prior follow a closed-form schedule that requires no validation labels, while a variant that selects them on a validation split is reported alongside throughout.

Contributions. (i) We show that class names from an independently pre-trained text encoder are an effective prior even for a strong self-supervised vision encoder: a closed-form alignment alone already outperforms a linear probe at 1 shot and, with the class names moved part of the way toward their support means, under distribution shift at every shot count. (ii) We identify *partial transport* as the decisive design choice: a map trained to carry images to their class names underperforms the linear probe when applied fully and outperforms it when applied partially, for every training objective we tried. (iii) We introduce TAP, a practical method with frozen encoders, a single short training run and no validation labels, which gains 8.4 points over a linear probe at 1 shot on ImageNet, 3.7 points at 1 shot over eleven datasets and 2.5 points under distribution shift at 16 shots.

2 METHOD

Setup. A few-shot task provides K labeled images per class for C classes, together with the class names. Let f_{img} be a frozen self-supervised vision encoder and f_{txt} a frozen text encoder, of output dimensions D_{img} and D_{txt}. The two were never trained together, so their feature spaces are not comparable. For image $\boldsymbol{s}$ with label $\boldsymbol{l}$ we use unit-normalized features

$$\boldsymbol{x} = \frac{f_{\text{img}}(\boldsymbol{s})}{\|f_{\text{img}}(\boldsymbol{s})\|} \quad \text{and} \quad \boldsymbol{t} = \frac{f_{\text{txt}}(\hat{\boldsymbol{l}})}{\|f_{\text{txt}}(\hat{\boldsymbol{l}})\|}, \tag{1}$$

where $\hat{\boldsymbol{l}}$ is a fixed prompt template containing the class name (Appendix A). A vision-only classifier uses only the image features; our method, TAP (Text-Anchored Partial transport), also uses the C

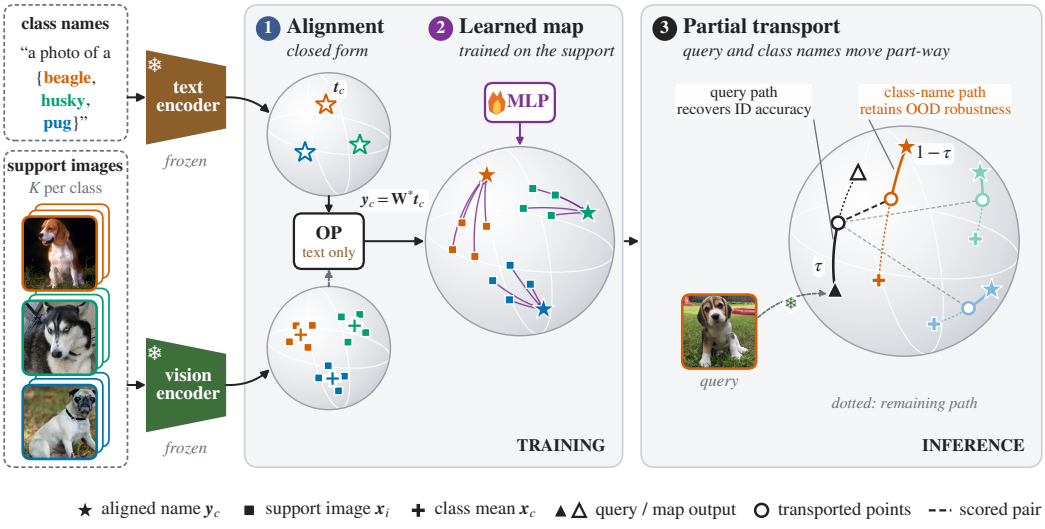


Figure 1: **TAP workflow.** Frozen text and self-supervised vision encoders embed the class names and the K support images in separate spaces. (1) Orthogonal Procrustes (OP) maps each text feature $\boldsymbol{t}_c$ into the visual space in closed form, $\boldsymbol{y}_c = \mathbf{W}^\star \boldsymbol{t}_c$, preserving mutual distances. (2) A residual map, a small MLP, learns on the support to move image features toward the aligned names. (3) At inference the query $\boldsymbol{x}$ moves a fraction τ toward the learned endpoint and each $\boldsymbol{y}_c$ moves $1-\tau$ toward its support mean $\boldsymbol{x}_c$. The cosine of the two is mixed with $\langle \boldsymbol{x}, \boldsymbol{y}_c \rangle$ by α, both scalars read off the support (Sec. 2.3).

class-name features $\boldsymbol{t}_c$, in three components (Fig. 1, Algorithms 1 and 2): a closed-form orthogonal map that places the class names in the visual space (Sec. 2.1), a map learned on the support examples that moves image features toward their class names (Sec. 2.2), and an inference procedure that transports queries and class names only partially along that map (Sec. 2.3).

2.1 Aligning the class names with the visual space

The image features $\boldsymbol{x}$ and the text features $\boldsymbol{t}$ live in different spaces, which may even differ in dimension. Let $\boldsymbol{x}_c$ be the normalized mean of the K support features of class c and $\boldsymbol{t}_c$ the normalized embedding of its prompt, stacked as rows of $\mathbf{X} \in \mathbb{R}^{C \times D_{\text{img}}}$ and $\mathbf{T} \in \mathbb{R}^{C \times D_{\text{txt}}}$. We estimate the semi-orthogonal map that best lays the text prototypes onto the image prototypes,

$$\mathbf{W}^\star = \underset{\mathbf{W} \in \mathbb{R}^{D_{\text{img}} \times D_{\text{txt}}}}{\arg\min} \left\| \mathbf{T}\mathbf{W}^\top - \mathbf{X} \right\|_F^2 \quad \text{s.t.} \quad \mathbf{W}^\top \mathbf{W} = \mathbf{I}_{D_{\text{txt}}}, \tag{2}$$

whose solution is $\mathbf{W}^\star = \mathbf{V}\mathbf{U}^\top$ for the thin SVD $\mathbf{X}^\top \mathbf{T} = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top$ (Schönemann, 1966). Procrustes matches the dimensions of the two networks and, being an isometric embedding for $D_{\text{txt}} \leq D_{\text{img}}$ (the case for all but two encoder pairs in the paper; Appendix F), carries the class names into the visual space with the distances between them preserved. Unlike LFA (Ouali et al., 2023), which estimates it on all $K \cdot C$ image-text pairs, we estimate it on the C prototype pairs. Image features are left unchanged, and every text feature is mapped and renormalized into an *aligned class name*,

$$\boldsymbol{y} = \frac{\mathbf{W}^\star \boldsymbol{t}}{\|\mathbf{W}^\star \boldsymbol{t}\|_2} \in \mathbb{R}^{D_{\text{img}}}. \tag{3}$$

The aligned similarity $\langle \boldsymbol{x}, \boldsymbol{y}_c \rangle$ can already serve as a classifier, but an orthogonal map can only match the class means. The class names land far from the images they describe, as Fig. 2a shows, and at 16 shots the classifier is 7.7 points behind a linear probe (Table 4). We therefore treat the alignment as the starting point of what follows.

2.2 Learning a map from images toward their class names

After alignment, image features and class names lie on the same unit sphere $S^{D_{\text{img}}-1}$. From the $K \cdot C$ support pairs we learn a map that pulls each image toward the name of its class:

$$\hat{\boldsymbol{x}}_1(\boldsymbol{x}) = \frac{\boldsymbol{x} + f_\theta(\boldsymbol{x})}{\|\boldsymbol{x} + f_\theta(\boldsymbol{x})\|_2}, \qquad \mathcal{L}(\theta) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}) \sim \Omega} \left[1 - \langle \hat{\boldsymbol{x}}_1(\boldsymbol{x}),\, \boldsymbol{y} \rangle \right], \tag{4}$$

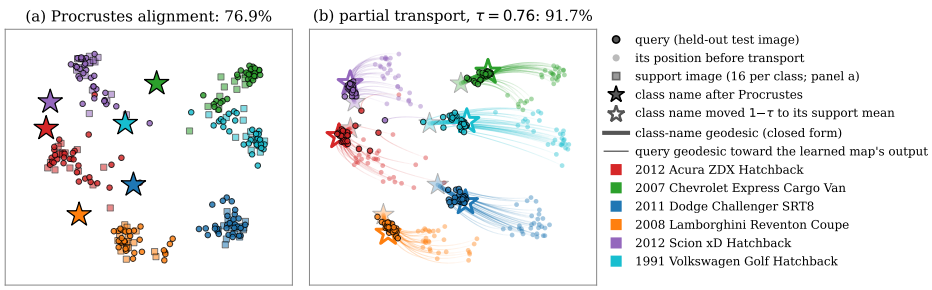


Figure 2: **Alignment alone (a) and partial transport (b)** on 16-shot StanfordCars, frozen DINOv3-B image features and CLIP-B/16 class names, 6 of 196 classes in one fixed PCA projection. (a) Procrustes places the class names (stars) in the image space but far from their images: 76.9% accuracy by nearest class name. (b) Each query moves a fraction τ toward the learned map's output (thin lines) and each class name the complementary fraction toward its class support mean (thick lines), at the scheduled τ=0.76: 91.7% accuracy, against 90.9% for a validated linear probe.

where f_θ is a residual MLP whose output layer is zero-initialized, so that $\hat{\boldsymbol{x}}_1 = \boldsymbol{x}$ before training. At that point TAP leaves every query untouched and only moves each class name toward its support mean (Sec. 2.3), the *identity map* row of the tables. The loss is cosine regression of each support image onto its aligned class name, the simplest objective that suffices. In Sec. 3.3 we test alternative objectives, cross-entropy over the aligned class names and flow matching (Lipman et al., 2023; Chen & Lipman, 2023), and find that, applied partially as described next, all reach accuracies within 1 point of one another.

2.3 Partial transport along the learned map

Applying the map fully degrades accuracy (Sec. 3.3), while not applying it at all reduces the classifier to the aligned class names on their class means. TAP takes an intermediate point along the map, on both sides. Let $\gamma(\boldsymbol{a}, \boldsymbol{b}; t)$ be the spherical geodesic from $\boldsymbol{a}$ to $\boldsymbol{b}$ at fraction $t \in [0, 1]$ (the slerp). For a query $\boldsymbol{x}$ and a transport fraction $\tau \in [0, 1]$,

$$\boldsymbol{z}^{(\tau)} = \gamma\big(\boldsymbol{x}, \hat{\boldsymbol{x}}_1(\boldsymbol{x}); \tau\big), \qquad \boldsymbol{z}_c^{(1-\tau)} = \gamma\big(\boldsymbol{y}_c, \boldsymbol{x}_c; 1-\tau\big) : \tag{5}$$

the query travels a fraction τ toward where the map would send it, and each class name travels the complementary fraction toward its class support mean $\boldsymbol{x}_c$; both moves are closed-form expressions (Fig. 2b). The transported score is their cosine similarity,

$$s_c^{\text{TAP}} = \left\langle \boldsymbol{z}^{(\tau)}, \boldsymbol{z}_c^{(1-\tau)} \right\rangle, \tag{6}$$

which, following common practice (Zhang et al., 2022; Silva-Rodríguez et al., 2024), is mixed with the closed-form similarity it started from:

$$s_c = (1-\alpha)\, s_c^{\text{TAP}} + \alpha\, \langle \boldsymbol{x}, \boldsymbol{y}_c \rangle. \tag{7}$$

Fig. 4 in Sec. 3.3 traces accuracy as a function of τ. Partial transport combines two sources of accuracy. The class names, moved toward their class means, retain the accuracy under distribution shift of the closed-form prior, and the query, moved toward the map's output, adds the in-distribution accuracy of a learned classifier; Sec. 3.3 separates the two contributions and shows that they hold for every training objective we tested. Partial transport is the feature-space counterpart of interpolating fine-tuned and zero-shot weights (Wortsman et al., 2022), applied to both the query and the class names.

Determining α and τ without held-out data. Selecting the two scalars on a validation split costs labeled images, the resource few-shot adaptation is short of, so we set both from the support alone. Let ρ be the *alignment cosine*, the mean cosine between each aligned class name and its class support mean, $\rho = \frac{1}{C}\sum_c \langle \boldsymbol{y}_c, \boldsymbol{x}_c \rangle$, a by-product of Eq. 2 that measures how far the alignment alone takes the names toward the images. We set

$$\alpha = \min\big(0.6,\ \tfrac{1}{K}\big), \qquad \tau = \rho + (1-\rho)\,\alpha. \tag{8}$$

The two scalars respond to different quantities, α to the number of shots and τ to the quality of the alignment. The prior carries weight α in inverse proportion to the shots, 0.6 at 1 shot and $1/16$ at

16. The query travels as far as the alignment already agrees with the images plus a share α of what is left. Hence τ is largest when K is smallest, which is when the complementary fraction $1 - \tau$ that moves each class name toward its class mean should be smallest, since at 1 shot that mean is a single image (Table 4). As K grows the schedule decays to ρ. Every TAP result in the comparison tables is reported under this *closed-form schedule* and, beside it, with (α, τ) *validated* on the split the baselines use, which shows the effect of selecting the two scalars on validation labels. The schedule transfers unchanged to three further image encoders and three further text encoders (Sec. 3.4).

Finally, when a classifier on the same features is at hand, its log posterior and that of TAP (the softmax of Eq. 7 at the text encoder's logit scale) can be averaged at equal weight; the two make different errors, and Sec. 3.1 shows that this combination is the strongest row at every shot count.

3 Experiments

Setup. Both encoders are frozen throughout, DINOv3-B/16 (Siméoni et al., 2025) for images and the CLIP-B/16 text tower (Radford et al., 2021) for class names, except in Sec. 3.4. We follow the CoOp protocol (Zhou et al., 2022) with $K \in \{1, 2, 4, 8, 16\}$ support images per class on its eleven datasets (listed in Appendix A), official test splits, 3 seeds and the same support for every method. TAP's map is a two-block residual MLP of width 1536 (11.8M parameters) trained with Adam at $2 \cdot 10^{-4}$. Accuracy under distribution shift is measured by adapting on K-shot ImageNet and evaluating, unchanged, on ImageNet-V2, -Sketch, -A and -R (Recht et al., 2019; Wang et al., 2019; Hendrycks et al., 2021b;a); their mean is the *OOD average*.

Baselines with a hyper-parameter select it on the CoOp validation split of $\min(K, 4)$ extra images per class, spending $K + \min(K, 4)$ labels, whereas TAP under the schedule spends K; TAP is also reported *validated* on that split. The vision-only baselines are cosine centroids, an ℓ_2-regularized linear probe and GDA (Wang et al., 2024). The text-anchored baselines are the Procrustes classifier of Sec. 2.1, LFA (Ouali et al., 2023) as released (Appendix A), a Tip-Adapter cache (Zhang et al., 2022) on the Procrustes score, and a Cross-modal probe (Lin et al., 2023), which adds each aligned class name to the support as an extra sample. Two further rows are ours: the Prior mixture, a log-linear blend of the probe's and the aligned-text posteriors with weight α from Eq. 8, and TAP with the identity map ($f_\theta \equiv 0$).

3.1 Few-shot classification on eleven datasets

Table 1 reports the mean over the eleven datasets of the CoOp benchmark at 1 to 16 shots, in three blocks: vision-only classifiers, baselines that use the class names through the Procrustes alignment, and TAP.

Table 1: **Few-shot accuracy on eleven datasets by number of shots.** Mean over the 11 CoOp datasets, frozen DINOv3-B features, CLIP-B/16 text, 3 seeds (seed std of the mean 0.1 to 1.1). The val. column marks rows that select hyper-parameters on a $\min(K, 4)$-image validation split and so spend $K + \min(K, 4)$ labels. Text rows use the Procrustes-aligned class names. Bold: best per column among CLIP-text rows; the grey row swaps in SigLIP2 text. Further rows, per-dataset results and paired gains: Appendix Tables 6, 19 and 20.

method	text	val.	1	2	4	8	16
Linear probe	–	–	59.99	67.50	74.17	79.40	83.39
Linear probe	–	✓	60.25	70.14	78.11	82.78	85.85
GDA	–	✓	59.69	70.19	78.22	82.61	85.24
LFA	CLIP	–	57.77	68.55	76.04	80.92	84.08
Tip-Adapter cache	CLIP	✓	62.42	69.86	75.75	79.47	82.17
Prior mixture	CLIP	✓	62.30	70.27	77.98	82.77	85.87
Cross-modal probe	CLIP	–	60.79	68.29	74.89	79.70	83.50
Cross-modal probe	CLIP	✓	62.71	71.58	78.59	82.83	85.90
TAP	CLIP	–	63.98	72.21	79.04	83.00	85.96
TAP	CLIP	✓	64.08	72.32	79.22	83.32	86.18
TAP + Cross-modal probe	CLIP	✓	**64.17**	**72.55**	**79.41**	**83.62**	**86.46**
TAP	SigLIP2	–	64.94	72.80	79.33	83.31	86.14

The Prior mixture row blends the aligned-name score into the validated probe and trains nothing more. It is 2 points ahead at 1 shot and indistinguishable from the probe afterwards. The class text

dominates the decision when a single image per class is the only visual evidence, and is outweighed as labeled images accumulate. The learned map extends the range of shot counts over which the prior helps. TAP under the schedule is above the validated probe at every shot count; the standard deviation of the per-seed paired difference is at most 0.5, so the gains through 4 shots are several standard deviations wide and those at 8 and 16 are within seed noise. Against the untuned probe, which spends the same K labels, it is ahead by 2.6 to 4.9 points at every shot count. At 1 shot it is also above every text-using baseline while spending fewer labels, and validated (α, τ) change it by a few tenths of a point. The gain over the validated probe fades with K, as expected of a prior, which indicates that TAP uses the class text as one such prior: its contribution is largest when the support carries little information about the class and decreases as the labeled evidence grows. Two controls confirm that the gain comes from the names and from their alignment: assigning the names to the wrong classes removes it, and replacing the Procrustes solution by a fixed random semi-orthogonal map costs 6.9 points at 1 shot, as Appendix Tables 9 and 10 show. At 16 shots the class names still contribute, but to accuracy under distribution shift rather than in distribution, as Sec. 3.2 shows. Swapping CLIP's text tower for SigLIP2's (grey row) adds a further point at 1 shot under the same schedule.

Per dataset (Appendix Tables 19 and 20), TAP's gain persists to 16 shots on Cars, SUN397 and ImageNet and turns negative on DTD and Aircraft, where the probe is stronger at higher shot counts (Sec. 5). Averaging TAP's log posterior with the Cross-modal probe's, two classifiers that make different errors, is the best row of Table 1 at every shot count.

3.2 IMAGENET AND ITS DISTRIBUTION SHIFTS

Table 2 adapts on K-shot ImageNet and evaluates on ImageNet and, unchanged, on its four shifts, at 1, 4 and 16 shots (all shot counts: Appendix Table 8), and Fig. 3 plots each method's 16-shot result as a point in the plane of ImageNet accuracy against the OOD average. The two axes separate two properties that a single accuracy conflates: how well a classifier fits the source distribution and how well it transfers under distribution shift. The two linear-probe rows of Table 2 already show the trade-off: selecting the regularization strength on four validation images per class raises ImageNet accuracy and lowers the OOD average (right and down in Fig. 3), and GDA and LFA lose even more under shift without gaining in distribution. The rows that use the class names, in contrast, gain under distribution shift, and every one of them except LFA is above the validated probe's OOD average at every shot count.

Table 2: **ImageNet and its distribution shifts by number of shots.** Frozen DINOv3-B features; left ImageNet top-1, right the average over ImageNet-V2, -Sketch, -A and -R; 3 seeds, rows and "val." as in Table 1. Bold: best per column among CLIP-text rows; grey rows swap in SigLIP2 text. Per-shift breakdown: Appendix Table 7.

			ImageNet			OOD average		
method	text	val.	1	4	16	1	4	16
Linear probe	–	–	47.42	64.40	74.43	39.97	54.39	61.58
Linear probe	–	✓	47.45	67.57	75.91	39.98	53.90	59.90
GDA	–	✓	47.42	67.63	74.91	40.10	51.66	56.83
LFA	CLIP	–	49.84	65.10	72.82	39.83	50.81	56.51
Tip-Adapter cache	CLIP	✓	53.67	67.79	74.13	46.33	57.43	61.92
Prior mixture	CLIP	✓	53.35	69.26	76.35	46.21	57.57	61.17
Cross-modal probe	CLIP	–	49.88	65.44	74.69	42.60	55.42	61.79
Cross-modal probe	CLIP	✓	55.00	70.21	76.39	46.68	57.62	61.51
TAP	CLIP	–	55.90	70.28	76.96	**48.48**	58.11	62.44
TAP	CLIP	✓	56.04	70.49	77.05	47.93	58.74	62.63
TAP + Cross-modal probe	CLIP	✓	**56.23**	**70.61**	**77.21**	48.43	**58.92**	**63.02**
TAP	SigLIP2	–	57.66	70.85	76.94	50.87	59.31	63.21
TAP + Cross-modal probe	SigLIP2	✓	58.54	71.76	77.40	50.95	60.26	63.60

TAP is above the validated probe on both ImageNet and the OOD average at every shot count, by $+8.4$ / $+8.5$ points at 1 shot and $+1.1$ / $+2.5$ at 16, its OOD average is above every baseline's at every shot count, and at 16 shots TAP and its combination are the only rows above both probe settings on both axes, as Fig. 3 shows. The in-distribution margin decreases with K as on the eleven-dataset mean, while the margin under shift does not close: at 16 shots what the class names add is coverage of the test distribution, which more labeled images from the source distribution do not supply. The gains are largest where the support images are least informative, at 1 shot and away from the source

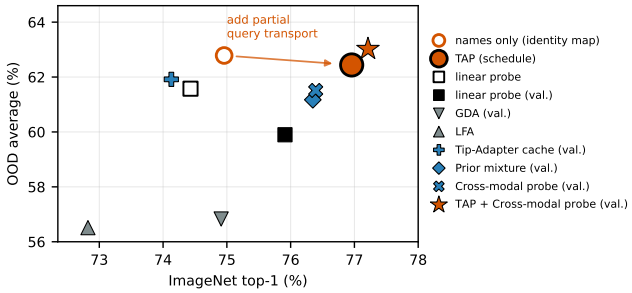


Figure 3: **In-distribution against out-of-distribution accuracy at 16 shots.** Each method of Table 2 as a point, 3 seeds, same support; hollow and filled squares are the untuned and validated linear probe. One shot: Appendix Fig. 6.

distribution, which is the pattern a text prior should produce and the evidence that TAP uses the class names as such a prior.

Finally, the two greyed-out rows swap CLIP's text tower for SigLIP2's with the same frozen vision model and the same schedule. Every 1-shot number and every OOD average rises, so the strength of the prior follows the quality of the text encoder, and Sec. 3.4 varies both encoders.

3.3 Partial versus full transport

Fig. 4 traces accuracy along the transport fraction τ on 16-shot ImageNet. Both curves rise above the linear probe over a wide interior range, ImageNet accuracy peaking near $\tau = 0.6$ and the OOD average near $\tau = 0.3$, and both fall below it as τ approaches one, while the schedule, set from the support alone, lands between the two optima. Doubling the map's width leaves the curves essentially unchanged (Appendix Table 17).

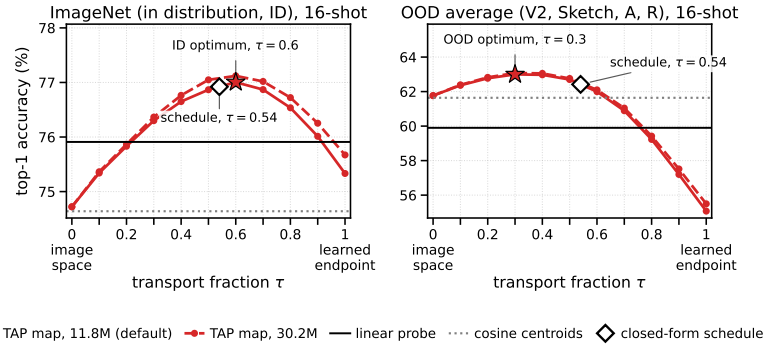


Figure 4: **Accuracy along the transport fraction τ on 16-shot ImageNet.** ImageNet top-1 (in distribution, ID) and OOD average for TAP's map at width 1536 (solid, 11.8M parameters) and 2560 (dashed); 3-seed means (seed std at most 0.22 at every τ), α=1/16. At τ=0 the query is untouched and the class names sit on their support means; at τ=1 the query is at the learned endpoint and the class names are the raw aligned text. Horizontal lines: the validated linear probe and cosine centroids. Stars mark the test-grid optima, the diamond the schedule's τ=0.54, set from the support alone; both curves are above both baselines for $\tau \in [0.22, 0.63]$. One shot: Appendix Fig. 7.

Table 3 trains the same residual architecture with four objectives: cosine regression, cross-entropy over the aligned class names, and flow matching along geodesics or straight lines. It scores each network twice, at the schedule's τ (partial transport) and at $\tau = 1$ (full transport). Used fully, no objective reaches the validated probe on either axis, and all are far below it under distribution shift. Stopped part-way, every one is above the probe on both, with a gain under shift of 7 to 13 points over its own endpoint. The four objectives differ by less than a point at the schedule's τ against an 8-point swing along τ in Fig. 4, and neither training length nor the interpolation path is the cause, as Appendix Tables 16 and 15 show. The transport fraction, not the objective, is the design variable that matters.

Table 3: **Partial against full transport for four training objectives.** 16-shot ImageNet, ImageNet top-1 / OOD average, 3 seeds (±: seed std). One residual architecture trained on the same support with each objective and scored twice: *partial transport* is TAP's inference under the schedule (τ=0.54, α=1/16), *full transport* the same network used as an ordinary adapter (τ=1, α=0); Δ_{OOD} is their difference on the OOD average. The first row has no learned map. Widths, learning rates and per-shift results: Appendix D.

		partial transport		full transport		
training objective	params (M)	ImageNet	OOD avg.	ImageNet	OOD avg.	Δ_{OOD}
none (identity map)	0	74.96	62.79 ± 0.34	68.20	57.17 ± 0.28	+5.6
cosine regression (TAP)	11.8	76.96	62.44 ± 0.14	74.97	54.01 ± 0.12	+8.4
cross-entropy on names	11.8	76.16	62.82 ± 0.41	72.91	55.75 ± 0.10	+7.1
flow matching, geodesic	15.7	76.48	62.69 ± 0.11	73.79	53.34 ± 0.15	+9.4
flow matching, straight line	15.7	76.12	62.67 ± 0.18	70.07	49.48 ± 0.40	+13.2
cosine regression	30.2	77.10	62.54 ± 0.22	75.33	54.50 ± 0.12	+8.0
cross-entropy on names	30.2	76.20	62.91 ± 0.16	73.12	56.14 ± 0.14	+6.8

Table 4: **Decomposition of the transport: class-name side, query side, both.** Same checkpoints, schedule's α: τ_{img} is the fraction the query travels toward the learned endpoint, τ_{proto} the fraction each class name travels toward its class mean. ImageNet top-1 / OOD average, 3 seeds; bold best, underlined second best per column.

			1 shot		16 shots	
variant	τ_{img}	τ_{proto}	ImageNet	OOD	ImageNet	OOD
Procrustes only (aligned names)	0	0	52.97	45.85	68.20	57.17
names only (identity map)	0	$1-\tau$	53.92	46.95	74.96	**62.79**
query only	τ	0	**56.06**	<u>48.18</u>	<u>76.45</u>	61.23
both (TAP)	τ	$1-\tau$	55.90	**48.48**	**76.96**	<u>62.44</u>
learned endpoint, same α	1	0	<u>56.03</u>	47.81	75.33	55.07
schedule τ (ImageNet)			0.76		0.54	

Table 4 switches each side of the transport on alone, on the same checkpoints, at 1 and 16 shots. At 16 shots moving only the names gives the best OOD average in the table and stays below the probe on ImageNet, moving only the query does the reverse, and moving both is best on ImageNet and within half a point of the best under shift. This locates the two gains of Sec. 3.2: at 16 shots the gain under distribution shift comes from the class names, and what the learned map adds is the in-distribution accuracy that takes the prior above the probe. At 1 shot the query side accounts for nearly all of the gain and moving the names adds a fraction of a point. Letting the two fractions vary independently gains at most 0.2 points, see Appendix Fig. 5.

3.4 Other image and text encoders

Appendix Table 11 repeats the ImageNet protocol at 1 and 16 shots, changing one encoder at a time (top block: image encoder, CLIP text fixed; bottom block: text encoder, DINOv3-B fixed). TAP uses the schedule with each pair's own alignment cosine, and the probe is validated as before.

For every image encoder TAP is above the probe on both axes at both shot counts, with the same shape as on DINOv3-B, largest at 1 shot and under shift. DINOv3-L gains slightly less than DINOv3-B and DINOv2-L least, while CAPI-L, trained by masked image modelling alone and with a much weaker probe, gains most, 5 points at 1 shot and 3 to 4 at 16. On all three the schedule is within 0.35 points in distribution and 0.8 under shift of validated selection (Appendix Table 12).

The text encoder determines how strong the prior is. The two encoders trained against images, SigLIP2 and CLIP, are 8 to 10 points above the probe at 1 shot, whereas the two trained on text alone, Qwen3-Embedding and mini-GTE, are within a point of it (validated results: Appendix Table 13). At 16 shots all four are above the probe, mostly under distribution shift, and within half a point of one another in distribution. The prior is thus a property of the encoder pair, as the SigLIP2 rows of Table 2 also showed.

Cost. TAP's adaptation is a 768×512 SVD (22 ms) plus one training run of the map (55 s) with no hyper-parameter search, against 1.6 s for the validated probe and 8.7 s for the validated Cross-modal probe; inference adds 30 MFLOPs per image, 0.08% of the encoder's FLOPs (Appendix Table 5).

4 Related work

Few-shot classifiers on frozen features. A simple classifier on a strong frozen embedding is hard to beat (Chen et al., 2019; Tian et al., 2020), the more so with self-supervised encoders (Oquab et al., 2024; Siméoni et al., 2025) and the self-distillation objectives behind them (Caron et al., 2021; Shalam & Korman, 2024), and Wang et al. (2024) make the same observation for CLIP. CLIP adapters keep both towers frozen and modify prompts, caches, features or logits (Zhou et al., 2022; Zhu et al., 2023; Zhang et al., 2022; Udandarao et al., 2023; Gao et al., 2024; Zhang et al., 2023; Silva-Rodríguez et al., 2024; Zhu et al., 2024; Li et al., 2023b; 2025b); Lin et al. (2023) treat the class text as an extra training sample of its class, and LP++ (Huang et al., 2024) blends a linear probe with CLIP's text logits. All of them need image and text features that are already comparable, which ties them to CLIP's image tower. TAP gives a self-supervised encoder the same text score through the closed-form alignment of Sec. 2.1, and the adapters that carry over are the baselines of Table 1.

Aligning encoders that were not trained together. Linear maps between separately trained embedding spaces go back to cross-lingual word vectors (Mikolov et al., 2013) and DeViSE (Frome et al., 2013). Relative representations, latent-space translation and ASIF (Moschella et al., 2023; Maiorca et al., 2023; Norelli et al., 2023) connect unimodal encoders through paired examples, Text-to-Concept (Moayeri et al., 2023) maps vision encoders linearly into CLIP space, and alignability grows with scale (Huh et al., 2024). LFA (Ouali et al., 2023) is the closest to our alignment step: a Procrustes map on paired image–text features, mainly between CLIP's own two towers and, through a fixed random projection, from self-supervised encoders to CLIP text. Secs. 2.1 and 3.1 give the two differences, prototype-level estimate and partial use. Projectors trained on millions of captions (Maniparambil et al., 2025; Merullo et al., 2023; Li et al., 2023a) and text towers trained against a locked image encoder (Zhai et al., 2022; Jose et al., 2025; Barsellotti et al., 2025) solve a related problem at a data scale that K images per class do not reach (Kumar et al., 2022).

Interpolating between a learned solution and a prior. Kumar et al. (2022) show that fine-tuning distorts pretrained features and can fall below a linear probe out of distribution, and WiSE-FT (Wortsman et al., 2022) repairs this by averaging fine-tuned and zero-shot weights, with the coefficient chosen on held-out data. Partial transport is the counterpart for a frozen encoder, in feature space rather than weight space: the learned solution is the map's endpoint, the prior is the aligned class name, and the operating point is a fraction τ along the geodesic between them, on both sides, read off the support. Geodesic mixup (Oh et al., 2023) interpolates image and text embeddings on the sphere to augment CLIP fine-tuning. Sec. 3.3 asks the same question of adapters on frozen features and finds the same failure at full strength, for cross-entropy as for flow matching (Lipman et al., 2023; Liu et al., 2023; Chen & Lipman, 2023; Esser et al., 2024; Liu et al., 2025; Li et al., 2025a).

5 Conclusion

Self-supervised vision encoders give stronger frozen few-shot features than the CLIP image tower, but, having no text tower, their adaptation discards the class names. An orthogonal alignment estimated from the support prototypes places the class names of an independently pre-trained text encoder in the visual space, where they act as a prior: the gain over a linear probe is largest at 1 and 2 shots and under distribution shift, and fades as labeled images accumulate. A map trained to carry images toward their class names extends this gain further, provided it is applied only part of the way: the transport fraction, not the training objective, is the design variable that matters, and it can be read off the support. The result holds across four image and four text encoders, and a stronger text encoder improves the classifier without any change to the vision model. Class names are thus a reliable source of few-shot accuracy for self-supervised encoders, and TAP delivers it with one short training run and no validation labels.

Limitations. At 1 and 2 shots, published CLIP adapters remain ahead on the eleven-dataset mean (Appendix Table 18); they start from CLIP's zero-shot classifier, which TAP's vision-only encoder lacks, and from 4 shots TAP is ahead of all of them. The 16-shot gain on that mean is within seed noise, as expected once labels suffice, while the gain under distribution shift remains. On DTD and Aircraft, TAP under the schedule falls behind the probe at higher shot counts, and averaging with a probe closes most of the gap. Our experiments cover vision transformers and English class names.

AI USE STATEMENT

In this work, we used generative AI tools for research execution and for writing. An AI coding assistant (Claude, via the Kiro CLI) was used, under the authors' direction, to write and run experiment, evaluation and analysis scripts, to aggregate results into tables and figures, and to draft and edit text; all experimental settings were specified by the authors and all reported numbers were re-derived from the raw run outputs and checked by the authors. Generative AI was not used to design the method, to produce or alter experimental results, or to generate reviews of related work; the ideas, claims and conclusions are the authors' own. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of AI tools.

ETHICS STATEMENT

This work studies few-shot image classification with publicly available pretrained encoders and public benchmark datasets. It does not involve human subjects, private data, or new data collection. Text-anchored classifiers inherit the biases of both the vision and the text encoder they are built from, and because TAP admits vision and text encoders that were never pretrained as a pair, the resulting combination has not been audited by anyone; practitioners should audit it for the deployment domain rather than assume the properties of either encoder alone.

REPRODUCIBILITY STATEMENT

All encoders are frozen, publicly released checkpoints (Appendix A lists the exact identifiers). Few-shot splits follow the CoOp protocol; the validation split used by the baselines, the hyperparameter grids, and TAP's closed-form schedule for its two inference scalars (Eq. 8) are given in Sec. 2.3 and Appendix A. Per-dataset numbers with standard deviations over three seeds are in Appendix E. Code, prompts, cached features and the scripts that produce every table and figure will be released.

REFERENCES

Luca Barsellotti, Lorenzo Bianchi, Nicola Messina, Fabio Carrara, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, and Rita Cucchiara. Talking to DINO: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *International Conference on Learning Representations (ICLR)*, 2023.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations (ICLR)*, 2019.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Timothée Darcet, Federico Baldassarre, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Cluster and predict latent patches for improved masked image modeling. *arXiv preprint arXiv:2502.08769*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Patrick Esser et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

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

Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

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. *International Journal of Computer Vision (IJCV)*, 132(2):581–595, 2024.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (J-STARS)*, 12(7), 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021b.

Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. LP++: A surprisingly strong linear probe for few-shot CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *International Conference on Machine Learning (ICML)*, 2024.

Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, Oriane Siméoni, Huy V. Vo, Patrick Labatut, and Piotr Bojanowski. DINOv2 meets text: A unified framework for image- and pixel-level vision-language alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C.H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, 2023a.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Zichun Liao, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Omniflow: Any-to-any generation with multi-modal rectified flows. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025a.

Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b.

Yayuan Li, Jintao Guo, Lei Qi, Wenbin Li, and Yinghuan Shi. Text and image are mutually beneficial: Enhancing training-free few-shot classification with clip. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025b.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023c.

Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Qihao Liu, Xi Yin, Alan Yuille, Andrew Brown, and Mannat Singh. Flowing from words to pixels: A noise-free framework for cross-modality evolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

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

Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent space translation via semantic alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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

Mayug Maniparambil, Raiymbek Akshulakov, Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Ankit Singh, and Noel E. O'Connor. Harnessing frozen unimodal encoders for flexible multi-modal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29847–29857, 2025. arXiv preprint title: Freeze-Align.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. In *International Conference on Learning Representations (ICLR)*, 2023.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

Mazda Moayeri, Keivan Rezaei, Maziar Sanjabi, and Soheil Feizi. Text-to-concept (and back) via cross-model alignment. In *International Conference on Machine Learning (ICML)*, 2023.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations (ICLR)*, 2023.

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

Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodolà, and Francesco Locatello. ASIF: Coupled data turns unimodal models to multimodal without training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Changdae Oh, Junhyuk So, Hoyoon Byun, YongTaek Lim, Minchul Shin, Jong-June Jeon, and Kyungwoo Song. Geodesic multi-modal mixup for robust fine-tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. Black box few-shot adaptation for vision-language models. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

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

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31:1–10, 1966.

Daniel Shalam and Simon Korman. Unsupervised representation learning by balanced self attention matching. In *European Conference on Computer Vision (ECCV)*. Springer, 2024.

Julio Silva-Rodríguez, Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. A closer look at the few-shot adaptation of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, et al. DINOv3, 2025.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: A good embedding is all you need? In *European Conference on Computer Vision (ECCV)*, 2020.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A hard-to-beat baseline for training-free clip-based adaptation. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *CVPR*, 2022.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18123–18133, 2022.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision (ECCV)*, 2022.

Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022.

Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15659–15669, 2023.

Yuhan Zhu, Yuyang Ji, Zhiyu Zhao, Gangshan Wu, and Limin Wang. Awt: Transferring vision-language models via augmentation, weighting, and transportation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

APPENDIX

A IMPLEMENTATION DETAILS

Datasets. The eleven datasets of the CoOp protocol (Zhou et al., 2022) are ImageNet (Deng et al., 2009), Caltech101 (Fei-Fei et al., 2004), OxfordPets (Parkhi et al., 2012), StanfordCars (Krause et al., 2013), Flowers102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), FGVCAircraft (Maji et al., 2013), SUN397 (Xiao et al., 2010), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019) and UCF101 (Soomro et al., 2012). We take the few-shot support splits, the validation splits, the test splits and the class names from CoOp's released files without modification. The four distribution shifts of Sec. 3.2 are ImageNet-V2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019), ImageNet-A (Hendrycks et al., 2021b) and ImageNet-R (Hendrycks et al., 2021a), evaluated with the class names of the ImageNet classes each of them covers.

Encoders and prompts. Vision, all frozen, CLS token unless stated: DINOv3-B/16 (`facebook/dinov3-vitb16-pretrain-lvd1689m`), DINOv3-L/16 (`facebook/dinov3-vitl16-pretrain-lvd1689m`) and DINOv3-S/16 (Siméoni et al., 2025); DINOv2-L/14 (`facebook/dinov2-large`) (Oquab et al., 2024); CAPI-L/14 (`capi_vitl14_lvd` from `facebookresearch/capi` via torch.hub, its 1024-d global representation) (Darcet et al., 2025); the CLIP image towers of Table 14 from open_clip (`openai` weights). Images are resized to 256 and center-cropped to 224 at test time; training uses random-resized crops of scale $[0.5, 1]$ with horizontal flips. Text, all frozen: CLIP-B/16 and CLIP-L/14 text towers (open_clip, `openai`) (Radford et al., 2021), SigLIP2-B/16 text tower (`ViT-B-16-SigLIP2`, `webli`, open_clip) (Tschannen et al., 2025), Qwen3-Embedding-0.6B (`Qwen/Qwen3-Embedding-0.6B`) (Yang et al., 2025) and mini-GTE (`prdev/mini-gte`) (Li et al., 2023c) from Hugging Face, mean-pooled. Prompts: on ImageNet the class prototype is the mean of CLIP's seven selected ImageNet templates; on the ten other datasets it is the CoOp dataset-specific template ("a photo of a {class}, a type of pet." and so on) (Zhou et al., 2022). The same prototypes feed every text-using method.

TAP's learned map. The residual map f_θ is the residual MLP of Li et al. (2024) without its time conditioning: 2 blocks of width 1536 with `SiLU` activations, input and output width equal to the encoder dimension, and a zero-initialized output layer, so that $\hat{\boldsymbol{x}}_1(\boldsymbol{x}) = \boldsymbol{x}$ before training (Sec. 2.2). It has 11.8M parameters at $D_{\text{img}} = 768$. It is trained on the KC support pairs by cosine regression onto the aligned class names (Eq. 4) with Adam at $\text{lr} = 2\text{e}{-4}$, for 200 epochs on all datasets except ImageNet, which uses 50 epochs, with batch size 32 for $K \leq 4$ and 128 for $K \geq 8$; the flow-matching rows of Table 3 keep the $\text{lr} = 5\text{e}{-5}$ at which they were tuned. Encoders are frozen; the OP map is estimated in closed form on the K-shot support and kept fixed while the map is trained. Nothing is learned on the text side: the class names move along the closed-form geodesic of Eq. 5. α and τ follow Eq. 8: $\alpha = \min(0.6, 1/K)$ and $\tau = \rho + (1 - \rho)\alpha$, with ρ the alignment cosine, the mean cosine between the OP-aligned class names and the normalized support class means, computed once per task from the support; no held-out images are used. The validated rows reported throughout use a grid search ($\alpha \in \{0, 0.1, \ldots, 1\}$, $\tau \in \{0, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, 1\}$) on the validation split of $\min(K, 4)$ images per class. The objective controls of Table 3 share this architecture at widths 1536 and 2560 (11.8M and 30.2M); the flow rows add time conditioning and train at learning rate $5 \cdot 10^{-5}$ against $2 \cdot 10^{-4}$ for the MLPs, and swapping the two learning rates changes no ordering. All other hyperparameters are fixed across datasets and shot counts.

Matched baselines (Sec. 3). All baselines and controls use the same frozen features and the same K-shot support. Baselines with a hyperparameter select it on a validation split of $\min(K, 4)$ labeled images per class, disjoint from the support (the CoOp split; on ImageNet, whose CoOp split is the test set, we draw it from the training set); TAP under the closed-form schedule does not use this split (Eq. 8). *Cosine centroids:* class means of ℓ_2-normalized support features, prediction by cosine similarity. *Linear probe:* multinomial logistic regression on ℓ_2-normalized features, ℓ_2 penalty C selected on the validation split from $\{10^{-2}, 10^{-1}, \ldots, 10^{3}\}$ (log-spaced, 6 values); features are the frozen, ℓ_2-normalized support features. *Linear probe (fixed):* the same with $C{=}1$ and no selection. We use $C{=}1$, the unselected default, rather than a value such as $C{=}10$ that happens to work well on

these features: any other fixed choice would itself be a selection made on test accuracy. The untuned probe is sensitive to it at high shot, C=10 reaching 85.63 against C=1's 83.39 on the eleven-dataset mean at sixteen shots (Table 18), and insensitive at low shot (60.11 against 59.99 at one shot). Every comparison in the main text is made against the *validated* probe, which does not depend on this choice. *GDA:* class means and a shared within-class covariance Σ estimated from the support, shrunk toward the isotropic matrix, $(1-\lambda)\Sigma + \lambda\,\mathrm{tr}(\Sigma)/D \cdot I$, with $\lambda \in \{0.05, 0.1, 0.2, 0.5, 0.8, 0.95\}$ selected on the validation split; prediction by the Gaussian discriminant $\boldsymbol{x}^\top \Sigma^{-1}\boldsymbol{\mu}_c - \frac{1}{2}\boldsymbol{\mu}_c^\top \Sigma^{-1}\boldsymbol{\mu}_c$; at one shot there is no within-class covariance and we set $\Sigma = I$, which is the cosine-centroid classifier. *Prior mixture (ours, Sec. 3):* the log-softmax of the probe's logits (with C already selected as above) mixed with the log-softmax of $s\,\langle \boldsymbol{x}, \boldsymbol{y}_c \rangle$, where $\boldsymbol{y}_c$ are the OP-aligned prototypes (closed form from the same support) and s is the frozen text encoder's pretrained logit scale (exp of CLIP's `logit_scale`: 100 for CLIP-B/16), not a tuned or learned quantity; α is $\min(0.6, 1/K)$ as for TAP (Eq. 8). The control therefore trains nothing and uses no held-out data beyond the probe's C. TAP's own loss is the cosine regression of Eq. 4 only; no cross-entropy or contrastive term is used. *Tip-Adapter cache (Zhang et al., 2022):* keys are the KC normalized support features, values their one-hot labels; the logit is the aligned text score $s\,\langle \boldsymbol{x}, \boldsymbol{y}_c \rangle$ plus α_t times the affinity $\exp(-\beta(1-\cos))$ summed over the cache, with $(\alpha_t, \beta) \in \{0.1, 0.3, 0.5, 1, 2, 4, 8, 16, 32, 64\} \times \{1, 2, 3, 4, 5.5, 7, 10\}$ selected on the validation split; nothing is fine-tuned. *Cross-modal probe (Lin et al., 2023):* the linear probe above trained on the KC support features plus the C aligned class names as one extra sample per class with sample weight w; the untuned row uses w=1, C=1, the validated row selects $(w, C) \in \{1, 2, 4, 8, 16\} \times \{0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 100\}$ on the validation split. *LFA (Ouali et al., 2023):* the released code (`feature_alignment.py`) applied to our features. LFA's own procedure for uni-modal encoders of unequal dimension is a fixed Gaussian random projection of the image features to the text dimension before alignment (its `gaussian_projection` step, used for its BYOL, Barlow Twins and MoCo v3 experiments); we follow it ($768 \to 512$). LFA then fits a closed-form Procrustes map on all KC one-to-one image–text pairs (`get_one_to_one_features`) and refines it for 200 steps of its re-ranking loss (Adam, lr $5 \cdot 10^{-4}$, β chosen by its own cross-validation); prediction is the cosine between the mapped query and the raw names. Both steps are LFA's, and they, not the prototype-level fit, separate LFA from our Procrustes-only row at one shot (49.84 against 52.97 on ImageNet).

Compute. Experiments were run on single NVIDIA A100 and L40S GPUs; the timings of Table 5 are on an L40S. Metrics are top-1 accuracy, reported as the mean over 3 seeds; seeds change the support/validation sample and the initialization of the learned map.

B Computational cost and scalability

Table 5 reports the measured adaptation and inference costs of every method of Table 2 on one L40S, all of them after the test features have been encoded once (Sec. 3.4). The paragraphs below read those numbers as functions of the number of classes C and the number of shots K.

Adaptation. Of TAP's 57 s, 55 s are the training of the residual map; the closed-form part is one text forward pass over the C class names (1.9 s) and the SVD of Eq. 2 (22 ms). Both scale with C alone: C short prompts to encode, one thin SVD of a $D_{\text{txt}} \times D_{\text{img}}$ matrix, and C class means to average. Together they are of the same order as fitting a linear probe with validation (1.6 s), which is why the rows that need nothing learned, the Procrustes-only classifier, the prior mixture and centroids with the aligned score, are as cheap to adapt as the cheapest baseline and are the whole of TAP at τ=0. Training the map is what separates TAP from those fractions of a second: its cost is linear in the number of support features and in the number of epochs, and it is unaffected by the closed-form schedule. Selecting (α, τ) on validation instead adds one scoring pass over the validation features per grid point, on top of the same training.

Inference. The head costs 9.2 μs and 30 MFLOPs per image, that is 0.08% of the FLOPs and 1.3% of the latency of the DINOv3-B forward pass that produced the feature. The two residual blocks account for almost all of it: the two geodesics and the product with the C transported names are the cost of a cosine classifier, so the per-image cost is flat in K and grows with C only in the final $C \times D_{\text{img}}$ product, as for a linear probe. In FLOPs this is about 20 times a probe (1.5 MFLOPs) and about half a Tip-Adapter cache (58 MFLOPs), whose cost, unlike TAP's, grows with KC because

Table 5: **Cost after feature caching**, ImageNet 16-shot, DINOv3-B features, 1000 classes, one NVIDIA L40S; median of repeated timings after a warm-up. FLOPs count multiplications and additions separately (2 per multiply-add); Table 14 lists multiply-adds. Adaptation: one fit, and as validated in Table 2 (the probe's C grid, Tip-Adapter's (α_t, β) grid, the cross-modal probe's (w, C) grid; TAP's validated (α, τ) grid adds 0.09 s). Methods that use class text share a one-off 1.9 s pass of the 1000 names through the CLIP text tower, counted in the TAP rows and not in the other text rows. Inference: cost per image on top of the shared encoder forward (batch 256); the encoder itself is given for reference. Memory: parameters and cached tensors of the head.

	adaptation (s)		inference, per image		memory
method	one fit	validated	μs	MFLOPs	MB
Linear probe	0.12	1.6	0.16	1.5	2.9
GDA	0.003	–	0.16	1.5	2.9
Tip-Adapter cache	0	0.06	2.5	58	50
LFA	3.0	–	0.6	2.3	3.0
Cross-modal probe	0.10	8.7	0.17	1.5	2.9
TAP	57	57	9.2	30	51
TAP + cross-modal probe	57	66	10.1	31	54
DINOv3-B forward (reference)			*730*	*35 864*	

every support feature is kept at test time; in wall-clock the cache is faster (2.5 against 9.2 μs), being one large matrix product. The transported class names are computed once per task, not per query.

Memory and parameters. 51 MB holds the map, the OP matrix and the C transported names, against 2.9 MB for a probe and 50 MB for the 16-shot Tip-Adapter cache. The 11.8M trainable parameters are the map alone; nothing in the encoders is updated, and no encoder forward pass is repeated during adaptation. A narrower map lowers both memory and adaptation time at a small cost in accuracy (Table 17), and the linear residual map of Sec. 3.3 has about twenty times fewer parameters (0.6M).

C Algorithms

Algorithm 1 TAP adaptation

Require: support set $\Omega = (\{\boldsymbol{s}_i\}_{i=1}^N, \{\boldsymbol{l}_i\}_{i=1}^N)$ of $N = KC$ images, K per class
Require: frozen encoders $f_{\text{img}}, f_{\text{txt}}$
Require: residual map f_θ with a zero-initialized output layer
compute image and class-name embeddings $\{\boldsymbol{x}_i\}_{i=1}^N$ and $\{\boldsymbol{t}_i\}_{i=1}^N$ (Eq. 1)
average the support features of each class and normalize, giving the prototypes $\{\boldsymbol{x}_c\}_{c \in C}$
estimate the OP map $W^\star$ from the C prototype pairs by SVD (Eq. 2, Sec. 2.1)
align the class names to the visual space: $\boldsymbol{y}_i \leftarrow W^\star \boldsymbol{t}_i / \|W^\star \boldsymbol{t}_i\|$ for $i = 1, \ldots, N$ (Eq. 3)
compute the alignment cosine $\rho \leftarrow \frac{1}{C} \sum_{c \in C} \langle \boldsymbol{y}_c, \boldsymbol{x}_c \rangle$
for minibatches $(\boldsymbol{x}, \boldsymbol{y})$ of support pairs **do**
  compute the endpoint $\hat{\boldsymbol{x}}_1(\boldsymbol{x}) = (\boldsymbol{x} + f_\theta(\boldsymbol{x}))/\|\boldsymbol{x} + f_\theta(\boldsymbol{x})\|$
  update θ by cosine regression onto the aligned class name, $1 - \langle \hat{\boldsymbol{x}}_1(\boldsymbol{x}), \boldsymbol{y} \rangle$ (Eq. 4)
end for
set $\alpha \leftarrow \min(0.6, 1/K)$, $\tau \leftarrow \rho + (1 - \rho)\alpha$ (Eq. 8; no held-out data)

Algorithm 2 Few-shot inference: partial transport

Require: query s; class labels $\{l_c\}_{c\in C}$; encoders $f_{\text{img}}, f_{\text{txt}}$; residual map f_θ; map $W^\star$; class prototypes $\{\boldsymbol{x}_c\}_{c\in C}$; params τ, α;

1: **compute** embeddings $\boldsymbol{x}$ and $\{\boldsymbol{t}_c\}_{c\in C}$ (Eq. 1)
2: **align** the class names to the visual space: $\boldsymbol{y}_c \leftarrow W^\star \boldsymbol{t}_c / \|W^\star \boldsymbol{t}_c\|$ for every $c \in C$
3: **compute** the endpoint $\hat{\boldsymbol{x}}_1(\boldsymbol{x})$ of the learned map (Eq. 4)
4: **transport** the query a fraction τ along the geodesic toward it, $\boldsymbol{z}^{(\tau)} \leftarrow \gamma(\boldsymbol{x}, \hat{\boldsymbol{x}}_1(\boldsymbol{x}); \tau)$ (Eq. 5)
5: **transport** every class name the complementary fraction toward its class prototype, $\boldsymbol{z}_c^{(1-\tau)} \leftarrow \gamma(\boldsymbol{y}_c, \boldsymbol{x}_c; 1-\tau)$, in closed form (Eq. 5)
6: **compute** the transported score $\langle \boldsymbol{z}^{(\tau)}, \boldsymbol{z}_c^{(1-\tau)} \rangle$ and mix it with $\langle \boldsymbol{x}, \boldsymbol{y}_c \rangle$ by α, giving s_c (Eq. 7)
7: **predict** class $c = \arg\max_{c\in C} s_c$

D Ablations

Table 6: **Closed-form baselines and TAP by number of shots** (complements Table 1). Left: 11-dataset mean. Middle and right: ImageNet and its OOD average after K-shot adaptation on ImageNet. All rows use the same frozen DINOv3-B features, CLIP-B/16 text and support sets; 3 seeds. Closed-form rows use $\alpha = \min(0.6, 1/K)$ and the schedule τ; the identity-map row is TAP with the learned map removed (aligned names moved a fraction $1-\tau$ toward the class support means, query unchanged); the cache row is Tip-Adapter's key-value cache on DINOv3-B features with validated (α, β). The two TAP rows repeat Tables 1 and 2; the two combination rows average validated TAP's log posterior with the validated probe's and with the validated cross-modal probe's at equal weight. Which combination is better depends on the axis: on the eleven-dataset mean the cross-modal probe wins at 1, 2 and 8 shots and the plain probe at 4 and 16, by at most 0.11; on ImageNet the cross-modal probe wins at every shot count, and under shift at all but eight. Per-shift columns at sixteen shots: Table 7.

	11 datasets (mean)					ImageNet					ImageNet OOD average				
method	1	2	4	8	16	1	2	4	8	16	1	2	4	8	16
Cosine centroids	59.69	68.52	75.60	79.55	81.96	47.42	58.56	66.99	71.98	74.64	40.10	48.95	55.42	59.56	61.64
Linear probe (validated)	60.25	70.14	78.11	82.78	85.85	47.45	59.05	67.57	72.63	75.91	39.98	48.33	53.90	57.82	59.90
Centroids + Procrustes score	61.49	69.46	75.98	79.68	82.04	51.47	61.03	67.73	72.24	74.72	44.84	51.79	56.43	59.91	61.77
TAP, identity map	62.27	69.54	75.92	79.30	81.51	53.92	62.63	69.00	72.88	74.96	46.95	53.37	58.08	61.13	62.79
Tip-Adapter cache	62.42	69.86	75.75	79.47	82.17	53.67	62.11	67.79	71.70	74.13	46.33	52.79	57.43	60.39	61.92
Prior mixture	62.30	70.27	77.98	82.77	85.87	53.35	62.14	69.26	73.60	76.35	46.21	52.68	57.57	60.13	61.17
TAP (schedule)	63.98	72.21	79.04	83.00	85.96	55.90	64.37	70.28	74.17	76.96	48.48	54.19	58.10	60.90	62.44
TAP (validated)	64.08	72.32	79.22	83.32	86.18	56.04	64.27	70.49	74.28	77.05	47.93	54.03	58.74	61.18	62.63
TAP (validated) + probe	63.77	72.29	79.49	83.58	86.57	55.10	64.01	70.11	74.25	77.10	47.99	53.86	58.26	61.59	62.78
TAP (validated) + cross-modal probe	64.17	72.55	79.41	83.62	86.46	56.23	64.55	70.61	74.44	77.21	48.43	54.35	58.92	61.42	63.02

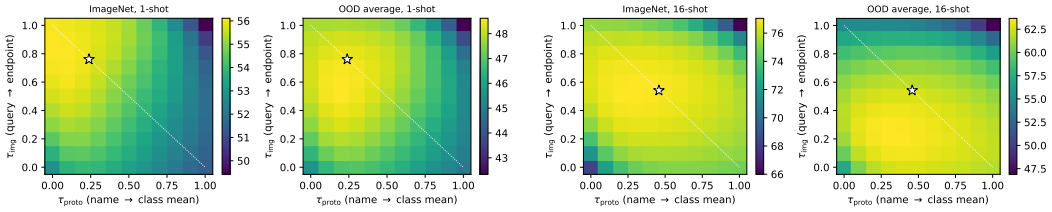


Figure 5: **Independent transport fractions (test-grid oracle, analysis only).** ImageNet (left of each pair) and OOD average (right) over an 11×11 grid of $(\tau_{\text{img}}, \tau_{\text{proto}})$ with the schedule's α, 3 seeds, for 1 shot (left pair) and 16 shots (right pair). The star is TAP's coupled schedule $(\tau, 1-\tau)$, the dotted line the coupling $\tau_{\text{proto}} = 1 - \tau_{\text{img}}$. Grid optima, selected on the test set and therefore not a method result: 1 shot, ImageNet 56.11 / 48.14 at $(0.8, 0.0)$ and OOD 48.62 at $(0.6, 0.2)$ (TAP: 55.90 / 48.48); 16 shots, ImageNet 77.04 / 61.91 at $(0.6, 0.3)$ and OOD 63.67 at $(0.2, 0.3)$ (TAP: 76.96 / 62.44). In-distribution accuracy is governed by τ_{img} (plateau at 0.4–0.7 for 16 shots, 0.6–1.0 for 1 shot); OOD accuracy falls once τ_{img} exceeds about 0.3 (16 shots) or 0.7 (1 shot); τ_{proto} has a broad optimum at 0.2–0.5 and hurts beyond it, more so at 1 shot where the class mean is a single image. One coupled scalar lands within 0.1–0.2 of the two-dimensional optimum in distribution.

D.1 Distribution-shift performance as a function of shot count

Table 8 tests how the text prior changes with label availability on ImageNet. The margin of TAP under the schedule over the validated probe falls from $+8.44$ in distribution and $+8.50$ under shift at one shot to $+1.05$ and $+2.54$ at sixteen: a larger fraction of the one-shot gain remains under shift. Which text-anchored row is best moves with K. TAP under the schedule holds the best OOD average

Table 7: **Class text against the strongest classifiers on the same frozen DINOv3-B features, per shift.** 16-shot ImageNet and the four shifts, evaluated without further adaptation; 3-seed means ± seed std. The labels column counts the labels each method spends per class: 20 marks the methods that select a hyper-parameter (probe C, GDA shrinkage, cache scalars, cross-modal weight, TAP's (α, τ) when validated) on the $\min(K, 4)$-image split, 16 the methods that select nothing. Text-anchored rows use CLIP-B/16 text except the last two. Eleven-dataset means and other shot counts: Tables 6 and 8.

method	labels	ImageNet	V2	Sk	A	R	OOD avg.
vision-only classifiers							
Cosine centroids	16	74.64 ± 0.22	65.35 ± 0.13	53.86 ± 0.12	51.71 ± 0.86	75.64 ± 0.42	61.64 ± 0.27
Linear probe	20	75.91 ± 0.28	66.65 ± 0.25	51.82 ± 0.19	48.78 ± 0.81	72.34 ± 0.60	59.90 ± 0.12
GDA (Wang et al., 2024)	20	74.91 ± 0.17	65.36 ± 0.31	50.74 ± 0.28	44.88 ± 1.36	66.35 ± 1.28	56.83 ± 0.62
class text, prior work							
Procrustes only	16	68.20 ± 0.04	60.32 ± 0.26	48.30 ± 0.35	49.29 ± 0.50	70.77 ± 0.27	57.17 ± 0.28
LFA (Ouali et al., 2023)	16	72.82 ± 0.04	63.78 ± 0.15	47.96 ± 0.46	46.26 ± 0.68	68.06 ± 0.20	56.51 ± 0.29
Tip-Adapter cache (Zhang et al., 2022)	20	74.13 ± 0.27	65.66 ± 0.43	53.80 ± 0.54	52.89 ± 1.26	75.31 ± 1.04	61.92 ± 0.78
Cross-modal probe (Lin et al., 2023)	16	74.69 ± 0.26	65.68 ± 0.17	54.11 ± 0.20	51.70 ± 0.81	75.68 ± 0.44	61.79 ± 0.30
Cross-modal probe (Lin et al., 2023)	20	76.39 ± 0.30	67.32 ± 0.28	53.58 ± 0.76	50.51 ± 0.65	74.63 ± 1.29	61.51 ± 0.68
class text, nothing learned (ours)							
Centroids + Procrustes score	16	74.72 ± 0.17	65.47 ± 0.20	53.96 ± 0.13	51.85 ± 0.89	75.77 ± 0.44	61.77 ± 0.26
TAP, identity map	16	74.96 ± 0.11	66.23 ± 0.22	54.88 ± 0.26	53.23 ± 0.70	76.80 ± 0.37	62.79 ± 0.34
Prior mixture	20	76.35 ± 0.26	67.26 ± 0.20	53.25 ± 0.25	50.21 ± 0.51	73.97 ± 0.80	61.17 ± 0.23
class text, learned map (ours)							
TAP, schedule	16	76.96 ± 0.23	67.89 ± 0.04	54.52 ± 0.14	52.76 ± 0.55	74.59 ± 0.24	62.44 ± 0.14
TAP, validated	20	77.05 ± 0.17	68.06 ± 0.13	54.86 ± 0.29	52.87 ± 0.53	74.71 ± 0.46	62.63 ± 0.26
TAP (val.) + probe (val.)	20	77.10 ± 0.16	68.16 ± 0.06	54.71 ± 0.52	53.10 ± 1.04	75.16 ± 0.76	62.78 ± 0.55
TAP (val.) + cross-modal probe (val.)	20	77.21 ± 0.19	68.30 ± 0.06	55.02 ± 0.28	53.31 ± 0.69	75.45 ± 0.40	63.02 ± 0.30
TAP, full transport (τ=1, α=0)	16	74.97 ± 0.06	65.24 ± 0.21	48.57 ± 0.30	42.52 ± 0.53	59.71 ± 0.36	54.01 ± 0.12
other text encoders, TAP under the schedule (ours)							
TAP, SigLIP2-B/16 text	16	76.94 ± 0.15	68.10 ± 0.20	54.92 ± 0.05	53.70 ± 0.55	76.11 ± 0.30	63.21 ± 0.17
TAP, mini-GTE text	16	76.47 ± 0.20	67.47 ± 0.33	53.50 ± 0.10	51.81 ± 0.72	73.51 ± 0.07	61.57 ± 0.17

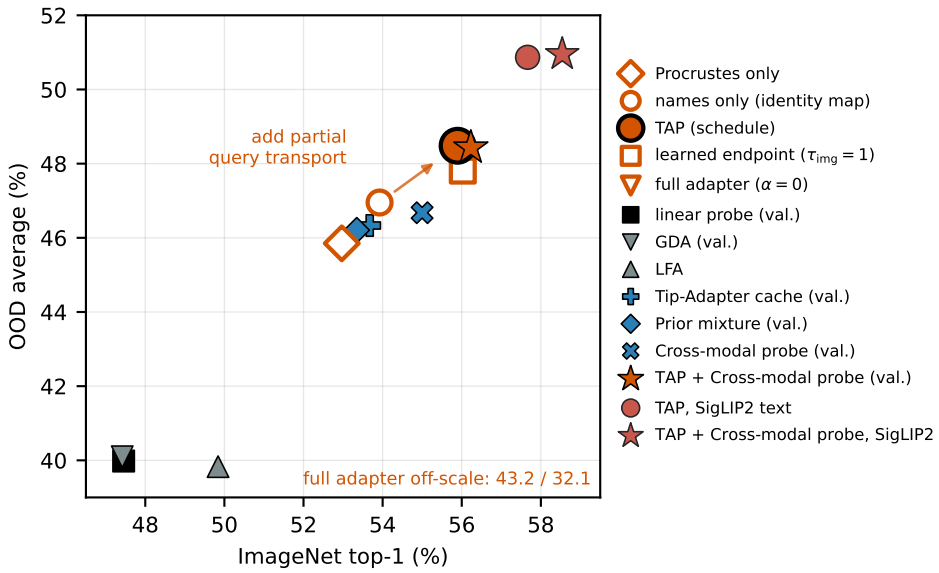


Figure 6: **ImageNet against its shifts at one shot**, the counterpart of Fig. 3. Same axes and markers, with the Procrustes-aligned names alone and the two SigLIP2-text rows added. Nothing trades off here: the text-using methods improve both axes, LFA and the fully used map excepted, and TAP is at the corner. The learned map used fully is off the scale at 43.2 / 32.1. The probe and GDA coincide at 47.4 / 40.0.

at one shot, the only column where no combination row is ahead; the cross-modal combination is the best row in distribution at every shot count and under shift at every shot count but one and eight, where the schedule and the plain-probe combination lead. At sixteen shots the identity map is 0.95 *below* the probe in distribution while reaching 62.79 under shift, behind only the cross-modal combination's 63.02 and ahead of TAP itself: the split described in Sec. 3.2, where the learned map buys in-distribution accuracy and the transported names buy accuracy under shift. The prior mixture, which blends the probe's score with the aligned-text score at a single global weight instead of transporting, stays 2.55 and 2.27 points behind TAP at one shot and 0.61 and 1.27 at sixteen. The

two vision-only rows separate for the same reason in reverse: GDA's within-class covariance is worth $+0.27$ over cosine centroids in distribution at sixteen shots and costs 4.81 under shift, and at one shot, with no covariance to estimate, the two classifiers coincide. Whether the names have to be the correct ones, rather than any well-conditioned set of class-specific targets, is tested separately in Appendix D.2.

Table 8: **ImageNet and its shifts by shot count.** Source ImageNet top-1 and OOD avg., the mean over ImageNet-V2, -Sketch, -A and -R; frozen DINOv3-B features, CLIP-B/16 text, 3-seed means, bold best per column. The per-shift columns at sixteen shots are Table 7. GDA and cosine centroids coincide at one shot, where a single image per class leaves no within-class covariance to estimate.

	1-shot		2-shot		4-shot		8-shot		16-shot	
method	source	OOD	source	OOD	source	OOD	source	OOD	source	OOD
Cosine centroids	47.42	40.10	58.56	48.95	66.99	55.42	71.98	59.56	74.64	61.64
Linear probe (validated)	47.45	39.98	59.05	48.33	67.57	53.90	72.63	57.82	75.91	59.90
GDA	47.42	40.10	58.97	46.52	67.63	51.66	72.43	55.04	74.91	56.83
Procrustes only	52.97	45.85	59.96	50.90	64.43	54.38	66.93	56.30	68.20	57.17
Centroids + Procrustes score	51.47	44.84	61.03	51.79	67.73	56.43	72.24	59.91	74.72	61.77
Prior mixture (ours)	53.35	46.21	62.14	52.68	69.26	57.57	73.60	60.13	76.35	61.17
TAP, identity map (ours)	53.92	46.95	62.63	53.37	69.00	58.08	72.88	61.13	74.96	62.79
TAP, schedule (ours)	55.90	**48.48**	64.37	54.19	70.28	58.11	74.17	60.90	76.96	62.44
TAP, validated (ours)	56.04	47.93	64.27	54.03	70.49	58.74	74.28	61.18	77.05	62.63
TAP (val.) + probe (val.) (ours)	55.10	47.99	64.01	53.86	70.11	58.26	74.25	**61.59**	77.10	62.78
TAP (val.) + cross-modal probe (val.) (ours)	**56.23**	48.43	**64.55**	**54.35**	**70.61**	**58.92**	**74.43**	61.42	**77.21**	**63.02**
Δ TAP (schedule) $-$ probe	$+8.44$	$+8.50$	$+5.32$	$+5.86$	$+2.71$	$+4.20$	$+1.54$	$+3.08$	$+1.05$	$+2.54$

D.2 Control: class names assigned to the wrong classes

Section 3.4 shows that the gain is not tied to a particular text embedding space. It does not separate names that describe their class from any well-conditioned set of class-specific targets. This control keeps everything about the text prototypes identical (encoder, prompts, norms, pairwise distances) and breaks only the correspondence between classes and names. We draw one fixed derangement π of the 1000 ImageNet classes, so that no class keeps its own name, and train TAP with the embedding of class $\pi(c)$ as the target of class c; the Procrustes step, the learned map, the training recipe and both selections of (α, τ) are unchanged. The experiment is run at one shot, where the prior contributes most (Table 8), with three training seeds and the same permutation, under the recipe of Sec. 3, so the correct-name rows are the paper's own (Table 2).

Table 9: **TAP with the class names assigned to the wrong classes.** ImageNet 1-shot, DINOv3-B image and CLIP-B/16 text, 3 seeds; standard deviations for ImageNet and the OOD average. "Shuffled" rows use the same text prototypes assigned to the wrong classes by one fixed derangement; everything else, including the recipe and both selections of (α, τ), is the paper's. The val. column and the vision-only references are as in Table 8.

method	val.	ImageNet	-V2	-Sketch	-A	-R	OOD avg
Cosine centroids (vision-only)	–	47.42 ± 0.69	41.64	32.05	33.65	53.05	40.10 ± 0.95
Linear probe (vision-only)	✓	47.45 ± 0.67	41.65	32.02	33.33	52.93	39.98 ± 1.05
Procrustes only, correct names	–	52.97	47.09	35.71	40.34	60.27	45.85
TAP, shuffled names	–	41.64 ± 0.68	35.85	26.00	23.20	39.08	31.03 ± 0.44
TAP, shuffled names	✓	47.01 ± 0.73	41.18	31.66	33.27	52.49	39.65 ± 1.05
TAP, correct names	–	55.90 ± 0.12	49.56	**39.20**	**41.72**	**63.44**	**48.48** ± 0.30
TAP, correct names	✓	**56.04** ± 0.11	**49.56**	39.01	40.62	62.53	47.93 ± 0.33

A wrong prior is worse than no prior. Under the schedule the shuffled model reaches 41.64 in distribution and 31.03 under shift, 5.8 and 9.1 points below cosine centroids on the same support and 14.3 and 17.5 below TAP with the correct names; used at full strength it falls to 22.93 / 12.21, against 43.23 / 32.15 for the correct-name endpoint. The text prototypes therefore help through their correspondence with the classes, not by supplying a well-separated set of class-specific targets, and the eight points TAP gains at one shot are not an artifact of regressing onto separated unit vectors.

The two selections of (α, τ) behave differently here. The schedule (Eq. 8) reads the support only, and the permutation is nearly invisible there: ρ falls from 0.40 to 0.35, so the schedule still gives the aligned score weight α=0.6 and still transports the query τ=0.74 of the way toward a wrong name, which accounts for the 14.3-point loss. Validated selection detects the permutation: it selects τ=0

in two of three seeds, which leaves each name on its class support mean, and recovers the centroid classifier to within 0.41 in distribution and 0.45 under shift (47.01 / 39.65). With an uninformative prior the best operating point of Eq. 6 is the vision-only classifier, and one validation image per class, the budget at one shot, suffices to find it. This is a second reason for reporting every TAP number both under the schedule and validated.

D.3 Control: a random map in place of the alignment

This control asks whether the closed-form alignment is needed, since the learned map could in principle discover the correspondence between images and names by itself. On the default pair the text space (512-d) is smaller than the image space (768-d), so the control replaces $\mathbf{W}^\star$ by a fixed random semi-orthogonal 768×512 map, the same for every seed, and changes nothing else: same recipe (Sec. 3), same closed-form schedule and same validated selection. The two arms of Table 10 are therefore matched, and the Procrustes arm is the paper's own result.

Table 10: **TAP with the Procrustes alignment replaced by a fixed random semi-orthogonal map.** DINOv3-B image and CLIP-B/16 text on ImageNet, 3 seeds, ImageNet top-1 (left) and OOD average (right). The val. column is as in Table 8; Δ is the paired difference between the two validated arms. With the random map the alignment cosine is ρ=0.001 at every shot count, against 0.40 to 0.51 with Procrustes.

		ImageNet					OOD average				
method	val.	1	2	4	8	16	1	2	4	8	16
Linear probe (vision-only)	✓	47.45	59.05	67.57	72.63	75.91	39.98	48.33	53.90	57.82	59.90
TAP, fixed random map	–	43.28	58.18	67.19	72.05	74.67	32.99	46.54	55.53	59.58	61.69
TAP, fixed random map	✓	49.17	59.61	68.18	72.84	76.32	40.79	45.75	55.48	57.50	60.85
TAP, Procrustes	–	55.90	64.37	70.28	74.17	76.96	48.48	54.19	58.11	60.90	62.44
TAP, Procrustes	✓	**56.04**	**64.27**	**70.49**	**74.28**	**77.05**	**47.93**	**54.03**	**58.74**	**61.18**	**62.63**
Δ *(validated)*		+6.87	+4.66	+2.31	+1.44	+0.73	+7.14	+8.28	+3.26	+3.68	+1.78

Without the alignment there is no prior, only a classifier. Under the random map the aligned similarity $\langle \boldsymbol{x}, \boldsymbol{y}_c \rangle$ is noise (ρ=0.001), and validated selection responds by setting α to 0 in eight of the fifteen splits and to 0.1 in four more, whereas with Procrustes α never falls below 0.2. What remains is a residual map that regresses images onto a thousand fixed arbitrary unit targets, that is, a classifier with a random code per class. At sixteen shots this classifier is adequate, 76.32 against the validated probe's 75.91, but it has no prior, so its gap to TAP follows the label budget as the prior does throughout the paper: 6.87 in distribution and 7.14 under shift at one shot, 0.73 and 1.78 at sixteen.

Under the schedule the cost is larger at every shot count, 2.12 to 12.62 points in distribution, for a mechanical reason: $\tau = \rho + (1-\rho)\alpha$ reduces to α when ρ vanishes, so the query barely moves while each name travels $1-\alpha$ of the way to its class mean, and Eq. 6 reduces to the centroid classifier plus a share α of noise. Unlike the wrong names of Appendix D.2, a missing alignment is visible on the support, and both selections respond to it. The alignment is therefore not a shortcut that spares the map some work. It supplies the correspondence that makes the text score a prior rather than an arbitrary label code, and the learned map refines that correspondence instead of discovering it.

D.4 Encoder scaling

Table 14 is the starting point that the rest of the paper builds on: at comparable vision-backbone compute, linear probes on DINOv3 features are above linear probes on CLIP features at every evaluated scale, and at ViT-B the DINOv3-B probe is 7.8 points above the CLIP-B/16 probe on ImageNet and 11.3 points above on the OOD average. This is the motivation of Sec. 1: the stronger few-shot representation in this comparison is not the CLIP image tower. What TAP adds on top of these features, per image encoder and shot count, is Table 12; the choice of text encoder with the image encoder fixed is Table 11 in the main text. All rows follow the protocol of Table 2: adaptation on 16-shot ImageNet, the probe's C selected on the validation split, evaluation without further adaptation on the four shifts, 3-seed means.

TAP on DINOv3-L, DINOv2-L and CAPI-L at every shot count. Table 12 extends Table 11 to $K \in \{1, 2, 4, 8, 16\}$ and adds the Procrustes-only classifier (aligned names, no learned map). The schedule's (α, τ) follow Eq. 8 from each encoder's own alignment cosine. The validation split

Table 11: **The prior is a property of the pair of encoders.** ImageNet / OOD average, 3 seeds; validated probe against TAP under the closed-form schedule, Δ paired over seeds. Top: one frozen image encoder per row with CLIP-B/16 text, self-supervised distillation (Siméoni et al., 2025; Oquab et al., 2024) or masked image modelling alone (Darcet et al., 2025). Bottom: one frozen text encoder per row on DINOv3-B features, two trained against images and two on text alone. All shot counts and validated (α, τ) for the image encoders: Appendix Table 12; validated results for the text encoders: Appendix Table 13.

	1 shot			16 shots		
encoder	probe	TAP	Δ	probe	TAP	Δ
image encoder, CLIP-B/16 text						
DINOv3-B/16 (default)	47.45 / 39.98	55.90 / 48.48	$+8.4$ / $+8.5$	75.91 / 59.90	76.96 / 62.44	$+1.0$ / $+2.5$
DINOv3-L/16	59.99 / 59.28	66.43 / 65.63	$+6.4$ / $+6.4$	81.41 / 74.68	82.12 / 76.23	$+0.7$ / $+1.5$
DINOv2-L/14	60.13 / 57.36	63.61 / 60.38	$+3.5$ / $+3.0$	78.87 / 68.50	79.10 / 69.56	$+0.2$ / $+1.1$
CAPI-L/14 (masked im.)	33.96 / 20.96	39.09 / 26.41	$+5.1$ / $+5.4$	67.40 / 41.83	71.03 / 44.64	$+3.6$ / $+2.8$
text encoder, DINOv3-B/16 features (the probe uses no text)						
SigLIP2-B/16 (im.–text)	47.45 / 39.98	57.66 / 50.87	$+10.2$ / $+10.9$	75.91 / 59.90	76.94 / 63.21	$+1.0$ / $+3.3$
CLIP-B/16 (im.–text)	47.45 / 39.98	55.90 / 48.48	$+8.4$ / $+8.5$	75.91 / 59.90	76.96 / 62.44	$+1.0$ / $+2.5$
Qwen3-Embedding-0.6B	47.45 / 39.98	48.44 / 40.59	$+1.0$ / $+0.6$	75.91 / 59.90	76.51 / 61.56	$+0.6$ / $+1.7$
mini-GTE	47.45 / 39.98	46.31 / 38.76	-1.1 / -1.2	75.91 / 59.90	76.47 / 61.57	$+0.6$ / $+1.7$

for these encoders is built in the encoder's own feature space with the same image selection as for DINOv3-B ($\min(K, 4)$ images per class, disjoint from the support).

Table 12: **TAP on DINOv3-L/16, DINOv2-L/14 and CAPI-L/14 at every shot count**, ImageNet / OOD average, CLIP-B/16 text, 3 seeds. Δ is the paired gain over the linear probe on the same features; the probe selects its C on the validation split (val. ✓), the Procrustes-only classifier and TAP under the schedule use none (val. –). Recipe of Table 2.

encoder	K	probe (val. ✓)	Procrustes only (val. –)	TAP (val. –)	Δ	TAP (val. ✓)	Δ
DINOv3-L	1	59.99 / 59.28	64.11 / 62.50	66.43 / 65.63	$+6.4$ / $+6.4$	66.11 / 65.20	$+6.1$ / $+5.9$
	2	69.51 / 66.80	70.11 / 66.99	73.24 / 70.48	$+3.7$ / $+3.7$	73.18 / 70.47	$+3.7$ / $+3.7$
	4	76.16 / 71.28	73.76 / 69.95	77.80 / 73.72	$+1.6$ / $+2.4$	77.77 / 73.51	$+1.6$ / $+2.2$
	8	79.62 / 73.88	75.56 / 71.12	80.34 / 75.34	$+0.7$ / $+1.5$	80.33 / 75.22	$+0.7$ / $+1.3$
	16	81.41 / 74.68	76.53 / 71.92	82.12 / 76.23	$+0.7$ / $+1.5$	82.05 / 75.63	$+0.6$ / $+1.0$
CAPI-L	1	33.96 / 20.96	35.83 / 24.47	39.09 / 26.41	$+5.1$ / $+5.4$	40.55 / 27.08	$+6.6$ / $+6.1$
	2	44.35 / 27.63	41.92 / 27.39	49.90 / 32.11	$+5.5$ / $+4.5$	52.32 / 34.13	$+8.0$ / $+6.5$
	4	54.30 / 33.62	46.27 / 30.18	60.83 / 38.74	$+6.5$ / $+5.1$	61.93 / 39.70	$+7.6$ / $+6.1$
	8	61.93 / 38.54	49.05 / 31.90	65.54 / 41.20	$+3.6$ / $+2.7$	66.55 / 42.44	$+4.6$ / $+3.9$
	16	67.40 / 41.83	50.76 / 32.73	71.03 / 44.64	$+3.6$ / $+2.8$	71.35 / 45.32	$+4.0$ / $+3.5$
DINOv2-L	1	60.13 / 57.36	62.43 / 59.58	63.61 / 60.38	$+3.5$ / $+3.0$	63.46 / 60.06	$+3.3$ / $+2.7$
	2	68.16 / 62.72	68.27 / 63.89	70.12 / 64.66	$+2.0$ / $+1.9$	69.95 / 63.86	$+1.8$ / $+1.1$
	4	73.74 / 65.70	72.01 / 66.52	74.80 / 67.35	$+1.1$ / $+1.6$	74.78 / 67.28	$+1.0$ / $+1.6$
	8	76.76 / 67.53	73.67 / 67.78	77.27 / 68.73	$+0.5$ / $+1.2$	77.34 / 68.46	$+0.6$ / $+0.9$
	16	78.87 / 68.50	74.41 / 68.27	79.10 / 69.56	$+0.2$ / $+1.1$	79.28 / 69.26	$+0.4$ / $+0.8$
DINOv3-B	1	47.45 / 39.98	52.97 / 45.85	55.90 / 48.48	$+8.4$ / $+8.5$	56.04 / 47.93	$+8.6$ / $+7.9$
	2	59.06 / 48.33	59.96 / 50.90	64.37 / 54.19	$+5.3$ / $+5.9$	64.27 / 54.04	$+5.2$ / $+5.7$
	4	67.57 / 53.90	64.43 / 54.38	70.28 / 58.11	$+2.7$ / $+4.2$	70.49 / 58.74	$+2.9$ / $+4.8$
	8	72.63 / 57.82	66.93 / 56.30	74.17 / 60.90	$+1.5$ / $+3.1$	74.28 / 61.18	$+1.6$ / $+3.4$
	16	75.91 / 59.90	68.20 / 57.17	76.96 / 62.44	$+1.0$ / $+2.5$	77.05 / 62.63	$+1.1$ / $+2.7$

Table 13: **Text encoders: TAP under the closed-form schedule and with validated (α, τ).** DINOv3-B features, ImageNet / OOD average, 3 seeds; the validated probe is 47.45 / 39.98 at one shot and 75.91 / 59.90 at sixteen. Validation changes the sixteen-shot result by under 0.3 in distribution for every tower.

	1 shot		16 shots	
text encoder	schedule	validated	schedule	validated
SigLIP2-B/16	57.66 / 50.87	58.48 / 50.72	76.94 / 63.21	77.14 / 63.06
CLIP-B/16 (default)	55.90 / 48.48	56.04 / 47.93	76.96 / 62.44	77.05 / 62.63
Qwen3-Embedding-0.6B	48.44 / 40.59	49.12 / 42.42	76.51 / 61.56	76.31 / 61.55
mini-GTE	46.31 / 38.76	48.54 / 42.09	76.47 / 61.57	76.48 / 61.62

Table 14: **Linear probes on CLIP and on DINOv3 features, 16-shot ImageNet.** Top-1 accuracy (%) on ImageNet and the OOD average over ImageNet-V2, -Sketch, -A and -R, the probe's C selected on the validation split, 3-seed means. GFLOPs are the image encoder's forward pass. The self-supervised features are the stronger starting point at every scale; what TAP adds on top of them is Table 12.

probe features	image encoder	GFLOPs	ImageNet	OOD avg.
CLIP	CLIP-B/32	4.5	61.90	37.28
	CLIP-B/16	17.6	68.15	48.63
	CLIP-L/14	80.0	75.16	63.15
DINOv3	DINOv3-S/16	4.5	66.39	41.49
	DINOv3-B/16	17.6	75.91	59.90
	DINOv3-L/16	70.0	81.41	74.68

D.5 Transport curves at one shot

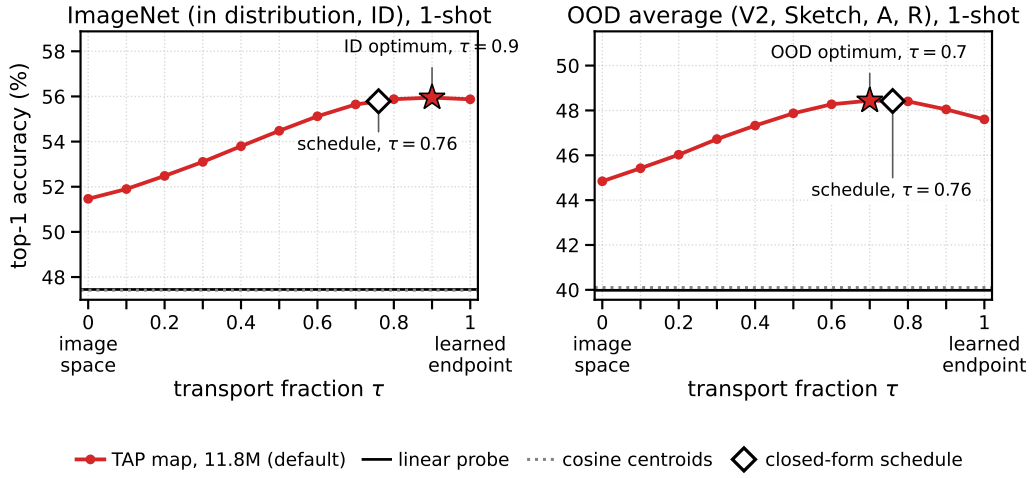


Figure 7: **Accuracy along the transport fraction τ, ImageNet 1-shot** (TAP's map, 11.8M; α=0.6 held fixed; 3-seed means, seed std at most 0.6). Same construction as Fig. 4. With one image per class the closed-form prior carries most of the weight, and the curve is above the probe and the centroids at every τ; the optima move to the right (τ=0.9 in distribution, 0.7 under shift) and the collapse at the learned endpoint is mild (0.8 OOD points below the optimum). The schedule's τ=0.76 sits between the two optima, within 0.2 of both.

D.6 Interpolation geometry: does the path matter?

TAP moves the query and the names along great circles. Table 15 keeps the checkpoints and the schedule's (α, τ) of Table 2 and changes only how the two endpoints are combined: normalized linear interpolation of the vectors instead of slerp; interpolation of the two endpoint scores, $(1-\tau)\langle \boldsymbol{x}, \boldsymbol{x}_c\rangle + \tau\langle \hat{\boldsymbol{x}}_1, \boldsymbol{y}_c\rangle$; and a free mixture of the four similarities between $\{\boldsymbol{x}, \hat{\boldsymbol{x}}_1\}$ and $\{\boldsymbol{y}_c, \boldsymbol{x}_c\}$ with weights on the simplex (step 0.1) selected on the validation split or, as an oracle, on the test set. Every row is mixed with the aligned-name score by the same α. Slerp and linear interpolation agree within 0.1, so the shape of the path is immaterial; what matters is where transport stops. Interpolating scores rather than vectors is 0.4–0.7 lower in distribution, and no reweighting of the four similarities, even with test-selected weights, improves on TAP on both axes at once (the oracle gains at most 0.2 in distribution and gives back 0.1 to 0.4 under shift): the gain is a property of the stopping point, not of a weighting the schedule failed to find. The selected weights (about 0.2 centroids, 0.5–0.6 transported endpoint, 0.2 aligned names at sixteen shots) are close to what slerp at τ=0.54 implies.

D.7 Training length: full and partial transport at every stopping point

Table 16 trains TAP's map for 5, 10, 25 and 50 epochs on 16-shot ImageNet and scores each checkpoint both ways. Used fully, the map is below the probe on both axes at every training length, furthest at 5 epochs, and it closes the gap slowly; stopping training early does not rescue it. Used

Table 15: **How the endpoints are combined.** ImageNet / OOD average, DINOv3-B, CLIP-B/16 text, 3 seeds; same checkpoints and (α, τ) as Table 2. The last row selects the mixture weights on the test set (analysis only).

combination of the endpoints	1 shot (τ=0.76)	16 shots (τ=0.54)
slerp (TAP)	55.90 / 48.48	76.96 / 62.44
normalized linear interpolation	55.96 / 48.44	76.97 / 62.38
score interpolation	55.21 / 48.27	76.53 / 62.52
four-similarity mixture, weights validated	55.53 / 48.19	76.95 / 62.49
four-similarity mixture, weights test-selected	56.11 / 48.06	77.00 / 62.32
full transport	43.23 / 32.15	74.97 / 54.01

partially under the schedule, the same checkpoints are above the probe under shift from 5 epochs on and on both axes from 10 epochs on, and gain 1.3 points in distribution from 5 to 50 epochs.

Table 16: **Full and partial transport as a function of training length.** ImageNet 16-shot, DINOv3-B, CLIP-B/16 text, 3 seeds; ImageNet / OOD average. Partial: the schedule's (α, τ); full: τ=1, α=0. Validated linear probe: 75.91 / 59.90.

epochs	partial transport	full transport
5	75.69 / 62.21	66.06 / 48.02
10	76.25 / 62.48	72.38 / 52.61
25	76.82 / 62.52	74.54 / 53.74
50	76.96 / 62.44	74.97 / 54.01

D.8 Residual-map architecture

The residual stack of TAP's learned map (Sec. 2.2) is two blocks of width 1536. Table 17 varies the width on 16-shot ImageNet at 3 seeds and replaces the stack by a single residual linear layer. Capacity buys little: from width 768 to 2560, an eightfold increase in parameters, ImageNet moves 0.55 and the OOD average 0.30, and the linear map at 0.6M is 1.7 below the default in distribution while holding its OOD average. Two blocks of width 1536 is the default because it is where the in-distribution curve flattens, not because wider hurts.

Table 17: **Capacity of TAP's map**, ImageNet 16-shot, ImageNet / OOD average, 3 seeds; the first row replaces the MLP by a single residual linear layer trained with the same objective; TAP under the closed-form schedule (val. –) and with validated (α, τ) (val. ✓), the marker of Tables 1–2. The probe selects C on the validation split.

width	params (M)	TAP (val. –)	TAP (val. ✓)
linear residual map	0.6	75.23 / 62.35	75.43 / 63.03
768	3.6	76.55 / 62.24	76.54 / 62.77
1536 (default)	11.8	76.96 / 62.44	77.05 / 62.63
2128	21.4	77.01 / 62.55	76.98 / 63.13
2560	30.2	77.10 / 62.54	77.17 / 62.49
linear probe	0.8	75.91 / 59.90	

E Detailed results

Table 18 gives the eleven-dataset mean of every row at every shot count; Tables 19 and 20 give the per-dataset results of TAP and of its combination with the probe, of the matched probe, and of the paired difference between TAP and the probe, with seed standard deviations.

F Orthogonal Procrustes: degenerate cases

Rank deficiency ($C < D_{\text{txt}}$). On DTD, Pets and EuroSAT the number of classes is below the text dimension, so $\mathbf{M} = \mathbf{X}^\top \mathbf{T}$ has rank at most C. We do not truncate the SVD: the singular vectors of the zero singular values form an arbitrary orthonormal completion, every such completion minimizes Eq. 2, and $\mathbf{W}^\star$ remains a full isometry. The support determines how the $\operatorname{rank}(\mathbf{M})$ text directions it spans are placed in the visual space; the remaining directions are embedded orthonormally in

Table 18: **Few-shot classification on 11 datasets, mean accuracy**. Vision-only rows and TAP use the same frozen DINOv3-B features and support; the probe's C and GDA's shrinkage are selected on the validation split, TAP and the prior mixture use the closed-form schedule of Eq. 8. CLIP rows are published numbers. Bottom: paired gain of TAP (CLIP-B/16 text) over the probe, computed over 11 datasets $\times$ 3 seeds and reported as mean $\pm$ std over seeds of the 11-dataset mean; per-dataset values in Table 20.

method	text encoder	params (M)	1-shot	2-shot	4-shot	8-shot	16-shot
vision only, DINOv3-B							
Cosine centroids	–	86	59.69	68.52	75.60	79.55	81.96
Linear probe (fixed C=1)	–	86	59.99	67.50	74.17	79.40	83.39
Linear probe (fixed C=10)	–	86	60.11	69.91	77.62	82.41	85.63
Linear probe	–	86	60.25	70.14	78.11	82.78	85.85
GDA	–	86	59.69	70.20	78.22	82.62	85.25
jointly trained pair, CLIP-B/16 image tower (context; published numbers)							
CoOp	CLIP-B/16	86+63	67.2	70.5	73.0	76.4	79.0
TIP-Adapter-F	CLIP-B/16	86+63	**70.9**	**73.7**	75.6	78.3	80.7
TIMO	CLIP-B/16	86+63	66.78	68.69	71.63	74.20	76.86
ProGrad	CLIP-B/16	86+63	69.5	72.4	74.7	77.4	79.9
text-anchored DINOv3-B							
Centroids + Procrustes score	CLIP-B/16	86+63	61.49	69.46	75.98	79.68	82.04
Prior mixture (ours)	CLIP-B/16	86+63	62.30	70.27	77.98	82.77	85.87
TAP (ours)	CLIP-B/16	86+63	63.98	72.21	79.04	83.00	85.96
TAP + cross-modal probe (ours)	CLIP-B/16	86+63	**64.17**	**72.55**	**79.41**	**83.62**	**86.46**
Δ TAP $-$ probe			+3.73	+2.07	+0.93	+0.22	+0.11
Δ (TAP + cross-modal probe) $-$ probe			+3.92	+2.41	+1.30	+0.84	+0.61

Table 19: **Per-dataset few-shot results on 11 datasets for TAP and for TAP + probe.** Mean $\pm$ std over 3 seeds. The first block is the per-dataset detail of the TAP row of Table 18, under the closed-form schedule; the second averages TAP's log posterior with the validated linear probe's at equal weight, with (α, τ) and the probe's C selected on the $\min(K, 4)$ split, and has its eleven-dataset mean in Table 6. The matched probe and the paired gains are in Table 20.

	dataset	1-shot	2-shot	4-shot	8-shot	16-shot	mean
TAP, CLIP-B/16	OxfordPets	78.22 ± 3.56	83.97 ± 1.63	90.52 ± 1.18	93.30 ± 0.11	94.24 ± 0.41	88.05
	DescribableTextures	47.20 ± 3.37	57.60 ± 1.55	65.76 ± 1.39	71.39 ± 0.31	76.16 ± 1.15	63.62
	EuroSAT	65.81 ± 4.30	73.25 ± 5.06	83.54 ± 1.80	88.27 ± 1.94	91.13 ± 0.09	80.40
	Caltech101	87.64 ± 1.48	92.51 ± 2.30	96.15 ± 0.67	97.04 ± 0.59	97.72 ± 0.25	94.21
	FGVCAircraft	35.35 ± 0.37	44.77 ± 1.76	56.13 ± 0.11	63.61 ± 0.81	71.49 ± 1.34	54.27
	UCF101	61.40 ± 2.24	70.31 ± 0.65	77.39 ± 1.19	82.25 ± 0.82	84.93 ± 0.74	75.25
	OxfordFlowers	97.54 ± 0.81	99.31 ± 0.25	99.78 ± 0.10	99.80 ± 0.04	99.88 ± 0.00	99.26
	StanfordCars	60.98 ± 1.87	75.48 ± 0.92	83.99 ± 0.83	88.86 ± 0.12	92.02 ± 0.23	80.27
	Food101	62.97 ± 1.31	72.56 ± 0.38	79.29 ± 0.63	83.00 ± 0.55	86.22 ± 0.34	76.81
	SUN397	50.80 ± 1.59	60.22 ± 0.55	66.59 ± 0.28	71.37 ± 0.28	74.81 ± 0.23	64.76
	ImageNet	55.90 ± 0.12	64.37 ± 0.46	70.28 ± 0.17	74.17 ± 0.22	76.96 ± 0.23	68.34
	mean	**63.98**	**72.21**	**79.04**	**83.00**	**85.96**	**76.84**
TAP (val.) + probe (val.)	OxfordPets	77.61 ± 3.96	83.01 ± 2.07	90.05 ± 1.18	93.12 ± 0.13	94.28 ± 0.39	87.62
	DescribableTextures	47.14 ± 3.43	58.77 ± 1.66	66.41 ± 1.36	72.62 ± 0.79	77.09 ± 1.12	64.41
	EuroSAT	67.09 ± 6.06	73.07 ± 5.03	84.76 ± 1.59	88.65 ± 1.25	91.43 ± 0.43	81.00
	Caltech101	87.83 ± 1.58	92.43 ± 2.26	96.35 ± 0.60	97.05 ± 0.63	97.92 ± 0.08	94.32
	FGVCAircraft	35.72 ± 0.48	46.67 ± 0.44	58.88 ± 0.60	67.79 ± 0.04	76.74 ± 0.58	57.16
	UCF101	60.62 ± 1.26	70.54 ± 0.53	77.21 ± 1.19	82.41 ± 1.31	84.65 ± 0.41	75.09
	OxfordFlowers	97.70 ± 0.66	99.11 ± 0.36	99.58 ± 0.24	99.62 ± 0.23	99.19 ± 0.23	99.04
	StanfordCars	60.40 ± 1.42	75.25 ± 1.43	84.32 ± 0.72	89.28 ± 0.30	92.44 ± 0.12	80.34
	Food101	62.11 ± 0.84	72.28 ± 0.42	79.62 ± 0.65	82.91 ± 0.28	86.25 ± 0.42	76.64
	SUN397	50.18 ± 2.20	60.04 ± 0.60	67.12 ± 0.39	71.62 ± 0.37	75.23 ± 0.25	64.84
	ImageNet	55.10 ± 0.92	64.01 ± 0.51	70.11 ± 0.19	74.25 ± 0.28	77.10 ± 0.16	68.12
	mean	**63.77**	**72.29**	**79.49**	**83.58**	**86.57**	**77.14**

a data-independent orientation and do not affect the class similarities, which involve only the C prototypes.

Table 20: **Per-dataset results for the matched linear probe on DINOv3-B, and the paired gain of TAP (CLIP-B/16 text) over it.** Mean ± std over 3 seeds; the gain is computed seed by seed on the same support set.

	dataset	1-shot	2-shot	4-shot	8-shot	16-shot	mean
Linear probe (validated)	OxfordPets	74.05 ± 3.13	81.99 ± 1.27	89.77 ± 1.48	93.32 ± 0.35	94.19 ± 0.52	86.67
	DescribableTextures	46.71 ± 3.77	57.43 ± 1.64	66.06 ± 0.59	72.34 ± 0.54	77.23 ± 1.41	63.95
	EuroSAT	61.43 ± 3.41	68.59 ± 6.62	82.33 ± 2.74	87.11 ± 1.26	90.01 ± 1.52	77.89
	Caltech101	86.73 ± 1.28	92.35 ± 2.72	96.20 ± 0.39	97.01 ± 0.72	97.67 ± 0.12	93.99
	FGVCAircraft	34.84 ± 0.25	46.60 ± 0.44	57.48 ± 0.25	66.47 ± 0.46	74.84 ± 0.94	56.05
	UCF101	58.59 ± 0.88	69.29 ± 0.61	76.91 ± 0.63	82.29 ± 1.23	84.53 ± 0.58	74.32
	OxfordFlowers	97.24 ± 0.71	99.31 ± 0.32	99.78 ± 0.02	99.76 ± 0.07	99.84 ± 0.08	99.19
	StanfordCars	53.84 ± 0.82	70.66 ± 0.74	80.06 ± 1.27	86.66 ± 0.19	90.88 ± 0.10	76.42
	Food101	57.01 ± 1.68	69.75 ± 0.88	78.33 ± 0.88	82.55 ± 0.40	85.49 ± 0.32	74.63
	SUN397	44.84 ± 1.18	56.53 ± 0.59	64.74 ± 0.38	70.44 ± 0.23	73.73 ± 0.31	62.06
	ImageNet	47.45 ± 0.67	59.05 ± 0.74	67.57 ± 0.27	72.63 ± 0.35	75.91 ± 0.28	64.52
	mean	**60.25**	**70.14**	**78.11**	**82.78**	**85.85**	**75.43**
Δ TAP − probe, paired	OxfordPets	$+4.17 \pm 0.51$	$+1.97 \pm 1.38$	$+0.75 \pm 0.39$	-0.03 ± 0.45	$+0.04 \pm 0.56$	$+1.38$
	DescribableTextures	$+0.49 \pm 0.57$	$+0.18 \pm 0.33$	-0.30 ± 0.87	-0.95 ± 0.37	-1.06 ± 0.31	-0.33
	EuroSAT	$+4.38 \pm 5.21$	$+4.67 \pm 1.64$	$+1.21 \pm 1.33$	$+1.16 \pm 0.69$	$+1.12 \pm 1.46$	$+2.51$
	Caltech101	$+0.91 \pm 0.29$	$+0.16 \pm 0.51$	-0.05 ± 0.41	$+0.03 \pm 0.34$	$+0.04 \pm 0.37$	$+0.22$
	FGVCAircraft	$+0.51 \pm 0.52$	-1.83 ± 2.03	-1.35 ± 0.20	-2.86 ± 1.13	-3.35 ± 1.50	-1.78
	UCF101	$+2.80 \pm 1.39$	$+1.01 \pm 0.54$	$+0.48 \pm 0.77$	-0.04 ± 0.41	$+0.41 \pm 0.38$	$+0.93$
	OxfordFlowers	$+0.30 \pm 0.26$	$+0.00 \pm 0.47$	-0.00 ± 0.11	$+0.04 \pm 0.04$	$+0.04 \pm 0.08$	$+0.08$
	StanfordCars	$+7.14 \pm 1.22$	$+4.83 \pm 0.77$	$+3.93 \pm 0.45$	$+2.19 \pm 0.29$	$+1.14 \pm 0.33$	$+3.85$
	Food101	$+5.96 \pm 0.50$	$+2.81 \pm 0.54$	$+0.96 \pm 0.25$	$+0.46 \pm 0.16$	$+0.73 \pm 0.22$	$+2.18$
	SUN397	$+5.96 \pm 0.43$	$+3.69 \pm 0.30$	$+1.86 \pm 0.16$	$+0.93 \pm 0.11$	$+1.08 \pm 0.17$	$+2.70$
	ImageNet	$+8.44 \pm 0.56$	$+5.32 \pm 0.35$	$+2.71 \pm 0.11$	$+1.54 \pm 0.13$	$+1.05 \pm 0.07$	$+3.81$
	mean	$+3.73$	$+2.07$	$+0.93$	$+0.22$	$+0.11$	$+1.41$

Dimension reduction ($D_{\text{txt}} > D_{\text{img}}$). This case arises for DINOv3-S ($D_{\text{img}} = 384$) with CLIP-B/16 text ($D_{\text{txt}} = 512$) in the vision-scaling sweep and for DINOv3-B ($D_{\text{img}} = 768$) with Qwen3-Embedding-0.6B ($D_{\text{txt}} = 1024$); mini-GTE and SigLIP2 text are 768-d, so with DINOv3-B the map is square, and CLIP-B/16 text (512-d) gives the isometric-embedding case of Eq. 2. We keep the SVD solution $\mathbf{W}^\star = \mathbf{V}\mathbf{U}^\top$, now with orthonormal rows ($\mathbf{W}\mathbf{W}^\top = \mathbf{I}_{D_{\text{img}}}$). Under this constraint $\|\mathbf{T}\mathbf{W}^\top\|_F^2$ is no longer constant, so $\mathbf{W}^\star$ maximizes the cross term $\mathrm{tr}(\mathbf{W}\mathbf{T}^\top\mathbf{X})$ rather than minimizing the least-squares objective of Eq. 2; it is the standard semi-orthogonal Procrustes solution and is the map we use. It projects the text space onto a D_{img}-dimensional subspace before rotating it, so only the geometry within that subspace is preserved, and the renormalization in Eq. 3 is no longer a no-op.

G Visualizations

G.1 Transport trajectories

Fig. 2 shows the actual 768-d states projected with one fixed linear map (PCA on the class means of both modalities at the start and at the end of the two paths and on the support means), so distances along a path are faithful up to the projection; the projection preserves the modality-axis motion of the names at near full scale while compressing the within-class spread into a small blob, which is why a hollow star can appear one or two blob radii from its squares although it sits inside the class in 768-d. It uses the same construction on StanfordCars, where at 16 shots the alignment cosine is ρ=0.74 and the closed-form schedule of Eq. 8 gives τ=0.76. Data and plotting scripts are released with the code (`figs/teaser/make_teaser.py` for the panels above, `figs/teaser/make_teaser_tap.py` for Fig. 2).