

Rethinking Cross-Modality Prompt Transfer: Geometric Priors over Linguistic Semantics

Ningyuan Zhang, Jie Lu*, Keqiuyin Li, Guangquan Zhang

Australian Artificial Intelligence Institute, University of Technology Sydney

*Corresponding Author

ningyuan.zhang@student.uts.edu.au

{jie.lu, keqiuyin.li, guangquan.zhang}@uts.edu.au

Abstract

Cross-Modality Prompt Transfer (CMPT) has been explored to address the initialization instability and data scarcity challenges inherent in prompt tuning. Its empirical success is typically interpreted as the transfer of linguistic semantics acquired during language-based prompt pretraining. In this paper, we revisit this interpretation and explore what actually transfers when language-pretrained prompts are adapted to the visual modality. Through controlled empirical analyses, we show that transfer performance does not consistently correlate with source-task linguistic performance. Instead, transferability is better explained by the *geometric priors* of prompt vectors, particularly stable rank and spatial direction relative to target features. To further validate this interpretation, we directly construct prompts with favorable geometry without relying on language pretraining. The resulting prompts achieve competitive transfer performance against language-pretrained prompts across diverse visual tasks, supporting prompt geometry as a stronger empirical explanation of transferability.

1 Introduction

As the scale of pretrained Transformers (Vaswani et al., 2017) continues to increase, finetuning them on downstream tasks has become prohibitively expensive. Parameter-Efficient Transfer Learning (PETL; Ding et al., 2023) has been explored to mitigate this challenge. Among PETL methods, prompt tuning (Lester et al., 2021) stands out for its exceptional parameter efficiency and simplicity, as it optimizes only a small set of vectors prepended to the layer inputs. Despite its effectiveness, prompt tuning is highly sensitive to initialization and often struggles when target tasks suffer from data scarcity (Wang et al., 2024b; Su et al., 2022).

To address these limitations, prompt transfer (Vu et al., 2022; Su et al., 2022; Asai et al., 2022) pretrains soft prompts on a data-rich source task

and uses them to initialize prompts for a target task. However, conventional prompt transfer assumes that the source and target tasks share the same modality, which limits its applicability when no modality-compatible source task is available. Cross-Modality Prompt Transfer (CMPT; Zhang et al., 2025) relaxes this constraint by transferring prompts pretrained on language tasks to non-language target tasks through learnable projectors, achieving substantial improvements in data-scarce settings. Despite its empirical success, the underlying mechanism remains poorly understood.

Consistent with prior work, a natural interpretation of CMPT is that language-pretrained prompts carry linguistic semantics that benefit the target task (Zhang et al., 2025). In this paper, we revisit this interpretation by asking which properties learned by a soft prompt during language pretraining remain useful after transfer to the visual modality. Through controlled empirical studies, we observe a decoupling between source-task linguistic performance and transfer performance: prompts with stronger source performance do not necessarily transfer better. These results suggest that source-task linguistic performance alone does not reliably explain cross-modality prompt transferability.

What, then, drives CMPT’s empirical success? Our analysis identifies the *geometric priors* of prompt vectors as a stronger empirical explanation. Specifically, we identify two geometric descriptors associated with transfer performance: (i) stable rank, which reflects the directional diversity and spectral spread of the prompt vectors; and (ii) spatial direction, which captures the orientation of prompt vectors relative to target features. Prompts with higher stable rank and spatial directions biased away from the dominant target-feature directions tend to transfer more effectively.

This geometric insight suggests that language pretraining is not the only way to obtain transferable prompts. To further validate our analysis, we

construct prompts directly from geometric priors. The proposed paradigm, Geometric Prompt Initialization (GPI), autonomously synthesizes prompts with high stable rank and controlled spatial direction and then transfers them to visual tasks and models through the standard CMPT projection-transfer pipeline. GPI achieves competitive performance against language-pretrained prompts without requiring source prompt pretraining.

In summary, our contributions are: (1) We revisit the semantic-transfer interpretation of CMPT and show that source-task linguistic performance does not reliably predict cross-modality transferability, challenging the view that language-pretrained prompts benefit visual tasks by carrying transferable linguistic semantics. (2) We identify stable rank and spatial direction as two geometric descriptors that provide a stronger empirical explanation of cross-modality prompt transferability. (3) We validate this geometric insight via GPI, showing that prompts constructed directly from geometric priors can compete with language-pretrained prompts across diverse visual tasks. Together, these findings support prompt geometry as a stronger empirical explanation of transferability and offer an efficient alternative for transferable prompt initialization.

2 Related Works

2.1 Prompt Tuning

Prompt tuning (Lester et al., 2021; Li and Liang, 2021; Wang et al., 2023) adapts pretrained Transformers by optimizing only a small set of continuous prompt vectors. Originally developed for language models, prompt tuning has been extended to the visual domain through Visual Prompt Tuning (VPT; Jia et al., 2022), which prepends prompts to the inputs of Vision Transformers (ViTs; Dosovitskiy et al., 2021). Subsequent methods improve VPT from different perspectives: E²VPT (Han et al., 2023) introduces learnable key-value prompts into self-attention modules, VFPT (Zeng et al., 2024) incorporates spatial and frequency domain information, and VAPT (Le et al., 2026) generates input-adaptive prompts from image features. Prompt tuning has also been extended to pretrained vision-language models such as CLIP (Radford et al., 2021): MiraGe (Shi et al., 2025) adapts CLIP’s text and image branches through multi-modal prompts for generalizable AI-generated image detection, while SCP (Wang et al., 2024a) updates CLIP’s text prompts at test time to improve

zero-shot robustness under distribution shifts.

Prior work improves prompt tuning through architectural redesigns (Shi and Lipani, 2024; Han et al., 2023; Le et al., 2026) and target-side adaptation strategies (Zeng et al., 2024). In contrast, our work focuses on prompt initialization: We ask why prompts pretrained on language tasks can benefit target visual tasks. Rather than proposing another VPT architecture, we analyze which properties of language-pretrained prompts remain useful after cross-modality transfer and show that their transferability is better explained by geometric priors.

2.2 Prompt Transfer

Prompt transfer seeks to alleviate the initialization instability of prompt tuning by pretraining prompts on data-rich source tasks and using them to initialize prompts for downstream tasks. SPoT (Vu et al., 2022) demonstrated that prompt pretraining can improve transfer across language tasks. Subsequent work extended prompt transfer to cross-model settings (Su et al., 2022), showing that pretrained prompts can remain effective even when the source and target language models differ.

Cross-Modality Prompt Transfer (Zhang et al., 2025) further extends prompt transfer across modalities by adapting language-pretrained prompts to another modality. Its empirical success motivates the interpretation that linguistic semantics acquired during source pretraining can benefit non-text target tasks. Our work revisits this interpretation. Through controlled analyses, we show that source linguistic performance does not reliably predict cross-modality transfer performance. Instead, transferability is more strongly associated with geometric properties of the prompt vectors, particularly stable rank and spatial direction relative to target features. To further validate this interpretation, we construct prompts directly from these geometric priors without relying on language pretraining. The resulting Geometric Prompt Initialization (GPI) shows that prompts with favorable geometry can match or outperform language-pretrained prompts under the same transfer pipeline.

3 Cross-Modality Prompt Transfer

CMPT (Zhang et al., 2025) alleviates the initialization instability and data scarcity issues of prompt tuning by transferring prompts pretrained on data-rich language tasks. It consists of two phases:

Pretraining. A source prompt p_s is pretrained on a data-rich language task using a pretrained language model h_s . Here, $p_s \in \mathbb{R}^{N_l \times N_p \times d_h}$, where N_l , N_p , and d_h denote the number of transformer layers, the number of prompt vectors, and the hidden dimension of h_s , respectively.

Transfer. The pretrained source prompt p_s is adapted to the target visual task and model h_t through N_l learnable linear projectors $\{P^{(i)}\}_{i=1}^{N_l}$, with the following objective:

$$\begin{aligned} Z^{(1)} &= h_t^{(1)} \left(\left[P^{(1)}(p_s^{(1)}), X_t \right] \right) \\ Z^{(i)} &= h_t^{(i)} \left(\left[P^{(i)}(p_s^{(i)}), Z^{(i-1)} \right] \right) \\ \arg \min_{\{P^{(i)}, C\}} \mathcal{L} \left(C(Z^{(N_l)}), Y_t \right) \end{aligned} \quad (1)$$

where $[\cdot, \cdot]$ denotes vector concatenation; $h_t^{(i)}$ and $p_s^{(i)}$ denote the i -th layer of the target model and the i -th layer of the source prompt, respectively; $Z^{(i)}$ is the hidden state produced by the i -th target layer; X_t and Y_t are the target inputs and labels; and C denotes the learnable classification head.

This formulation raises the central question of this paper: which properties of p_s make it transferable from language to vision? A common interpretation is that p_s transfers linguistic semantics acquired during language-based prompt pretraining. In the following sections, we revisit this interpretation and demonstrate that the geometric priors of p_s provide a stronger empirical explanation of transferability than source-task linguistic performance.

4 Methodology

In this section, we conduct controlled analyses to examine the relationship between transfer performance and source-task linguistic performance. We then show that transferability is better explained by the prompt’s *geometric priors*, particularly stable rank and spatial direction. Finally, we validate this geometric explanation by transferring prompts constructed directly from these geometric priors.

Our analyses are conducted on six VTAB-1K datasets (Zhai et al., 2019), spanning the natural, specialized, and structured VTAB groups while keeping the controlled analyses tractable. The source prompts, each with $N_p = 20$, are pretrained on SNLI (Bowman et al., 2015) using RoBERTa-base (Liu et al., 2019) or GPT2-base (Radford et al., 2019). The target model is a ViT-base (Dosovitskiy

et al., 2021) pretrained on ImageNet-21k (Deng et al., 2009) with a supervised objective.

4.1 Linguistic Semantics

A common assumption in transfer learning and prompt transfer is that transferability arises from knowledge acquired during source pretraining (Long et al., 2015; Vu et al., 2022). We use source-task performance as an empirical proxy for the amount of linguistic semantics acquired during prompt pretraining. In CMPT, this assumption implies that prompts with higher source-task performance should also transfer more effectively to target vision tasks. To test this hypothesis, we evaluate prompts pretrained using different fractions of SNLI data and different training durations:

- E1 and E64: pretrained on **100%** of SNLI for **1** and **64** epochs, respectively.
- E100 and E6400: pretrained on **1%** of SNLI for **100** and **6400** epochs, respectively.

By varying the amount of pretraining data, we vary the linguistic supervision available to the prompts. The **1%** series receives substantially less linguistic supervision and therefore reaches lower source accuracy, while maintaining the same number of gradient updates as the **100%** series.

Figure 1a shows the transfer performance of these prompts relative to their source performance. It yields a key observation: **Higher source accuracy does not reliably predict better transfer performance**. For example, although RoBERTa-E6400 achieves the lowest source accuracy among the RoBERTa prompts, it yields transfer performance comparable to or better than the other RoBERTa prompts. Similarly, GPT2-E6400 outperforms GPT2-E64 on five of the six target tasks despite having substantially lower source accuracy. Crucially, the RoBERTa prompts with source accuracy above 80% do not consistently outperform the prompt pretrained on random data, which achieves only random-guess-level SNLI accuracy ($\sim 33.3\%$). These results show that source-task performance on SNLI is not a reliable predictor of language-to-vision prompt transfer performance.

4.2 Geometric Priors

4.2.1 Stable Rank

Empirical Studies. The preceding finding motivates us to examine whether language-pretrained prompts are transferable due to their non-semantic

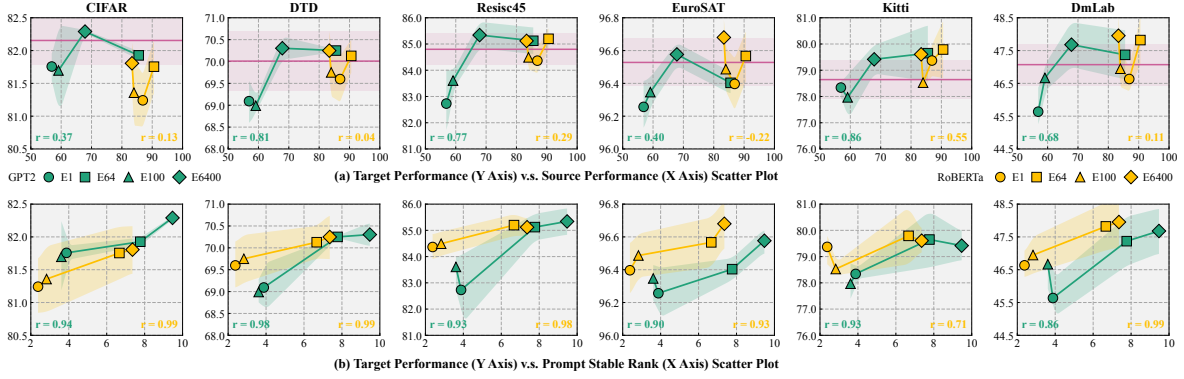


Figure 1: **(a) Transfer Performance vs. Source Performance:** Source-task linguistic performance does not consistently predict transfer performance. The horizontal line denotes the transfer performance of a prompt trained with RoBERTa for 64 epochs on randomly generated tokens and labels, with the number of training samples matched to that of SNLI. The r value denotes the Pearson correlation coefficient between the x and y axes, with larger values indicating stronger correlation. **(b) Transfer Performance vs. Prompt Stable Rank:** Transfer performance is more consistently correlated with prompt stable rank.

Norm	Prompt	Stb. Rank	CIFAR _{24}	DTD _{23}	Resisc45 _{25}	EuroSAT _{26}	Kitti _{20}	DmLab _{26}
7.85	E1	2.38	81.14 ₍₂₎	69.51 ₍₃₎	84.41 ₍₃₎	96.40 ₍₂₎	79.52 ₍₃₎	47.34 ₍₂₎
	E1+Rank(E64)	6.53	81.77 ₍₆₎	69.92 ₍₅₎	84.64 ₍₅₎	96.51 ₍₆₎	79.71 ₍₅₎	47.70 ₍₆₎
	E64	6.67	81.40 ₍₅₎	70.35 ₍₇₎	84.67 ₍₆₎	96.58 ₍₆₎	78.94 ₍₄₎	47.80 ₍₆₎
	E64+Rank(E1)	2.40	81.24 ₍₃₎	69.59 ₍₁₎	84.43 ₍₂₎	96.45 ₍₂₎	78.83 ₍₄₎	47.02 ₍₂₎
31.08	E1	2.38	81.45 ₍₃₎	69.61 ₍₃₎	84.56 ₍₀₎	96.43 ₍₁₎	79.15 ₍₂₎	46.89 ₍₁₎
	E1+Rank(E64)	6.53	81.81 ₍₅₎	69.84 ₍₅₎	85.34 ₍₈₎	96.68 ₍₇₎	79.82 ₍₆₎	47.92 ₍₇₎
	E64	6.67	81.93 ₍₈₎	70.05 ₍₆₎	85.30 ₍₆₎	96.70 ₍₇₎	78.99 ₍₅₎	47.69 ₍₇₎
	E64+Rank(E1)	2.40	81.39 ₍₀₎	69.83 ₍₂₎	84.53 ₍₂₎	96.44 ₍₁₎	79.08 ₍₃₎	47.14 ₍₁₎

Table 1: **Singular-value-swap analysis.** We swap the singular values of E1 and E64, yielding E1+Rank(E64) and E64+Rank(E1) while preserving the original singular vectors. The resulting prompts are further normalized to have a uniform vector norm, set to the average norm of E1 (7.85) or E64 (31.08). The best result for each target task is shown in **bold**. Each result is averaged across 8 seeds. $(\cdot)$ denotes the number of seed-level wins against the rank-modified counterpart. $\{\cdot\}$ denotes the number of seed-level wins of Rank(E64) within each target task.

geometric properties. Petrov et al. (2024) showed that a prompt’s ability to modify model outputs is bounded by its rank. Motivated by this, we measure the stable rank of pretrained prompts, defined as the ratio between the sum of squared singular values and the largest squared singular value (Rudelson and Vershynin, 2007). Stable rank captures how evenly the prompt energy is distributed across singular directions. We use it as a continuous proxy for the algebraic rank. As shown in Figure 1b, we observe a stronger correlation between transfer performance and stable rank.

However, the prompts in Figure 1 differ not only in stable rank, but also in other geometric properties, such as vector norm and subspace orientation. Therefore, Figure 1b does not isolate the effect of stable rank. To this end, we perform a singular-value-swap analysis between E1 and E64,

both pretrained with RoBERTa. Specifically, we first swap their singular values to exchange their spectral spread while preserving their orthogonal bases. We then normalize the prompt vectors to a uniform norm at each layer, using the average layer-wise norm of either E1 or E64. As a result, the modified prompts mainly differ in stable rank.

As shown in Table 1, **higher stable rank generally leads to better transfer performance.** When stable rank is isolated as the primary variable, 23 of 24 pairs show that higher stable rank yields higher target accuracy. At the seed level, 144 of 192 comparisons show that prompts with higher stable rank outperform their lower-rank counterparts, yielding a p -value of 1.22×10^{-12} under a one-sided binomial sign test. Moreover, the best transfer performance across all target tasks is achieved using the stable rank of the E64 prompt.

Theoretical Analysis. To formally ground these observations, we provide a theoretical analysis based on Petrov et al. (2024), who showed that prepending prompt vectors is equivalent to adding a bias term to the output of an attention block:

$$\mathcal{O}_p = \tilde{\mathcal{O}} + \beta_p \quad (2)$$

where $\tilde{\mathcal{O}}$ denotes the original output without prompt tokens, β_p is the additive bias introduced by the prompt vectors, and $\mathcal{O}_p$ is the modified output. Here, $\mathcal{O}_p, \tilde{\mathcal{O}}, \beta_p \in \mathbb{R}^{N_I \times d_h}$, where N_I denotes the number of input embeddings.

The i -th row of the bias matrix β_p , denoted by β_p^i , is a weighted sum of the prompt vectors p projected by the attention block’s value matrix W_V . The weights are given by the attention scores $A(\cdot, \cdot)$ computed between the i -th input embedding x^i and each prompt vector $[p^1, p^2, \dots, p^{N_p}]$:

$$\beta_p^i = \sum_{j=1}^{N_p} A(x^i, p^j) \times p^j W_V \quad (3)$$

Extending Equation (3) to CMPT gives:

$$\beta_p^i = \sum_{j=1}^{N_p} A(x^i, p_s^j W_P) \times p_s^j W_P W_V \quad (4)$$

where W_P is the weight matrix of the learnable linear prompt projector used in CMPT. The basic *rank product inequality* implies that the expressiveness of the bias term β_p is bounded by:

$$r(p_s W_P W_V) \leq \min(r(p_s), r(W_P), r(W_V)) \quad (5)$$

where $r(\cdot)$ denotes matrix rank. In Equation (5), $r(W_V)$ is fixed because the pretrained Vision Transformer is frozen. By contrast, $r(W_P)$ can evolve during transfer because the projection matrix W_P is learnable. Therefore, if the pretrained source prompt p_s has a higher rank, the theoretical upper bound on the dimensionality of the attention bias β_p increases accordingly. This offers an interpretation for why prompt matrices with greater stable rank (the empirical proxy for measuring algebraic rank in continuous space) are associated with better transfer performance, as observed empirically in Table 1. The full theoretical analysis is provided in Appendix E.

4.2.2 Spatial Direction

Stable rank captures the spectral spread of a prompt, but it does not describe the absolute orientation of

prompt vectors in the target space. We therefore introduce *spatial direction* to measure how prompt vectors are oriented relative to target features.

We use the principal components (PCs) of target features as a reference frame, as they characterize the dominant directions of the target data. Specifically, we feed target data into the ViT, extract layer-wise input embeddings, and apply SVD to obtain the target-feature orthogonal bases $V_t \in \mathbb{R}^{N_t \times d_h \times d_h}$. Given a prompt p_s , we compute its Center of Mass (CoM) with respect to V_t :

$$\begin{aligned} \text{CoM}^{(i)} &= \sum_{k=0}^{d_h-1} k \left(\frac{\sum_{x=1}^{N_p} C_{x,k}^{(i)}}{\sum_{k'=1}^{d_h} \sum_{x=1}^{N_p} C_{x,k'}^{(i)}} \right) \\ \text{CoM} &= \frac{1}{N_I} \sum_{i=1}^{N_I} \text{CoM}^{(i)} \end{aligned} \quad (6)$$

where $C_{x,k}^{(i)}$ denotes the entry in the x -th row and k -th column of the squared cosine-similarity matrix $C^{(i)} = \cos^2(p_s^{(i)}, V_t^{(i)}) \in \mathbb{R}^{N_p \times d_h}$. The superscript (i) denotes the i -th layer. The resulting CoM lies in $[0, d_h - 1]$ and indicates whether prompt vectors align more with the leading target-feature PCs (smaller) or with directions farther from the leading PCs (larger). This scalar respects the ordered structure of target-feature PCs and provides an interpretable measure of prompt orientation.

To isolate spatial direction as the sole independent variable, we construct two comparison sets:

1. V(E1) vs. V(E64): prompts constructed using the right singular vectors of E1 and E64. This set tests whether natural differences in prompt orientation affect transfer performance.
2. V(TOP20) vs. V(BOT20): prompts constructed using the top and bottom 20 PCs in V_t . This set magnifies the CoM discrepancy to further validate our geometric hypothesis.

All synthesized prompts are scaled to match the average layer-wise norm of the E64 prompt. Within each comparison set, prompts have the same stable rank and norm, leaving spatial direction (CoM) as the only varying geometric factor. The transfer results are reported in Table 2, from which two observations can be made.

First, prompts with directions farther from the dominant PCs (larger CoM) are generally more transferable. The clearest evidence comes from the controlled V(TOP20) vs. V(BOT20) comparison, where the CoM gap is manually enlarged. In

Prompt	CIFAR		DTD		Resisc45		EuroSAT		Kitti		DmLab	
	CoM	Acc.	CoM	Acc.	CoM	Acc.	CoM	Acc.	CoM	Acc.	CoM	Acc.
V(E1)	386.2	82.58 ₍₄₎	386.1	70.35 ₍₅₎	386.0	85.33 ₍₆₎	385.6	96.73 ₍₄₎	385.5	79.52 ₍₅₎	386.0	46.88 ₍₅₎
V(E64)	384.1	82.43 ₍₄₎	384.3	70.18 ₍₃₎	384.1	85.20 ₍₂₎	384.4	96.70 ₍₄₎	383.8	79.29 ₍₃₎	384.4	46.51 ₍₃₎
V(TOP20)	9.5	82.21 ₍₃₎	9.5	69.30 ₍₀₎	9.5	84.79 ₍₃₎	9.5	96.68 ₍₂₎	9.5	79.11 ₍₁₎	9.5	46.16 ₍₃₎
V(BOT20)	757.5	82.38 ₍₅₎	757.5	70.57 ₍₈₎	757.5	85.24 ₍₅₎	757.5	96.86 ₍₆₎	757.5	79.75 ₍₇₎	757.5	46.72 ₍₅₎

Table 2: **Transfer performance with respect to spatial direction** (CoM). Each result is averaged over 8 seeds. $(\cdot)$ denotes the number of seed-level wins.

this comparison, the higher-CoM prompt wins 36 of 48 seed-level comparisons ($p = 3.6 \times 10^{-4}$), providing strong evidence that directions farther from the dominant target PCs benefit transfer. The V(E1) vs. V(E64) comparison shows a similar but weaker trend: since their CoM values differ only slightly, the higher-CoM prompt wins 29 of 48 seed-level comparisons ($p = 0.097$), which should be interpreted as suggestive rather than conclusive evidence. Aggregating both comparisons gives 65 wins out of 96 seed-level comparisons ($p = 3.4 \times 10^{-4}$), supporting an overall positive correlation between CoM and transferability.

Second, language pretraining slightly shifts CoM toward a neutral state. V(E1) has a slightly larger CoM than V(E64), whose CoM is close to the neutral midpoint on every task ($\sim \frac{d_h-1}{2} = 383.5$). This suggests that language pretraining tends to de-bias the prompt’s spatial direction, potentially weakening its transferable orientation.

Overall, spatial direction provides complementary evidence for CMPT. Prompts that overlap heavily with dominant target PCs may provide redundant biases to the target model. By contrast, directions farther from the dominant PCs can provide complementary signals. Lu et al. (2024) also showed that constraining prompt updates to feature-orthogonal directions can reduce interference in continual learning. Together with the stable rank analysis, these results suggest that the geometric priors of the pretrained prompt provide a stronger empirical explanation of transferability than source-task linguistic performance.

4.3 Geometric Prompt Initialization

The preceding analyses suggest that the transferability of language-pretrained prompts is better explained by geometric priors than by source-task linguistic performance. If it is correct, then language pretraining should not be necessary for obtaining transferable prompts. Prompts constructed

directly with favorable geometry should also transfer effectively. We validate this hypothesis through Geometric Prompt Initialization (GPI), which synthesizes prompts with high stable rank and a spatial direction biased away from the target PCs.

To satisfy these two geometric conditions simultaneously, we treat the orthogonal basis of target features at layer l , denoted by $V_t^{(l)}$, as a vector pool. We then construct a matrix $U_c \in \mathbb{R}^{N_p \times d_h}$ and generate the prompt for layer l as $U_c V_t^{(l)}$. By controlling the structure of U_c , we control both the spatial direction and the stable rank of the resulting prompt matrix.

Controlling Spatial Direction. To bias the synthesized prompt away from the dominant target-feature directions, we first apply a Gaussian envelope to a random matrix $R \sim \mathcal{N}(0, 1) \in \mathbb{R}^{N_p \times d_h}$:

$$U_c = R \odot \underbrace{\exp \left(-0.5 \left(\frac{\mathbf{i} - c}{\sigma} \right)^2 \right)}_{\text{envelope}} \quad (7)$$

where $\mathbf{i} = [0, 1, \dots, d_h - 1]$ denotes the vector of principal component indices. σ controls the coverage of the envelope and is set to $d_h/2$. c controls the center of the envelope and is set to $d_h - 1$ to maximize the CoM.

The Gaussian envelope makes each synthesized vector a continuous weighted combination of the basis vectors in $V_t^{(l)}$. Unlike V(BOT20), which simply selects directions farthest from the leading PCs, GPI makes the synthesized prompts more similar to naturally pretrained representations, which also exist as linear combinations within the $V_t^{(l)}$ space.

Controlling Stable Rank. The matrix U_c obtained from Equation (7) is not guaranteed to have high stable rank. We therefore enforce row orthogonality via Löwdin orthogonalization:

$$U, \Sigma, V^T = \text{SVD}(U_c) \\ U_c = UV^T \quad (8)$$

This produces the orthogonal matrix closest to the enveloped matrix in Frobenius norm, increasing spectral spread while approximately preserving the intended spatial bias. The prompt at layer l is then computed as $U_c V_t^{(l)}$. We repeat this procedure for each target layer to obtain the final prompt. The complete algorithm is provided in Algorithm 1.

Algorithm 1 Geometric Prompt Initialization

Input: Target task $\mathcal{T}$, target model h_t

Output: Synthesized prompt $\hat{p}_s$

```

1: procedure GENERATE  $U_c(c, \sigma)$ 
2:   Sample  $R \sim \mathcal{N}(0, 1) \in \mathbb{R}^{N_p \times d_h}$ 
3:   Apply Equations (7) and (8) to obtain  $U_c$ 
4:   return  $U_c$ 
5: end procedure

6: Obtain  $V_t \in \mathbb{R}^{N_l \times d_h \times d_h}$  for  $\mathcal{T}$  and  $h_t$ 
7:  $c \leftarrow d_h - 1$ ;  $\sigma \leftarrow d_h/2$ 
8:  $\hat{p}_s \leftarrow \{\}$  ▷ Initialize an empty list
9: for layer  $l = 0$  to  $N_l - 1$  do
10:   $U_c^{(l)} \leftarrow \text{GENERATE } U_c(c, \sigma)$ 
11:   $p^{(l)} \leftarrow U_c^{(l)} V_t^{(l)}$ 
12:  Append  $p^{(l)}$  to  $\hat{p}_s$ 
13: end for
14: return  $\hat{p}_s \in \mathbb{R}^{N_l \times N_p \times d_h}$ 

```

Once the prompt is constructed, its magnitude is scaled by multiplying prompt vectors at the i -th layer by $\alpha \cdot n_t^{(i)}$, where $n_t^{(i)}$ is the average norm of target features at layer i , and α is selected by grid search on the validation split. Finally, the norm-scaled prompt is transferred to the target task using projection transfer formulated in Equation (1).

5 Experiments

5.1 Experimental Setup

Target Datasets. We evaluate whether prompts constructed from geometric priors can match or exceed language-pretrained prompts within the CMPT projection-transfer pipeline. Following prior CMPT work, we use the VTAB-1K benchmark (Zhai et al., 2019), which contains 19 data-scarce tasks grouped into three categories: (1) *Natural* tasks, which contain images captured using standard cameras: CIFAR100 (Krizhevsky, 2009), Caltech101 (Li et al., 2004), DTD (Cimpoi et al., 2014), Flowers102 (Nilsback and Zisserman, 2008), Pets (Parkhi et al., 2012), SVHN (Netzer et al., 2011), and SUN397 (Xiao et al.,

2010); (2) *Specialized* tasks, which contain images captured using specialized equipment (e.g., medical or satellite imaging): Patch Camelyon (Veeling et al., 2018), EuroSAT (Helber et al., 2018), Resisc45 (Cheng et al., 2017), and Diabetic Retinopathy (Dugas et al., 2015); and (3) *Structured* tasks, which require geometric understanding (e.g., distance estimation or object counting): Clevr (Johnson et al., 2017), DmLab (Zhai et al., 2019), KITTI (Geiger et al., 2012), dSprites (Matthey et al., 2017), and SmallNORB (LeCun et al., 2004). This evaluation is designed to verify our central claim: if CMPT is driven primarily by prompt geometry, then prompts constructed directly from geometric priors should transfer competitively against language-pretrained prompts. VTAB-1K provides a controlled, data-scarce, non-language testbed for isolating the contribution of language pretraining.

Baselines. The key comparison is with CMPT-Deep, to test whether language pretraining is necessary when target geometric priors are available, as GPI differs from CMPT-Deep only in prompt initialization. We also compare against representative parameter-efficient transfer learning baselines: (1) full finetuning (*Full*), which updates the entire model; (2) classic parameter-efficient transfer learning methods, including Bias (Rebuffi et al., 2017) and Adapter (Houlsby et al., 2019); (3) Visual Prompt Tuning with a deep architecture (VPT-Deep; Jia et al., 2022); and (4) recent deep visual prompt-tuning methods, including E²VPT (Han et al., 2023), VFPT (Zeng et al., 2024), and VAPT (Le et al., 2026). All experiments use ViT-B/16 (Dosovitskiy et al., 2021) under three pretraining objectives: supervised pretraining on ImageNet-21k (Deng et al., 2009), and self-supervised pretraining on ImageNet-1k using MAE (He et al., 2022) and MoCo v3 (Chen et al., 2021).

Hyperparameters. Each VTAB-1K training set is split into 800 training and 200 validation samples for hyperparameter tuning. A grid search is performed to determine the optimal learning rate, weight decay, and norm scalar α . All models are trained for 100 epochs with 3 seeds. The learning rate is warmed up over the first 10 epochs and decayed to zero using a cosine annealing schedule.

Reproducibility. To ensure reproducibility, we implement our method using the publicly available deep learning framework PyTorch (Paszke et al., 2019). We strictly adhere to PyTorch’s official

	ViT-Supervised				ViT-MAE				ViT-MoCo			
	<i>Natural</i> (7)	<i>Specialized</i> (4)	<i>Structured</i> (8)	AVG.	<i>Natural</i> (7)	<i>Specialized</i> (4)	<i>Structured</i> (8)	AVG.	<i>Natural</i> (7)	<i>Specialized</i> (4)	<i>Structured</i> (8)	AVG.
<i>Full</i>	75.88	83.36	47.64	65.57	59.31	79.68	53.82	61.29	71.95	84.72	51.98	66.22
Bias Adapter	73.30 ₍₃₎ 70.67 ₍₄₎	78.25 ₍₀₎ 77.80 ₍₀₎	44.09 ₍₂₎ 33.09 ₍₀₎	62.05 62.41	54.55 ₍₁₎ 54.90 ₍₃₎	75.68 ₍₁₎ 75.19 ₍₁₎	47.70 ₍₀₎ 38.98 ₍₀₎	56.11 52.47	72.89 ₍₃₎ 74.19 ₍₄₎	81.14 ₍₀₎ 82.66 ₍₁₎	53.43 ₍₄₎ 47.69 ₍₂₎	66.43 64.82
VPT-Deep	78.48 ₍₆₎	82.43 ₍₂₎	54.98 ₍₈₎	69.43	36.02 ₍₀₎	60.61 ₍₁₎	26.57 ₍₀₎	37.22	70.27 ₍₄₎	83.04 ₍₀₎	42.38 ₍₀₎	61.22
E ² VPT	80.01 ₍₆₎	84.43 ₍₃₎	57.39 ₍₈₎	71.42	59.52 ₍₄₎	77.80 ₍₁₎	44.65 ₍₃₎	57.11	76.47 ₍₄₎	87.28 ₍₂₎	54.91 ₍₆₎	69.67
VFPT	81.35 ₍₆₎	84.93 ₍₄₎	60.19 ₍₈₎	73.20	53.59 ₍₆₎	77.75 ₍₁₎	36.15 ₍₁₎	51.33	77.47 ₍₅₎	85.76 ₍₃₎	58.74 ₍₆₎	71.33
VAPT	81.43 ₍₆₎	85.13 ₍₄₎	59.34 ₍₈₎	72.91	59.23 ₍₅₎	80.73 ₍₂₎	47.24 ₍₂₎	58.71	77.69 ₍₆₎	83.95 ₍₂₎	60.74 ₍₇₎	71.87
CMPT-Deep	81.46 ₍₆₎	85.37 ₍₄₎	60.63 ₍₈₎	73.51 $\pm$ 0.59	55.54 ₍₄₎	81.37 ₍₂₎	54.18 ₍₄₎	60.41 $\pm$ 3.02	78.18 ₍₆₎	86.13 ₍₄₎	60.69 ₍₇₎	72.49 $\pm$ 0.64
GPI	81.96 ₍₆₎ [6]	85.53 ₍₄₎ [3]	61.19 ₍₈₎ [7]	73.97 $\pm$ 0.34	57.89 ₍₅₎ [6]	82.34 ₍₄₎ [3]	56.35 ₍₅₎ [7]	62.39 $\pm$ 2.06	78.50 ₍₆₎ [7]	86.25 ₍₄₎ [3]	61.65 ₍₇₎ [6]	73.04 $\pm$ 0.48

Table 3: **VTAB-1k Performance** using a **ViT-Base/16** backbone. $\{\cdot\}$ denotes the number of datasets within each group. $(\cdot)$ refers to the number of wins compared to full finetuning (*Full*). $[\cdot]$ refers to the number of wins compared to CMPT-Deep. GPI’s best results across all baselines are in **bold**.

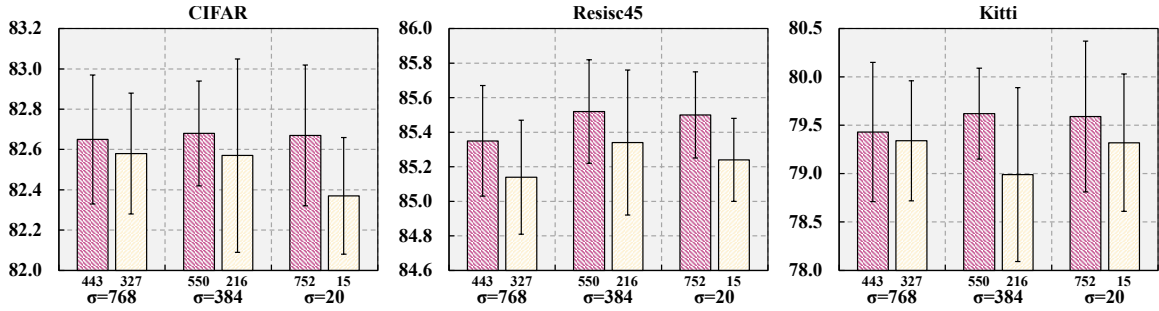


Figure 2: Ablation study on the spatial direction hyperparameters of GPI. The coverage σ is varied in $\{20, 384, 768\}$, ranging from a narrow envelope to a broad envelope over the target-feature basis. The center c is set to either 0 ($\square$) or 767 ($\blacksquare$) to induce the maximum CoM discrepancy (CoM values are marked under the bar).

guidelines to fix random seeds and minimize non-determinism. Our code is available at [this link](#).

5.2 Results and Analysis

Overall Performance. Table 3 reports the overall performance. Across the three backbones, GPI improves over CMPT-Deep by 0.46% on ViT-Supervised, 1.98% on ViT-MAE, and 0.55% on ViT-MoCo. GPI also outperforms CMPT-Deep on 48 of 57 tasks and in 121 of 171 seed-level comparisons, yielding a p -value of 2.81×10^{-8} (details are provided in Appendix D.4). These results suggest that language pretraining is not necessary for competitive transfer, as prompts constructed directly from geometry can exceed language-pretrained prompts across diverse scenarios. Appendix F provides a GPU-time comparison between GPI and SNLI pretraining, showing that GPI reduces the GPU time by approximately $9000\times$.

Ablation Study. We further ablate the two hyperparameters of GPI that control prompt spatial direction: the Gaussian width σ and the center position c . We vary $\sigma \in \{20, 384, 768\}$, where $\sigma = 20$ gives

the narrowest envelope because we use 20 prompt vectors per layer, and $\sigma = 768$ gives the widest envelope over the target-feature basis. We also vary $c \in \{0, d_h - 1\}$ to induce the largest possible CoM difference, corresponding to prompts biased toward the leading and farthest target-feature PCs, respectively. As shown in Figure 2, in the studied setting, two observations can be made. First, larger CoM consistently leads to better performance: under every σ value, setting $c = d_h - 1$ yields higher CoM and better accuracy than setting $c = 0$. Second, GPI is not overly sensitive to the exact choice of σ . When $c = d_h - 1$, performance remains stable across a wide range of σ values. These results suggest that, in the studied setting, GPI mainly requires a coarse directional bias away from the dominant PCs, rather than precise tuning of the spatial-direction hyperparameters.

Prompt Geometric Dynamics. Figure 3 shows how the prompt geometry evolves during SNLI prompt pretraining. We observe two trends: (i) prompt stable rank increases steadily as pretraining proceeds, indicating that source optimization

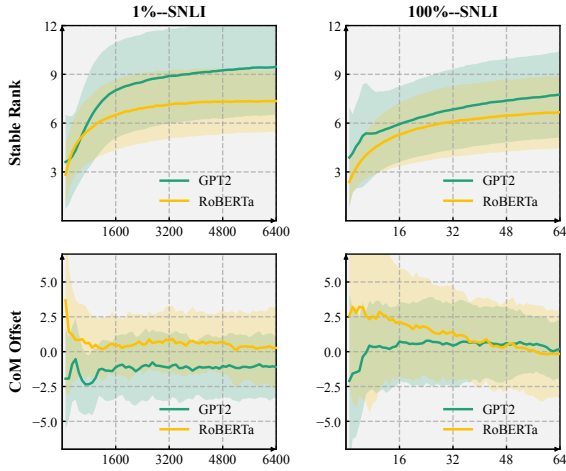


Figure 3: Evolution of prompt stable rank and CoM offset during source prompt pretraining. The CoM offset is computed as $\text{CoM} - \frac{d_h - 1}{2}$ using CIFAR as the target task. A positive offset indicates that prompt vectors are biased toward directions farther from the leading target-feature PCs, whereas a negative offset indicates stronger alignment with the leading PCs. The shaded region denotes the standard deviation across layers.

gradually expands the spectral spread of the prompt matrix; and (ii) the CoM offset progressively drifts toward zero, regardless of whether the prompt starts with a positive or negative spatial bias. This suggests that language pretraining tends to neutralize the spatial bias of prompts, driving their orientation toward the center of the target-feature basis.

These dynamics provide additional insight into why language-pretrained prompts are cross-modality transferable. Language pretraining indirectly reshapes prompt geometry by increasing stable rank, which benefits transfer. However, it also neutralizes directional bias, even though our spatial-direction analysis shows that a bias away from the dominant target PCs can be beneficial. Thus, source pretraining can be interpreted as an indirect and inefficient mechanism for producing transferable prompt geometry: it increases stable rank, but also drives spatial direction toward a neutral, median orientation rather than preserving or inducing the directional bias that benefits transfer. In Appendix F, we provide a case study to further examine how prompt geometry is shaped during language-based prompt pretraining.

6 Conclusions

In this paper, we revisit what transfers when language-pretrained prompts are adapted to the visual modality through Cross-Modality Prompt

Transfer (CMPT). Our controlled analyses show that, in the studied language-to-vision setting, transfer performance does not consistently correlate with source-task linguistic performance, while prompt geometry, particularly stable rank and spatial direction relative to target features, provides a stronger empirical explanation. To further validate this interpretation, we construct prompts directly from these geometric priors, without relying on language pretraining. The resulting paradigm, Geometric Prompt Initialization (GPI), achieves competitive performance against language-pretrained prompts across VTAB-1K tasks and ViT backbones while eliminating the cost of prompt pretraining. Together, these results support prompt geometry as a stronger empirical explanation of language-to-vision transferability and offer an efficient alternative for transferable prompt initialization.

Limitations

This work has several limitations. First, we use source performance as a proxy for linguistic semantics. This proxy may not capture all semantics encoded in a prompt, so our results cannot rule out every possible role of language pretraining. Instead, they show that source-task performance is not a reliable predictor of language-to-vision transferability. Second, our experiments focus on language-to-vision transfer. This setup does not cover all possible cross-modality scenarios, but it allows us to isolate the role of prompt geometry and compare directly with existing baselines. Whether the same geometric principles hold for other scenarios remains future work. Third, GPI removes source prompt pretraining but still relies on target feature statistics and validation-based norm selection. Thus, although it is cheaper than language prompt pretraining, it is not fully data-free. Fourth, our evidence supports prompt geometry within the studied CMPT projection-transfer pipeline but does not provide enough evidence to establish a pipeline-independent conclusion. Finally, stable rank and spatial direction may not capture all properties that influence transferability. Future work could investigate richer geometric or optimization-based explanations of CMPT.

Acknowledgments

The work was supported by the Australian Research Council (ARC) under Laureate Fellow grant FL190100149.

References

- Akari Asai, Mohammadreza Salehi, Matthew E Peters, and Hannaneh Hajishirzi. 2022. Attempt: Parameter-efficient multi-task tuning via attentional mixtures of soft prompts. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 6655–6672.
- Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 632–642. Association for Computational Linguistics (ACL).
- Xinlei Chen, Saining Xie, and Kaiming He. 2021. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649.
- Gong Cheng, Junwei Han, and Xiaoqiang Lu. 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883.
- Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.
- Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. [An image is worth 16x16 words: Transformers for image recognition at scale](#). In *International Conference on Learning Representations*.
- Emma Dugas, Jared, Jorge, and Will Cukierski. 2015. Diabetic retinopathy detection. <https://kaggle.com/competitions/diabetic-retinopathy-detection>. Kaggle.
- Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE.
- Cheng Han, Qifan Wang, Yiming Cui, Zhiwen Cao, Wenguan Wang, Siyuan Qi, and Dongfang Liu. 2023. E²vpt: An effective and efficient approach for visual prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17491–17502.
- Haoyu He, Jianfei Cai, Jing Zhang, Dacheng Tao, and Bohan Zhuang. 2023. Sensitivity-aware visual parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11825–11835.
- Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.
- Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2018. Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 204–207. IEEE.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.
- Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European conference on computer vision*, pages 709–727. Springer.
- Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Fei-Fei Li, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.
- Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images. <https://www.cs.toronto.edu/kriz/learning-features-2009-TR.pdf>.
- Minh Le, Anh Nguyen, Huy Nguyen, Chau Nguyen, Anh T Tran, and Nhat Ho. 2026. Revisit visual prompt tuning: The expressiveness of prompt experts. In *International Conference on Learning Representations*, volume 2026, pages 107257–107304.
- Yann LeCun, Fu Jie Huang, and Leon Bottou. 2004. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–104. IEEE.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt

- tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.
- Fei-Fei Li, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE.
- Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR.
- Yue Lu, Shizhou Zhang, De Cheng, Yinghui Xing, Nannan Wang, Peng Wang, and Yanning Zhang. 2024. Visual prompt tuning in null space for continual learning. *Advances in neural information processing systems*, 37:7878–7901.
- Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. 2017. dsprites: Disentanglement testing sprites dataset. <https://github.com/deepmind/dsprites-dataset/>.
- Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*.
- Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE.
- Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. *Pytorch: An imperative style, high-performance deep learning library*. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Bo Peng, Yuanwei Hu, Bo Liu, Ling Chen, Jie Lu, and Zhen Fang. 2026. Delving into spectral clustering with vision-language representations. In *The Fourteenth International Conference on Learning Representations*.
- Bo Peng, Jie Lu, Guangquan Zhang, and Zhen Fang. 2025. On the provable importance of gradients for autonomous language-assisted image clustering. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19805–19815. IEEE.
- Aleksandar Petrov, Philip Torr, and Adel Bibi. 2024. When do prompting and prefix-tuning work? a theory of capabilities and limitations. In *International Conference on Learning Representations*, volume 2024, pages 6031–6054.
- 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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30.
- Mark Rudelson and Roman Vershynin. 2007. Sampling from large matrices: An approach through geometric functional analysis. *J. ACM*, 54(4):21–es.
- Kuo Shi, Jie Lu, Shanshan Ye, Guangquan Zhang, and Zhen Fang. 2025. Mirage: Multimodal discriminative representation learning for generalizable ai-generated image detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 353–361.
- Zhengxiang Shi and Aldo Lipani. 2024. Dept: Decomposed prompt tuning for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, volume 2024, pages 51862–51882.
- Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, Lei Hou, Maosong Sun, and Jie Zhou. 2022. On transferability of prompt tuning for natural language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3949–3969,

- Seattle, United States. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. 2018. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pages 210–218. Springer.
- Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. 2022. Spot: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5039–5059.
- Qifan Wang, Yuning Mao, Jingang Wang, Hanchao Yu, Shaoliang Nie, Sinong Wang, Fuli Feng, Lifu Huang, Xiaojun Quan, Zenglin Xu, and Dongfang Liu. 2023. [APrompt: Attention prompt tuning for efficient adaptation of pre-trained language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9147–9160, Singapore. Association for Computational Linguistics.
- Ran Wang, Hua Zuo, Zhen Fang, and Jie Lu. 2024a. Towards robustness prompt tuning with fully test-time adaptation for clip’s zero-shot generalization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8604–8612.
- Yuzhu Wang, Lechao Cheng, Chaowei Fang, Dingwen Zhang, Manni Duan, and Meng Wang. 2024b. Revisiting the power of prompt for visual tuning. In *International Conference on Machine Learning*, pages 50233–50247. PMLR.
- Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE.
- Runjia Zeng, Cheng Han, Qifan Wang, Chunshu Wu, Tong Geng, Lifu Huang, Ying N Wu, and Dongfang Liu. 2024. Visual fourier prompt tuning. *Advances in Neural Information Processing Systems*, 37:5552–5585.
- Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, André Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. 2019. [The visual task adaptation benchmark](#). *CoRR*, abs/1910.04867.
- Ningyuan Zhang, Jie Lu, Keqiuyin Li, Zhen Fang, and Guangquan Zhang. 2025. Release the powers of prompt tuning: cross-modality prompt transfer. In *International Conference on Learning Representations*, volume 2025, pages 49285–49312.

Appendix

A Datasets

To obtain VTAB-1K, we use the script provided by SPT (He et al., 2023) to download all datasets as PNG images. The information for each dataset is listed in Table 4.

Group	Dataset	Classes	Train	Val	Test
Natural	CIFAR-100	100			10,000
	Caltech101	102			6,084
	DTD	47			1,880
	Flowers102	102	800	200	6,149
	Pets	37			3,669
	SVHN	10			26,032
	SUN397	397			21,750
Specialized	Patch Camelyon	2			32,768
	EuroSAT	10	800	200	5,400
	Resisc45	45			6,300
	Diabetic Retinopathy	5			42,670
Structured	DmLab	6			22,735
	Kitti	4			711
	SmallNORB_azi	18			12,150
	SmallNORB_ele	9	800	200	12,150
	dSprites_loc	16			73,728
	dSprites_ori	16			73,728
	Clevr_dist	6			15,000
	Clevr_count	8			15,000

Table 4: **Information on the VTAB-1K datasets.** During hyperparameter search, the model is trained on the 800 training samples and evaluated on the 200 validation samples. During final training, the model is trained on all 1000 training and validation samples and evaluated on the test set.

For image preprocessing on VTAB-1K, we follow the default settings and do not use any data augmentation, except for the following standard preprocessing steps:

1. Resizing images to 224×224 ;
2. Converting images to PyTorch tensors and rescaling pixel values to $[0, 1]$;
3. Normalizing images using the ImageNet mean and standard deviation values: mean = (0.485, 0.456, 0.406) and std = (0.229, 0.224, 0.225).

B Source Prompt Pretraining

The source prompt is pretrained on SNLI using either a pretrained RoBERTa-base or GPT2-base language model.

Input Embedding Formats. Each SNLI sample contains two sentences. For prompt tuning with RoBERTa-base, we use the following input embedding structure:

[Prompt, MASK, CLS, S1, SEP, S2, SEP]

For prompt tuning with GPT2-base, we use the following input embedding structure:

[Prompt, S1, S2]

where MASK, CLS, and SEP denote the special tokens of the RoBERTa tokenizer, and S1 and S2 denote the two sentences in an SNLI sample. Positional embeddings are added to all input embeddings except the prompt embeddings.

Output Feature Pooling. For RoBERTa, we take the last-layer hidden-state vector corresponding to the MASK token and pass it to the pretrained language modeling head to obtain the logits. For GPT2, we take the final hidden-state vector from the last layer and pass it to the pretrained language modeling head to obtain the logits.

Hyperparameters. The hyperparameters for pre-training the source prompt on SNLI are listed in Table 5.

Epochs	LR	LR Decay?	LR Warm-Up	WD	Batch Size
64	0.001	No	0	0	16

Table 5: Hyperparameters for prompt pretraining.

C Target Prompt Transfer

Input Embedding Formats. In our experiments, we use a uniform prompt length of 20 and adopt the following input embedding format:

[CLS, Prompt, Image]

Similarly, positional embeddings are added to all input embeddings except the prompt embeddings.

Output Feature Pooling. For ViT-supervised and ViT-MoCo, we use [CLS] pooling: the last-layer hidden-state vector corresponding to the CLS token is passed to the classification head. For ViT-MAE, we average all last-layer hidden-state vectors except the one corresponding to the CLS token, and pass the averaged vector to the classification head.

D Technical Details

D.1 Figure 1

Details of the prompts shown in Figure 1 are listed in Table 6. SNLI Acc. denotes the source accuracy of each prompt. Stb. Rank denotes the stable rank of each prompt, aggregated across all layers. For each prompt, the first row reports the mean target

accuracy, and the second row reports the standard deviation.

D.2 Statistical Significance of Table 1

We evaluate whether the advantage of higher stable rank is statistically reliable at the seed level. For each target dataset, we compare prompts that share the same singular vectors and norm but differ in their singular-value profile: E1 vs. E1+Rank(E64) and E64 vs. E64+Rank(E1) under both norm settings. Since each comparison is repeated across 8 seeds, this yields $2 \times 2 \times 8 = 32$ paired comparisons per dataset and 192 paired comparisons in total. As shown in Table 1, prompts equipped with the higher-rank singular-value profile of E64 win 144 out of 192 seed-level comparisons. Under the null hypothesis that higher-rank and lower-rank prompts are equally likely to win, a one-sided binomial sign test gives $p = 1.22 \times 10^{-12}$. This provides strong evidence that the improvement is not due to random seed variation. The trend is also consistent across individual datasets, with the higher-rank profile winning 24/32 comparisons on CIFAR, 23/32 on DTD, 25/32 on Resisc45, 26/32 on EuroSAT, 20/32 on Kitti, and 26/32 on DmLab. These results support our claim that stable rank is an independent geometric factor associated with stronger cross-modality prompt transferability.

D.3 Statistical Significance of Table 2

We assess whether the advantage of larger CoM is statistically reliable at the seed level. For each dataset, we compare two pairs of prompts whose main difference lies in spatial direction: V(E1) vs. V(E64) and V(TOP20) vs. V(BOT20). Each comparison is repeated across 8 seeds, yielding 48 paired comparisons for each prompt pair and 96 paired comparisons in total. As shown in Table 2, the higher-CoM prompt wins 29 out of 48 comparisons in V(E1) vs. V(E64) and 36 out of 48 comparisons in V(TOP20) vs. V(BOT20). Aggregating both comparison sets, the higher-CoM prompt wins 65 out of 96 seed-level comparisons. Under the null hypothesis that higher-CoM and lower-CoM prompts are equally likely to win, a one-sided binomial sign test gives $p = 3.37 \times 10^{-4}$. This result indicates that the benefit of larger CoM is statistically significant and not merely due to random seed variation. Notably, the stronger significance in V(TOP20) vs. V(BOT20) (36/48 wins, $p = 3.59 \times 10^{-4}$) suggests that the effect becomes clearer when the spatial-direction gap is larger.

These results support our claim that prompts biased farther away from the dominant target-feature PCs tend to provide more effective cross-modality initialization.

D.4 Per-Task Results of Table 3

The per-task hyperparameters of GPI are listed in Table 7. The number of prompt vectors per layer is universally set as 20 across all scenarios. We do not use any prompt dropouts and set the value of drop path rate as 0 across all scenarios. The per-task target results of GPI and CMPT are listed in Table 8.

Statistical Significance. To assess whether the improvements of GPI over CMPT-Deep are consistent beyond average accuracy, we conduct a seed-level paired comparison across all VTAB-1K tasks and backbone settings. Each task is evaluated with three random seeds, yielding $19 \times 3 \times 3 = 171$ paired comparisons in total. As shown by the superscripts in Table 8, GPI outperforms CMPT-Deep in 121 out of 171 seed-level comparisons. A one-sided binomial sign test under the null hypothesis that GPI and CMPT-Deep are equally likely to win gives $p = 2.81 \times 10^{-8}$, indicating that the advantage of GPI is statistically significant. The trend is also consistent within each backbone: GPI wins 43/57 comparisons on ViT-Supervised ($p = 7.69 \times 10^{-5}$), 42/57 on ViT-MAE ($p = 2.3 \times 10^{-4}$), and 36/57 on ViT-MoCo ($p = 0.031$). These results confirm that the effectiveness of GPI is not driven by isolated datasets or random-seed variation, but reflects a robust advantage of geometry-based initialization over language-pretrained prompt transfer.

E Theoretical Analysis

In this section, we provide a theoretical interpretation of why prompt stable rank is associated with cross-modality transferability. Our analysis builds on the formulation of Petrov et al. (2024), which shows that prepended prompt vectors primarily influence an attention layer by introducing an additive bias term to the attention output. We adapt this formulation to CMPT and show that the expressive capacity of this bias is bounded by the rank of the pretrained prompt. For simplicity, we analyze a single-head attention layer and omit the layer index. The same argument applies independently to each layer in deep prompt transfer.

Model	Prompt	SNLI Acc.	CIFAR	DTD	Resisc45	EuroSAT	Kitti	DmLab
RoBERTa	E1	86.89	81.24 0.40	69.60 0.52	84.37 0.48	96.40 0.16	79.37 0.59	46.63 0.37
	E64	90.57	81.75 0.40	70.13 0.37	85.20 0.41	96.57 0.15	79.79 1.09	47.82 0.73
	E100	83.97	81.36 0.50	69.75 0.55	84.49 0.45	96.49 0.13	78.53 0.24	46.95 0.56
	E6400	83.33	81.81 0.36	70.25 0.51	85.12 0.32	96.68 0.14	79.61 0.40	47.95 0.78
GPT2	E1	56.89	81.76 0.11	69.09 0.54	82.73 1.32	96.26 0.15	78.34 0.20	45.64 0.59
	E64	85.54	81.93 0.06	70.25 0.03	85.12 0.51	96.40 0.07	79.65 1.06	47.37 0.75
	E100	59.11	81.70 0.69	68.99 0.16	83.61 0.35	96.35 0.09	77.97 0.84	46.67 0.40
	E6400	67.86	82.29 0.02	70.30 0.27	85.34 0.50	96.58 0.07	79.42 0.54	47.68 0.68

Table 6: Details of Figure 1.

E.1 Prompt-Induced Bias in Attention

Let the original input embeddings be $X = [x^1, x^2, \dots, x^{N_I}] \in \mathbb{R}^{N_I \times d_h}$, where N_I denotes the number of input tokens and d_h is the hidden dimension. For a single-head attention layer with query, key, and value projection matrices W_Q , W_K , and W_V , the attention weight from token x^i to token x^r is

$$A(x^i, x^r) = \frac{e^{(x^i W_Q)(x^r W_K)^\top}}{\sum_{m=1}^{N_I} e^{(x^i W_Q)(x^m W_K)^\top}} \quad (9)$$

The corresponding output for the i -th token is

$$o_i = \sum_{r=1}^{N_I} A(x^i, x^r) x^r W_V \quad (10)$$

After prepending a prompt matrix $p = [p^1, p^2, \dots, p^{N_p}] \in \mathbb{R}^{N_p \times d_h}$, the attention layer operates on $[p, X]$. We focus on the outputs corresponding to the original input tokens, since prompt-token outputs are discarded in deep prompt tuning. The output for token x^i becomes

$$o_i^p = \sum_{r=1}^{N_I} A_p(x^i, x^r) x^r W_V + \sum_{j=1}^{N_p} A_p(x^i, p^j) p^j W_V \quad (11)$$

where $A_p(\cdot, \cdot)$ denotes attention weights normalized over both prompt and input tokens. Equation (11) can be written as

$$o_i^p = \tilde{o}_i + \beta_p^i \quad (12)$$

where $\tilde{o}_i$ denotes the reweighted contribution from the original input tokens, and

$$\beta_p^i = \sum_{j=1}^{N_p} A_p(x^i, p^j) p^j W_V \quad (13)$$

is the additive bias induced by the prompt vectors. Thus, the prompt modifies the attention output through directions lying in the span of the projected prompt vectors $\{p^j W_V\}_{j=1}^{N_p}$.

E.2 Implication for CMPT

In CMPT, the source prompt p_s is not directly inserted into the target model. Instead, it is first mapped into the target hidden space by a learnable linear projector with weight matrix W_P . The prompt-induced bias in Equation (13) therefore becomes

$$\beta_{\text{CMPT}}^i = \sum_{j=1}^{N_p} A_p(x^i, p_s^j W_P) p_s^j W_P W_V \quad (14)$$

Equivalently, the bias vectors produced across all input tokens lie in the row space of

$$p_s W_P W_V \quad (15)$$

Task	ViT Supervised					ViT MAE					ViT MoCo v3				
	BS	O	LR	WD	α	BS	O	LR	WD	α	BS	O	LR	WD	α
Caltech101	64	Adam	0.005	0.001	0.5	64	AdamW	0.05	0.1	1	64	Adam	0.01	0.0001	1
CIFAR	64	Adam	0.005	0.001	0.75	64	AdamW	0.05	0.2	1	64	Adam	0.01	0.0001	1
DTD	64	Adam	0.01	0.001	1	64	AdamW	0.001	0	1	64	Adam	0.02	0.0001	1
Flowers102	64	Adam	0.005	0.001	1	64	AdamW	0.0005	0.0001	0.5	64	Adam	0.05	0	1
Pets	64	Adam	0.01	0.001	1	64	AdamW	0.0005	0.1	1	64	Adam	0.05	0.0002	1
SUN397	64	Adam	0.01	0.001	1	64	AdamW	0.05	0.1	0.25	64	Adam	0.01	0.0001	1
SVHN	4	Adam	0.5	0	0.5	64	AdamW	0.0001	0	0.25	64	Adam	0.005	0.0001	1
Patch Camelyon	64	Adam	0.001	0.0001	0.75	64	AdamW	0.0005	0	0.25	64	Adam	0.001	0.001	1
Resisc45	64	Adam	0.01	0.001	1.5	64	AdamW	0.0005	0.001	0.25	64	Adam	0.05	0.0001	1.5
EuroSAT	64	Adam	0.01	0.001	0.75	64	AdamW	0.0005	0.0001	0.25	64	Adam	0.05	0.0001	1
Diabetic Retinopathy	64	Adam	0.0005	0	0.5	64	AdamW	0.0001	0.1	1	64	Adam	0.005	0	0.25
DmLab	64	Adam	0.001	0.01	0.5	64	AdamW	0.01	0.01	1	64	Adam	0.005	0.001	1
Kitti	64	Adam	0.0005	0.1	1	64	AdamW	0.0005	0.1	1	64	Adam	0.001	0.01	1
SmallNORB_azi	64	Adam	0.0005	0.0001	0.25	64	AdamW	0.00005	0.01	0.25	64	Adam	0.01	0	0.25
SmallNORB_ele	64	Adam	0.0002	0.1	0.75	64	AdamW	0.00005	0.1	0.25	64	Adam	0.001	0	1
dSprites_loc	64	Adam	0.001	0	1	16	AdamW	0.00005	0.01	1	64	Adam	0.001	0.0001	1
dSprites_ori	64	Adam	0.0005	0.0001	0.5	16	AdamW	0.00005	0.01	0.5	64	Adam	0.05	0	1
Clevr_dist	64	Adam	0.0005	0.1	0.1	64	AdamW	0.0001	0.1	0.1	64	Adam	0.001	0	1.5
Clevr_count	64	Adam	0.0005	0.0001	0.1	64	AdamW	0.0001	0.001	1	64	Adam	0.001	0.0001	1

Table 7: Hyperparameters of GPI. BS, O, LR, and WD stand for batch size, optimizer, learning rate, and weight decay values, respectively. α denotes the prompt norm scaler.

Therefore, the dimensionality of the additive bias subspace is bounded by the standard *rank product inequality*:

$$\begin{aligned} r(p_s W_P W_V) &\leq \\ \min(r(p_s), r(W_P), r(W_V)) \end{aligned} \quad (16)$$

where $r(\cdot)$ denotes matrix rank.

This bound provides a theoretical explanation for the empirical importance of prompt stable rank. The value matrix W_V is fixed because the pre-trained target Transformer remains frozen during transfer. The projector W_P is learnable, but it can only transform the subspace already spanned by the source prompt p_s . Consequently, if p_s has limited rank or is dominated by only a few singular directions, the resulting attention bias is restricted to a low-dimensional subspace, limiting the range of output modifications available to the prompt. By contrast, a prompt matrix with greater spectral spread provides a richer set of directions before projection, increasing the effective dimensionality of the bias subspace that CMPT can exploit.

Although Equation (16) is stated in terms of algebraic rank, our empirical stable-rank measure serves as a continuous proxy for this usable subspace dimensionality. A prompt with low stable rank may be formally full-rank but still behave like a low-dimensional prompt because most of its spectral mass is concentrated in a small number of directions. This explains why prompts with higher stable rank tend to yield stronger transfer

performance in Table 1: they allow the transferred prompt to bias the target attention layer along a more diverse set of directions.

F Discussions

What Shapes the Prompt Geometry? To better understand what shapes prompt geometry during language-based prompt pretraining, we conduct a case study by pretraining prompts under 5 different source-side policies detailed in Table 9.

Policy	Data	Backbone	Embedding
100%–SNLI	Full SNLI	Pretrained	Pretrained
Rand. Data	Random	Pretrained	Pretrained
Rand. Data & Model	Random	Randomized	Randomized
Rand. Data & Backbone	Random	Randomized	Pretrained
Rand. Data & Embedding	Random	Pretrained	Randomized

Table 9: Source-side pretraining policies used to analyze what shapes prompt geometry. Under each policy, the prompt is pretrained with RoBERTa for 64 epochs. “Backbone” and “Embedding” denote the corresponding modules of RoBERTa.

Figure 4 shows the geometric dynamics of prompts under these pretraining policies. Two observations can be made: First, the increase in stable rank depends on both real language data and the pretrained language model. When the data is randomized, stable rank can only reach below 4 rather than nearly 7 as in **100%–SNLI**. When the model is fully randomized, or when only the backbone is randomized, stable rank further decreases and remains around 1.5. This suggests that both real

		Natural								Specialized					Structured								
		Cattech101	CIFAR100	DTD	Flowers 102	Pets	Sun397	SVHN	AVG.	Patch Camelyon	Resisc45	EuroSAT	Retinopathy	AVG.	DMLab	KITTI	SmallNORB/azi	SmallNORB/ele	dSprites/loc	dSprites/ori	Clevr/dist	Clevr/count	AVG.
SUP	Full	87.7	68.9	64.3	97.2	86.9	38.8	87.4	75.88	79.7	84.2	95.7	73.9	83.36	41.7	65.5	25.7	29.1	57.5	46.7	58.6	56.3	47.64
	CMPT-Deep	91.1 ^{0.15}	81.8 ^{0.40}	70.1 ^{0.37}	98.6 ^{0.08}	90.1 ^{0.13}	52.9 ^{0.84}	85.6 ^{0.24}	81.46	83.9 ^{0.54}	85.2 ^{0.41}	96.6 ^{0.19}	75.8 ^{0.26}	85.37	47.8 ^{0.89}	79.8 ^{1.33}	31.1 ^{0.21}	43.7 ^{1.40}	89.1 ^{1.39}	53.8 ^{1.47}	62.2 ^{0.44}	77.5 ^{0.49}	60.63
	GPI	91.6 ³ _{0.28}	82.7 ³ _{0.17}	71.0 ³ _{0.24}	98.7 ² _{0.07}	90.0 ² _{0.13}	54.0 ³ _{0.29}	85.8 ² _{0.05}	81.96 ¹⁸	84.1 ² _{0.54}	85.5 ³ _{0.24}	96.9 ² _{0.19}	75.6 ¹ _{0.08}	85.53 ⁸	48.4 ² _{0.01}	80.3 ² _{0.14}	32.0 ³ _{0.81}	44.6 ² _{0.96}	89.6 ³ _{1.44}	55.0 ² _{0.52}	62.4 ² _{0.27}	77.3 ¹ _{0.09}	61.19 ¹⁷
MAE	Full	84.2	24.6	56.9	72.7	74.4	15.8	86.6	59.31	81.8	72.3	94.0	70.6	79.68	45.2	75.3	30.2	33.0	72.5	47.5	59.8	67.0	53.82
	CMPT-Deep	80.9 ^{6.94}	32.2 ^{0.94}	59.0 ^{1.53}	73.3 ^{0.70}	50.3 ^{19.99}	16.6 ^{1.54}	76.5 ^{1.05}	55.54	86.1 ^{0.91}	71.1 ^{2.10}	93.9 ^{1.23}	74.5 ^{0.26}	81.37	42.9 ^{2.83}	82.0 ^{1.07}	20.4 ^{1.33}	23.1 ^{6.18}	83.9 ^{3.30}	43.4 ^{1.58}	61.6 ^{1.78}	76.2 ^{2.11}	54.18
	GPI	85.5 ² _{0.69}	28.1 ⁰ _{4.18}	60.5 ³ _{0.43}	74.3 ³ _{1.03}	60.4 ² _{12.92}	18.8 ³ _{1.16}	77.5 ² _{0.17}	57.89 ¹⁵	86.0 ¹ _{0.22}	73.1 ² _{1.21}	95.3 ³ _{0.26}	74.9 ³ _{0.07}	82.34 ⁹	41.4 ¹ _{1.21}	83.3 ³ _{0.14}	20.6 ² _{1.76}	33.3 ³ _{6.27}	86.9 ² _{0.82}	45.0 ² _{4.55}	62.9 ³ _{0.58}	77.4 ² _{1.51}	56.35 ¹⁸
MoCo	Full	91.0	57.6	64.6	91.6	79.9	29.1	89.8	71.94	85.1	83.1	96.4	74.2	84.70	44.6	77.9	31.5	36.9	63.8	49.0	56.9	55.2	51.98
	CMPT-Deep	91.8 ^{0.22}	76.4 ^{0.14}	70.8 ^{0.11}	92.6 ^{0.09}	89.5 ^{0.26}	40.6 ^{1.41}	85.7 ^{0.53}	78.18	86.3 ^{0.26}	85.5 ^{0.04}	96.9 ^{0.29}	75.8 ^{0.31}	86.13	49.0 ^{0.27}	82.5 ^{0.21}	28.4 ^{0.77}	44.3 ^{2.61}	85.4 ^{1.60}	52.3 ^{1.23}	63.8 ^{0.32}	79.9 ^{1.53}	60.69
	GPI	92.0 ² _{0.12}	76.6 ² _{0.10}	70.9 ² _{0.22}	92.7 ² _{0.07}	89.5 ¹ _{0.21}	41.3 ² _{0.23}	86.5 ¹ _{0.30}	78.50 ¹²	86.8 ³ _{0.61}	85.6 ² _{0.04}	97.0 ¹ _{0.11}	75.6 ¹ _{0.40}	86.25 ⁷	48.9 ¹ _{0.57}	83.1 ³ _{0.20}	29.5 ³ _{1.31}	45.6 ² _{1.06}	87.4 ² _{2.47}	54.0 ³ _{0.52}	63.3 ⁰ _{0.03}	81.3 ³ _{0.64}	61.65 ¹⁷

Table 8: **Detailed results on VTAB-1K across three ViT-B/16 backbones.** All results are averaged over three random seeds (42, 44, and 100); the smaller superscripts below each score denote the standard deviation across seeds. Superscripts on GPI scores denote the number of seed-level wins over CMPT-Deep.

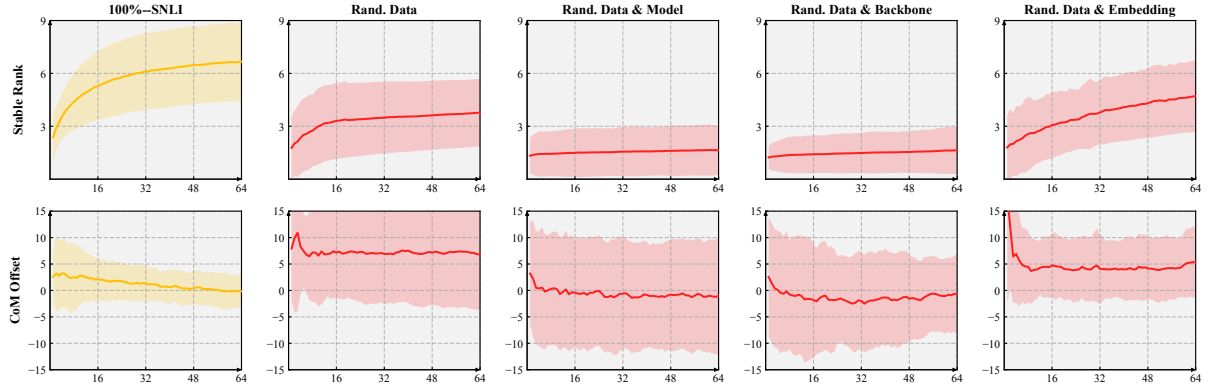


Figure 4: Geometric dynamics of prompts under different source pretraining policies. We track prompt stable rank and CoM offset across source pretraining epochs. The CoM offset is computed as $\text{CoM} - \frac{d_h - 1}{2}$ using CIFAR as the target task. A positive offset indicates that prompt vectors are biased toward directions farther from the leading target-feature PCs, whereas a negative offset indicates stronger alignment with the leading PCs. The shaded region denotes the standard deviation across layers.

text data and the pretrained language backbone contribute to shaping the spectral diversity of prompts.

Second, the evolution of spatial direction is less straightforward. We observe a biased CoM offset at epoch 64 for **Rand. Data** and **Rand. Data & Embedding**, but a near-neutral CoM for **Rand. Data & Model** and **Rand. Data & Backbone**, where the model backbone is randomized. This pattern suggests that the pretrained backbone may also influence directional drift, but the evidence is less conclusive than for stable rank. We therefore avoid attributing the evolution of spatial direction to a specific source-side factor.

Finally, we transfer the **E64** prompts obtained under these policies and report the results in Table 10. The results further support our main ar-

Setting	CIFAR
100%-SNLI	81.75
Rand. Data	82.15
Rand. Data & Model	81.28
Rand. Data & Backbone	81.46
Rand. Data & Embedding	82.05

Table 10: Transfer performance of **E64** prompts pretrained under different source-side policies. All prompts are transferred to ViT-Supervised.

gument. Prompts pretrained on random data with a pretrained RoBERTa backbone achieve performance slightly higher than prompts pretrained on full SNLI. This indicates that task-specific linguis-

tic supervision is not necessary for producing transferable prompts. However, when the pretrained backbone is randomized, performance decreases, as shown by both **Rand. Data & Model** and **Rand. Data & Backbone**. This is consistent with the geometric dynamics in Figure 4, where the stable rank remains below 3 when the backbone is randomized. Together, these results suggest that *linguistic resources, including real text data and pretrained language models, can help shape prompt geometry, particularly stable rank, but they are neither an efficient nor a necessary route to transferable prompts.*

Initialization Transfer without a Projector. To test whether GPI’s advantage depends on projection transfer, we remove the source-to-target projector and directly initialize the learnable target prompt with either RoBERTa-E64 or GPI. We compare this direct-initialization setting with the original projection-transfer pipeline on five VTAB-1K tasks using ViT-Supervised. The results are reported in Table 11.

Under direct initialization, E64 and GPI obtain five-task averages of 83.59 and 83.62, respectively; GPI improves by only +0.03 points and wins 10 of 15 seed-level comparisons. Under projection transfer, the corresponding averages are 82.70 and 83.31, giving GPI a +0.61-point gain and 13 of 15 wins. Thus, GPI remains a competitive direct target-prompt initialization, but its advantage is substantially stronger within the conventional projection-transfer pipeline.

Prompt-Length Ablation. We examine whether the directional prior remains effective as prompt length changes on CIFAR with ViT-Supervised. The results are reported in Table 12. We compare prompts centered on the leading target-feature direction ($c = 0$) with prompts centered away from it ($c = d_h - 1$) for $N_p \in \{10, 20, 50\}$. The directional margin is $\Delta_{\text{dir}} = \text{Acc}(c = d_h - 1) - \text{Acc}(c = 0)$.

For $N_p \in \{10, 20\}$, $c = d_h - 1$ has a positive mean margin over $c = 0$ for every tested σ , ranging from +0.07 to +0.43 points. For $N_p = 50$, the margins become negligible (+0.05, +0.05, and -0.07). One possible explanation is that direction matters more when shorter prompts span a limited initialized subspace, whereas longer prompts can cover more directions and become less sensitive to the coarse directional choice.

Efficiency Analysis. GPI provides a substantial efficiency advantage over language-based prompt pretraining. In our implementation, pretraining a source prompt on SNLI takes approximately 1700 GPU seconds per epoch, measured on an NVIDIA RTX A5500 24GB GPU. With the standard 64-epoch pretraining schedule, this amounts to about 108,800 GPU seconds, or roughly 30.2 GPU hours, for obtaining a language-pretrained prompt. By contrast, GPI only requires extracting target features, performing SVD to obtain the target-feature basis, and applying the Gaussian-envelope construction, which takes approximately 12 GPU seconds in total. This corresponds to more than a $9,000\times$ reduction in initialization cost. Therefore, GPI not only supports our geometric interpretation of CMPT, but also provides a much cheaper way to transferable prompt initialization.

Geometry as the Interface. The contrast between source linguistic performance and prompt geometry provides an insightful perspective on why CMPT works. When language and vision are explicitly aligned in a shared representation space, linguistic semantics can benefit visual tasks. Recent studies on cross-modally aligned vision-language models (Peng et al., 2026, 2025) demonstrate that textual semantics can be leveraged alongside visual representations to improve downstream image clustering. CMPT, however, differs fundamentally from this setting: source prompts are learned with an independently pretrained language model and then mapped to an independently pretrained vision model, without an explicitly aligned semantic space. Consequently, the preservation of source-side linguistic semantics after cross-modality projection cannot be assumed. By contrast, geometric properties such as spectral spread and spatial orientation are defined directly in representation space and are less tied to a specific source task or modality. From this perspective, CMPT can be better understood as transferring an initialization geometry that determines how flexibly the prompt can bias the target model, rather than necessarily transferring linguistic semantics themselves. This interpretation helps explain why language-pretrained prompts can benefit visual adaptation even when their source linguistic performance is not predictive of target performance.

Implications for Prompt Initialization. The success of GPI further suggests that source prompt pretraining is not the only way to obtain transfer-

Pipeline	Task	E64	GPI	Δ	GPI Wins
Projection	CIFAR	81.80 ± 0.40	82.73 ± 0.17	+0.93	3/3
	DTD	70.10 ± 0.37	71.03 ± 0.24	+0.93	3/3
	RESISC45	85.20 ± 0.41	85.53 ± 0.24	+0.33	3/3
	EuroSAT	96.60 ± 0.19	96.92 ± 0.19	+0.32	2/3
	KITTI	79.80 ± 1.33	80.32 ± 0.14	+0.52	2/3
Direct initialization	CIFAR	83.29 ± 0.11	83.50 ± 0.31	+0.21	2/3
	DTD	71.76 ± 0.35	71.74 ± 0.13	-0.02	2/3
	RESISC45	86.23 ± 0.32	86.16 ± 0.21	-0.07	2/3
	EuroSAT	96.91 ± 0.13	96.80 ± 0.31	-0.11	2/3
	KITTI	79.75 ± 1.29	79.89 ± 0.37	+0.14	2/3

Table 11: **Projection transfer versus direct target-prompt initialization.** Both settings use 20 prompt vectors per layer. Δ denotes GPI minus E64, and GPI Wins reports seed-level wins.

N_p	σ	$c = 0$	$c = d_h - 1$	Δ_{dir}
10	10	82.92	83.20	+0.28
	384	82.83	83.26	+0.43
	768	83.26	83.36	+0.10
20	20	82.37	82.67	+0.30
	384	82.57	82.68	+0.11
	768	82.58	82.65	+0.07
50	50	82.83	82.88	+0.05
	384	82.98	83.03	+0.05
	768	82.93	82.86	-0.07

Table 12: **Prompt-length ablation of GPI’s directional prior** on CIFAR with ViT-Supervised.

able prompts. If the main transferable properties are geometric, then they can be constructed directly rather than learned through language pretraining. This shifts the role of source pretraining from a necessary semantic learning stage to an expensive procedure that happens to produce useful prompt geometry. Such a perspective opens a simpler design space for prompt initialization: instead of searching for better source tasks or larger source models, one can directly initialize prompts with favorable rank and target-aware spatial orientation.

G AI Assistance Declaration

The authors used AI-assisted tools to improve the clarity, grammar, and readability of the manuscript. All core ideas, experimental designs, analyses, results, and conclusions were developed and verified by the authors. The authors take full responsibility for the content of the paper.

H Artifact Use and Licenses

We use all external artifacts in accordance with their intended research and evaluation purposes. SNLI is accessed through Hugging Face and is licensed under CC BY-SA 4.0. GPT-2 and RoBERTa are also accessed through Hugging Face and are released under MIT licenses. VTAB-1K is obtained through the Google Research task_adaptation repository, which is released under Apache-2.0. The ViT backbones are obtained from timm; the timm codebase is licensed under Apache-2.0, while pretrained weights may be subject to the terms of their original training datasets. Our use of these artifacts is limited to research experimentation, prompt pretraining, feature extraction, and downstream evaluation, which is consistent with their intended research use.

I Artifact Documentation

We use SNLI to pretrain source prompts, VTAB-1K for target evaluation, and GPT-2, RoBERTa, and ViT as pretrained model artifacts. SNLI is an English natural language inference dataset consisting of sentence pairs labeled as entailment, contradiction, or neutral, and is used only to pretrain source prompts in the language modality. VTAB-1K is a visual adaptation benchmark covering natural, specialized, and structured visual tasks; we use it as a controlled non-language testbed for evaluating cross-modality prompt transfer under data-scarce conditions. GPT-2 and RoBERTa are English pretrained language models used for source prompt pretraining, while ViT backbones from timm are used as target visual models.