Beyond Adapter Retrieval: Latent Geometry-Preserving Composition via Sparse Task Projection

Pengfei Jin[1*], Peng Shu[2*], Sifan Song[1], Sekeun Kim[1], Qing Xiao[1], Cheng Chen[3,4], Tianming Liu[2], Xiang Li[1], Quanzheng Li[1†]

[1]Center of Advanced Medical Computing and Analysis, Massachusetts General Hospital and Harvard Medical School
[2]School of Computing, The University of Georgia
[3]Department of Electrical and Electronic Engineering, The University of Hong Kong
[4]School of Biomedical Engineering, The University of Hong Kong

Abstract

Recent advances in parameter-efficient transfer learning have demonstrated the utility of composing LoRA adapters from libraries of pretrained modules. However, most existing approaches rely on simple retrieval heuristics or uniform averaging, which overlook the latent structure of task relationships in representation space. We propose a new framework for adapter reuse that moves beyond retrieval, formulating adapter composition as a geometry-aware sparse reconstruction problem. Specifically, we represent each task by a latent prototype vector derived from the base model's encoder and aim to approximate the target task prototype as a sparse linear combination of retrieved reference prototypes, under an ℓ_1-regularized optimization objective. The resulting combination weights are then used to blend the corresponding LoRA adapters, yielding a composite adapter tailored to the target task. This formulation not only preserves the local geometric structure of the task representation manifold, but also promotes interpretability and efficient reuse by selecting a minimal set of relevant adapters. We demonstrate the effectiveness of our approach across multiple domains—including medical image segmentation, medical report generation and image synthesis. Our results highlight the benefit of coupling retrieval with latent geometry-aware optimization for improved zero-shot generalization.

Code — https://github.com/Jinpf314/Geometry-Aware-Adapter-Composition

Introduction

Foundation models such as CLIP (Radford et al. 2021), LLaMA (Touvron et al. 2023), SAM (Kirillov et al. 2023), and Stable Diffusion (Rombach et al. 2022) have demonstrated impressive generalization across modalities and tasks. These models are pretrained on massive datasets and can serve as adaptable backbones for downstream applications in vision, language, and healthcare (Shu et al. 2024; Zhao et al. 2024a; Rezayi et al. 2024; Yang et al. 2024). However, adapting them to specific tasks remains costly in terms of data, compute, and retraining cycles.

Low-Rank Adaptation (LoRA) (Hu et al. 2021) provides a scalable solution by fine-tuning a small set of task-specific parameters while freezing the base model. LoRA adapters can be cheaply trained, shared, and reused, enabling modular and community-driven adaptation. As open-source ecosystems like CivitAI and HuggingFace grow, large-scale adapter libraries have emerged—with over 100K public LoRAs reported for Stable Diffusion alone (Luo et al. 2025).

On the other hand, Retrieval-Augmented Generation (RAG) (Lewis et al. 2020) enhances model outputs by integrating an external retrieval step, grounding responses in factual data. This method is especially effective in zero-shot learning scenarios, where the model encounters tasks it has not seen during training. This trend raises a natural question: *Can we retrieve and reuse existing LoRA adapters to generalize to new tasks without additional training?*

Recent works have explored this idea from two angles. One line of research focuses on supervised or few-shot adapter fusion(Huang et al. 2023; Prabhakar et al. 2024), learning composition weights through gradient-based tuning or black-box optimization. While effective, these methods rely on task-specific labeled data, limiting their use in zero-shot settings. Another line centers on retrieval and reranking strategies(Zhao et al. 2024b; Ostapenko et al. 2024), selecting relevant adapters based on embedding similarity. However, these methods typically use uniform or heuristic similarity-based fusion, which fails to model the geometric structure of task relationships in latent space.

In this work, we shift the focus from retrieval to composition. We propose a latent geometry-aware framework that formulates adapter fusion as a sparse projection problem. Specifically, we represent each task by a prototype vector in representation space, and approximate the target prototype as a sparse linear combination of retrieved reference prototypes. The resulting weights—solved via ℓ_1-regularized optimization—are used to combine the corresponding LoRA adapters in parameter space. This formulation captures the structure of the task manifold, promotes interpretable reuse, and requires no supervision.

Figure 1 illustrates our pipeline. We maintain a vectorized adapter library (Adapter-VecDB), which stores pretrained LoRA modules alongside their latent task representations.

*These authors contributed equally.
†Corresponding author. Li.Quanzheng@mgh.harvard.edu

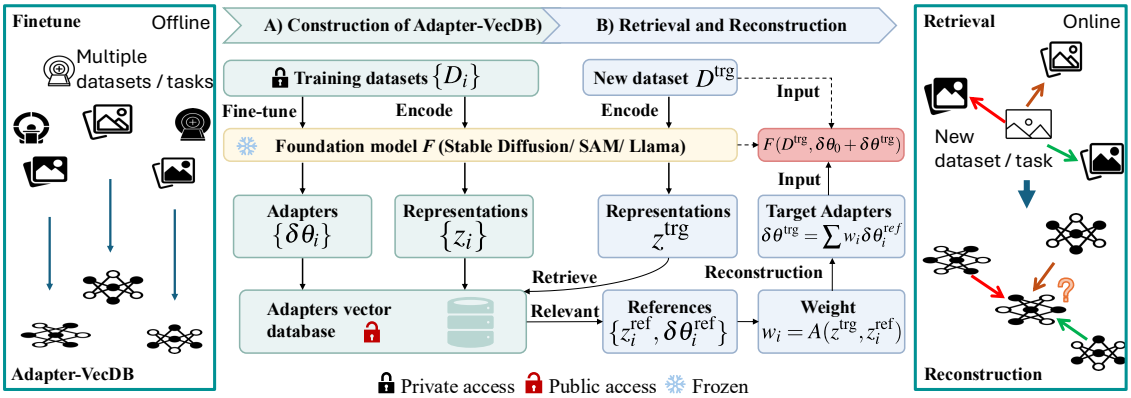

Figure 1: Overview of our latent geometry-aware adapter composition framework. Each task is represented by a prototype vector and stored in a vectorized adapter library (Adapter-VecDB), alongside its corresponding LoRA adapter. At inference time, we extract the target task prototype, retrieve the most relevant reference prototypes and adapters, and solve a sparse reconstruction problem in latent space. The resulting weights are then used to combine adapter parameters in a geometry-preserving manner.

Given a new task, we extract its prototype, retrieve the k most relevant adapters, and compute sparse combination weights via constrained reconstruction. Compared to averaging or similarity-based fusion, our method leverages the latent geometry of task space to guide composition.

We alidate our approach across three domains: medical image segmentation, radiology report generation, and image synthesis. Our method consistently outperforms baseline strategies under zero-shot evaluation, especially in distribution-shifted and low-resource settings.

Our main contributions are summarized as follows:

- We propose a novel framework for geometry-aware adapter composition, based on sparse projection over task prototypes.
- Our formulation reveals that different fusion methods can be viewed as special cases under different assumptions about prototype geometry, offering a unified lens to understand adapter composition.
- We demonstrate strong zero-shot performance on multiple domains using only pretrained adapters without any additional data or fine-tuning.

Preliminaries and Related Work

Retrieval and Zero-Shot Generalization

Retrieval-augmented generation (RAG) methods enhance large models by injecting external information at inference time, typically through dynamic retrieval of text or structured data (Ma et al. 2023; Peng et al. 2024). These approaches have improved low-resource performance and reduced hallucinations, especially in sensitive domains such as healthcare, where raw data access is limited (Seo et al. 2024; Parvez et al. 2022). However, RAG typically relies on textual corpora and does not support parameter-level reuse. Zero-shot learning, by contrast, aims to generalize to unseen tasks by leveraging semantic similarity or structural priors. Prior works such as DeViSE (Frome et al. 2013), GCN-ZL (Wang, Ye, and Gupta 2018), and DGP-ZL (Kampffmeyer et al. 2019) learn mappings from task descriptions to model parameters via semantic or graph-based embeddings. Unlike these methods, we focus on composing pretrained adapters using latent task prototypes—without additional data or model training.

Parameter Composition and Adapter Retrieval

Modular parameter composition has emerged as a flexible alternative to full-model fine-tuning. Early methods such as AdapterFusion (Pfeiffer et al. 2021) learn supervised fusion weights to combine task-specific adapters, while AdapterSoup (Chronopoulou et al. 2023) performs zero-shot fusion by averaging retrieved adapters selected via domain similarity. Recent retrieval-based approaches aim to improve flexibility and generalization: LoRA-Hub (Huang et al. 2023) combines LoRA modules using gradient-free optimization with few-shot supervision; LoRA-Retriever (Zhao et al. 2024b) employs a dual-encoder model to retrieve and weight LoRA adapters for zero-shot transfer, supporting both parameter fusion and MoE-style output averaging. Ostapenko et al.(Ostapenko et al. 2024) route tokens to adapters via prototype alignment. Stylus(Luo et al. 2025) retrieves adapters using prompt-VLM similarity and decomposes prompts into subtasks for fine-grained control. LoRA Soups (Prabhakar et al. 2024) optimize adapter combination via learnable concatenation with few-shot supervision.

Beyond retrieval-based composition, other works investigate parameter combination through expert gating and training-time routing. MoE-LoRA variants (Muqeeth, Liu, and Raffel 2023; Ma et al. 2024; Fan et al. 2025; Meng et al. 2025) use learned routers to combine adapter outputs dynamically at each layer. AdaMix (Wang et al. 2022) applies stochastic routing and consistency regularization during training, while UP-RLHF (Zhai et al. 2023) and IOP-FL (Jiang et al. 2023) optimize adapter selection through reinforcement or consistency-based objectives. Zoo-Tuning (Shu et al. 2021) adapts pretrained model weights to target tasks using learned combination coefficients, and Model

Soup (Wortsman et al. 2022) explores simple averaging of models or adapters without retraining. LoRA-Ensemble (Halbheer et al. 2024) similarly averages adapters at inference time but focuses on distributional robustness. While these approaches vary in supervision and assumptions, most require either training or access to large-scale evaluation, in contrast to our plug-and-play, data-free composition strategy based on latent task geometry.

Sparse Composition and Geometric Reconstruction

Sparse composition techniques—such as compressed sensing (Donoho 2006), dictionary learning (Mairal et al. 2010), and manifold learning (Roweis and Saul 2000)—reconstruct inputs from a small number of basis elements while preserving local geometry. These ideas have inspired applications in multi-task learning (Argyriou, Evgeniou, and Pontil 2008), subspace clustering (Elhamifar and Vidal 2013), and prototype-based adaptation in deep learning (Zhao, Fu, and He 2023), where task embeddings guide parameter reuse. However, few methods treat adapter composition explicitly as a geometry-aware sparse reconstruction problem. Our work fills this gap by formulating adapter fusion as an ℓ_1-regularized projection over task prototypes, enabling zero-shot reuse without additional supervision.

Method

We propose a three-stage retrieval-and-composition pipeline for zero-shot task adaptation. The key idea is to represent each task as a latent prototype vector in a shared subspace and to reconstruct the target task using a sparse combination of reference tasks. This latent-space reconstruction yields interpretable weights, which are then used to compose the corresponding adapters.

Task Prototype Construction

We begin by representing each task in a shared latent space via a fixed-length prototype vector. This prototype serves as a compact summary of the task's distribution and forms the basis for downstream adapter composition. We organize these prototypes, along with their corresponding adapters, in a centralized database we refer to as the *Adapter-VecDB*, analogous to the retrieval database in RAG systems. This database supports efficient lookup and geometric reasoning over previously trained adapters.

For a given dataset D_i, we compute its task prototype $z_i \in \mathbb{R}^d$ by averaging the feature representations of its constituent samples:

$$z_i = \frac{1}{|D_i|} \sum_{x_j \in D_i} E(x_j, \theta_0), \tag{1}$$

where $E(\cdot, \theta_0)$ denotes a pretrained encoder (e.g., CLIP (Radford et al. 2021), LLaMA (Touvron et al. 2023)) and θ_0 is the frozen backbone model. This mean-pooling strategy ensures that the prototype preserves coarse semantic and statistical structure while remaining efficient to store and compute.

In practice, we adopt the encoder from the foundation model itself to ensure compatibility with the LoRA adapters.

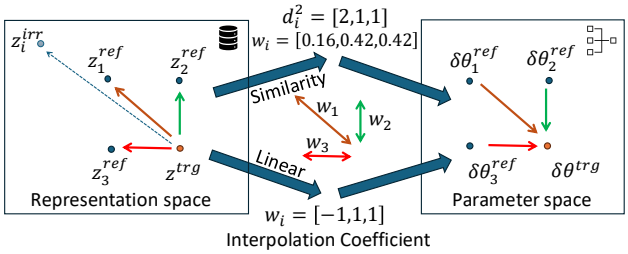

Figure 2: Overview of our adapter composition framework. (1) The input dataset is first encoded into a task prototype z^{trg}. (2) We retrieve a set of reference adapters and their prototypes $\{z_i^{\text{ref}}, \delta\theta_i^{\text{ref}}\}$ from the Adapter-VecDB. (3) Composition weights $\{w_i\}$ are computed in the prototype space using one of different strategies: similarity-based softmax weighting or Least-Squares Linear Combination (4) The final adapter $\delta\theta^{\text{trg}}$ is synthesized as a weighted sum in parameter space and applied to the frozen backbone.

Algorithm 1: Geometry-Aware Adapter Composition

Require: Foundation model $F(\cdot, \theta_0)$; Adapter-VecDB $\{(z_i^{\text{ref}}, \delta\theta_i^{\text{ref}})\}_{i=1}^N$; target dataset D^{trg}
Ensure: Adapted model $F(\cdot, \theta^{\text{trg}})$

1: $z^{\text{trg}} = \frac{1}{|D^{\text{trg}}|} \sum_{x_j \in D^{\text{trg}}} E(x_j, \theta_0)$
▷ Compute prototype for the target task
2: $\{z_i^{\text{ref}}\}_{i=1}^k = \text{top-}k\text{-NN}(z^{\text{trg}}, \{z_j^{\text{ref}}\})$
▷ Retrieve k nearest neighbors in the prototype space
3: $\{w_i\} = \mathcal{A}(\{z_i^{\text{ref}}\}, z^{\text{trg}})$
▷ Compute composition weights (e.g., similarity, LS, sparse)
4: $\delta\theta^{\text{trg}} = \sum_{i=1}^k w_i \delta\theta_i^{\text{ref}}$
▷ Compose the adapter in parameter space
5: $\theta^{\text{trg}} = \theta_0 + \delta\theta^{\text{trg}}$
▷ Output the final adapted model

This approach aligns with the strategy used in the encoder component of the MoE, where feature maps serve a pivotal role in the model architecture. When full access to task datasets is unavailable—such as for public or third-party adapters like those from Stable Diffusion—we instead compute z_i from a small set of descriptive examples, metadata, or natural language summaries associated with the adapter.

For simplicity and practicality in representing dataset features, we initially explored using various distribution distance metrics, such as the Chamfer distance, Nearest Neighbor Distance, Mean Distance, to measure similarities between datasets. However, these metrics did not show significant differences in dataset characteristics.

The collection $\{z_i\}$ across available source tasks defines the *prototype subspace*, which we treat as a geometric basis for reconstructing unseen tasks. This forms the foundation for our geometry-aware adapter composition described next.

Geometry-Aware Sparse Composition

Given a target task with prototype z^{trg}, our goal is to compose a new adapter by selecting and combining relevant pre-

trained adapters from the Adapter-VecDB. We treat this as a geometry-aware reconstruction problem in the task prototype space, where the target prototype is approximated as a linear combination of retrieved source prototypes.

Formally, let $\{z_i^{\text{ref}}\}_{i=1}^k$ be the prototypes of the top-k retrieved adapters from the Adapter-VecDB, and let $\{w_i\}$ denote the corresponding combination weights. We consider three strategies to compute the weights w_i, each reflecting a different assumption about the structure of the latent space.

Similarity-Based Weighting. A common heuristic is to assume that similarity in the latent space correlates with transferability. We compute distances using the squared ℓ_2 norm:

$$d^2(z_i, z^{\text{trg}}) = \|z_i - z^{\text{trg}}\|_2^2, \tag{2}$$

and define weights via a softmax over the negative distances:

$$w_i = \frac{\exp(-\lambda_1 d_i^2)}{\sum_j \exp(-\lambda_1 d_j^2)}, \tag{3}$$

where λ_1 is a temperature parameter controlling the sharpness of the weighting distribution.

Least-Squares Linear Combination. Rather than relying on pairwise similarity, we can explicitly reconstruct the target prototype using a linear combination of the retrieved references. The weights are chosen to minimize reconstruction error under a sum-to-one constraint:

$$\min_{\sum w_i=1} \left\| z^{\text{trg}} - \sum_{i=1}^k w_i z_i^{\text{ref}} \right\|_2^2. \tag{4}$$

This strategy allows the model to interpolate among retrieved prototypes in a globally consistent way, but may still overfit when too many adapters are used.

Sparse Projection (Ours). To promote compact and interpretable compositions, we further impose an ℓ_1 regularization term to encourage sparsity:

$$\min_{\sum w_i=1} \left\| z^{\text{trg}} - \sum_{i=1}^k w_i z_i^{\text{ref}} \right\|_2^2 + \lambda_2 \|w\|_1, \tag{5}$$

where λ_2 controls the trade-off between reconstruction fidelity and sparsity. This sparse formulation allows us to select a minimal subset of reference tasks that best explain the target, aligning with prior work in sparse coding and manifold reconstruction.

Different strategies yields a distinct set of weights $\{w_i\}$ that guide the construction of the final adapter. The **uniform averaging** approach corresponds to prior work such as AdapterSoup (Chronopoulou et al. 2023), where selected adapters are merged using equal weights. The **similarity-based weighting** strategy resembles that of LoRA-Retriever (Zhao et al. 2024b), which assigns weights proportional to prototype similarity; while these works include additional contributions such as retrieval pipelines or routing mechanisms, we isolate and compare their implicit weighting strategies in our framework. In contrast to these heuristics, we propose a novel **sparse projection** method that formulates adapter composition as a constrained optimization problem in the latent space.

Figure 2 illustrates these strategies. While similarity-based methods assign positive weights based on proximity, linear combinations can capture more nuanced structure—including negative contributions—by reconstructing the target from the geometry of the prototype manifold. The experimental section further demonstrates how these differences impact generalization and performance.

Adapter Reassembly

Once the weights $\{w_i\}$ are computed through geometry-aware composition in the prototype space, we apply them to merge the corresponding LoRA adapters in parameter space. Let $\delta\theta_i^{\text{ref}}$ denote the LoRA weight update associated with the i-th retrieved adapter. The composite adapter for the target task is then given by:

$$\delta\theta^{\text{trg}} = \sum_{i=1}^k w_i \delta\theta_i^{\text{ref}}. \tag{6}$$

This aggregated update is added to the frozen base model θ_0, yielding the final adapted model $F(\cdot, \theta_0 + \delta\theta^{\text{trg}})$.

Importantly, this process does not involve any gradient-based optimization or parameter tuning. The target adapter is generated entirely through linear combination of pretrained modules, guided by task geometry in representation space. Compared to training new adapters from scratch, our approach offers significant computational savings and eliminates the need for labeled data.

Moreover, the sparse nature of the weights enhances interpretability: only a small subset of reference adapters contributes to the final composition, allowing us to trace which source tasks influence the target.

The complete pipeline is summarized in Algorithm 1, which outlines the construction of task prototypes, retrieval of candidate adapters, computation of geometry-aware weights, and reassembly of the composite adapter.

Experiments

Experimental Settings and Comparison Methods

We evaluate our method across three domains—medical image segmentation, medical report generation, and image synthesis—using SAM (Kirillov et al. 2023), LLaMA 3.1 8B (Dubey et al. 2024), and Stable Diffusion v1.5 (Rombach et al. 2022). Each task poses a different challenge for adapter composition and zero-shot generalization.

For SAM and LLaMA, where no large-scale LoRA libraries exist, we simulate adapter reuse by training a small set of LoRAs on distinct medical datasets. Each dataset is treated as a task, with leave-one-out evaluation: LoRAs are trained on all but one dataset and composed to adapt to the held-out target. For Stable Diffusion, we follow Stylus (Luo et al. 2025) and use open LoRA library to build an Adapter-VecDB. At test time, we retrieve relevant adapters and compose them using different weighting strategies—without any fine-tuning—to generate images in a zero-shot fashion.

Dataset	Pre-trained	SFT	Zero-shot			
			Avg	Sim	Lin	Ours
Prostate-A (Liu, Dou, and Heng 2020)	-	95.4%	80.3%	87.8%	86.3%	**90.5%**
Prostate-B (Liu, Dou, and Heng 2020)	-	92.8%	77.5%	85.0%	83.4%	**86.0%**
Prostate-C (Liu, Dou, and Heng 2020)	-	90.5%	51.0%	59.8%	61.9%	**64.7%**
Prostate-D (Liu, Dou, and Heng 2020)	-	91.2%	74.9%	82.6%	86.7%	**90.3%**
Prostate-E (Liu, Dou, and Heng 2020)	-	92.7%	64.6%	56.9%	52.0%	**79.1%**
Prostate-F (Liu, Dou, and Heng 2020)	-	93.0%	82.2%	80.8%	82.4%	**90.3%**
AbdAtlas (Li et al. 2024)	-	85.1%	19.2%	22.9%	2.5%	**64.5%**
AutoPet (Gatidis and Kuestner 2022)	-	88.9%	52.7%	83.4%	85.9%	**87.2%**
Abdomen1k (Ma et al. 2021)	-	88.6%	36.6%	82.4%	80.5%	**82.6%**
RAOS (Luo et al. 2024)	-	88.5%	58.6%	80.5%	78.6%	**85.5%**

Table 1: Comparison of DICE scores on the medical image segmentation task across different testing datasets. All methods use the same k-NN retrieval pipeline; only the adapter combination strategy differs. Avg: uniform averaging (Chronopoulou et al. 2023); Sim: softmax similarity-weighted combination (Zhao et al. 2024b); Lin: convex least-squares projection; Ours: ℓ_1-regularized sparse projection (ours). The best performance in each zero-shot setting is highlighted in bold.

To ensure fair comparison, all methods share the same retrieval step (k-NN over prototypes), and only differ in how weights $\{w_i\}$ are computed:

- **Average (Avg):** Uniform averaging of retrieved adapters, following AdapterSoup (Chronopoulou et al. 2023).
- **Similarity-weighted (Sim):** Softmax-weighted combination based on prototype similarity (Eq. (3)), following Lora-Retriever (Zhao et al. 2024b), with $\lambda_1 = 1$.
- **Linear Combination (Lin):** Least-squares projection onto the subspace of retrieved prototypes (Eq. (4)), with convex weight constraint ($\sum w_i = 1$) but no sparsity.
- **Sparse Projection (Ours):** ℓ_1-regularized reconstruction (Eq. (5)) that additionally encourages sparse adapter selection. We use $\lambda_2 = 10$ for vision tasks and $\lambda_2 = 1$ for language tasks, reflecting adapter availability.

All experiments are conducted using 8 NVIDIA H100 80GB GPUs, with adapter training performed under frozen backbones. Additional implementation details, dataset statistics, extended results, and visualizations are provided in the Appendix.

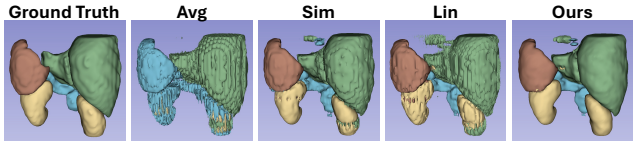

Figure 3: Visualization of segmentation results under different adapter fusion methods. Our method produces the most accurate and anatomically consistent segmentation, reducing both over- and under-segmentation artifacts.

	A	B	C	D	F
Sim	0.06	0.31	0.44	0.08	0.12
Lin	-1.13	0.67	0.26	0.11	1.09
Ours	-0.49	0.47	0.06	-0.03	0.99

Table 2: Weight distribution of our methods applied to Prostate-E Segmentation, with columns representing reference Prostate datasets.

Medical Image Segmentation

To evaluate our method in a high-variance domain, we consider prostate and abdominal CT segmentation using the SAM foundation model (Kirillov et al. 2023). We construct a collection of LoRA adapters, each trained on a distinct dataset from a different scanner, vendor, or population. This setting introduces significant domain shifts across datasets, making it a suitable testbed for zero-shot adaptation. We perform leave-one-out cross-validation: for each target dataset, we compose an adapter using those trained on all other datasets. The segmentation performance is measured using the Dice score (Carass et al. 2020).

Quantitative results. Table 1 reports the segmentation results for all methods. As expected, the pre-trained SAM model (without any adaptation) failed to produce meaningful results, highlighting the need for task-specific tuning. Supervised fine-tuning (SFT) serves as an upper bound but requires labeled data. Among zero-shot methods, our sparse projection approach consistently outperforms alternatives, achieving Dice scores close to SFT and significantly better than Avg or Sim baselines. The improvement is especially pronounced on out-of-distribution datasets. Visualization in Figure 3 further confirm the anatomical accuracy and visual fidelity of our method.

Interpretability of composition. Table 2 shows the weights assigned by different methods when adapting to the Prostate-E dataset. Similarity-based weights concentrate on Prostate-B/C, which are misleading in this case. The unregularized linear projection (Lin) yields unstable weights with large negative coefficients, indicating overfitting to noisy correlations. In contrast, our sparse projection selects a minimal and interpretable subset of relevant adapters, suppress-

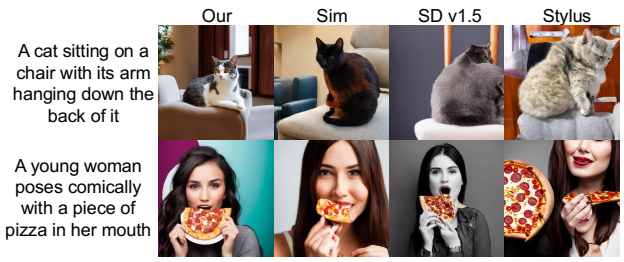

Figure 4: Examples of images generated by our method compared to Euclidean distance similarity, Stable Diffusion v1.5 baseline and Stylus. Our ensemble model demonstrates improved adherence to input captions and produces higher-quality images.

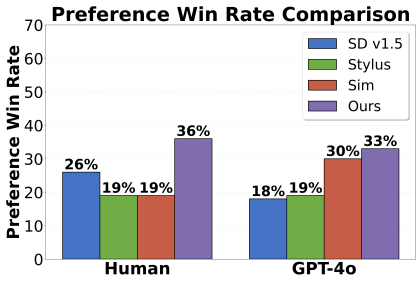

Figure 5: Human and AI evaluations. Our method obtains higher win rate on both tests.

ing unreliable components and improving generalization.

Medical Report Impression

We next evaluate our method in the domain of medical report summarization using Llama 3.1 8B as the foundation model (Dubey et al. 2024). Following Shi et al. (2024), we focus on the impression section of radiology reports across four datasets from Massachusetts General Hospital (MGH): CT-Head, CT-Abdomen, X-ray, and MR. Each dataset defines a task, and we conduct leave-one-out evaluation: for each test set, we compose adapters using the other three. We report ROUGE-L (Lin 2004) and BERTScore (Zhang et al. 2019) for lexical and semantic similarity.

Quantitative results. As shown in Table 3, our sparse projection (Ours) generally outperforms Avg, Sim, and Lin across most metrics and tasks. Notably, it also matches or even surpasses SFT in some cases, particularly when data distribution shift is moderate. The complete experiment results are shown in Appendix. While Lin occasionally attains the best ROUGE-L on a few tasks (e.g., CT-Abdomen and MR), our method delivers more balanced improvements overall. These improvements over Sim and Lin are especially meaningful given that all methods share the same retrieval pool, differing only in weight computation.

Interpretability of composition. To understand the behavior of each strategy, Table 4 shows the adapter weights when generating impressions for CT-Abdomen. Our method assigns most weight to CT-Head (82%) and modest weight to MR (18%), consistent with clinical intuition. This validates the ability of sparse projection to select semantically aligned sources and avoid overfitting.

Image Generation

To assess our method in generative settings, we apply it to Stable Diffusion v1.5 (Rombach et al. 2022), leveraging the large-scale community adapter library from the Civitai platform, which is also employed by Stylus (Luo et al. 2025). Each adapter corresponds to a specific visual concept or style, and each prompt defines a new image generation task. Given a prompt and its paired reference image from the dataset, we retrieve relevant adapters via CLIP-encoded prototype similarity, where the image serves as auxiliary context to improve retrieval quality, and compose them using different strategies.

We compare four methods: the base model (without LoRA), Stylus (Luo et al. 2025) (retrieval + CLIP reranking), similarity-weighted (Sim), and our sparse projection (Ours). For evaluation, we follow Stylus and sample 100 prompts from MS COCO (Lin et al. 2014). Each method generates one image per prompt, and we conduct both GPT-4o and human preference studies. In each evaluation trial, annotators (or GPT) are shown four images side-by-side—one from each method. Then, they are asked to select the best image based on alignment with the prompt as well as overall visual quality, including the absence of disfigured limbs and unrealistic object compositions.

Figure 5 shows that both GPT-4o and human raters consistently prefer our method over Stylus, Sim, and the base model. Qualitative examples in Figure 4 demonstrate that our method produces images that are more semantically consistent with the prompts and exhibit noticeably higher visual quality, with fewer disfigured limbs and unrealistic object compositions. These improvements are achieved even when building upon the pruned SD v1.5 backbone, highlighting the effectiveness of our adapter composition strategy. See Appendix for details on the setup and additional examples.

Ablation Study

We conduct ablation experiments to better understand the behavior and flexibility of our proposed sparse adapter composition framework. Specifically, we examine (1) whether simply using the most similar adapter is sufficient, (2) whether composition can improve supervised LoRA when in-domain adapters are available, and (3) how our method compares in terms of training cost.

Nearest adapter vs. composition. A natural baseline for zero-shot transfer is to retrieve only the most similar adapter (i.e., k=1) and directly apply it to the new task. Table 5 reports the DICE scores when using only the nearest adapter for each test set. While this works reasonably well for some tasks (e.g., Prostate-A, D, F), it fails completely for others (e.g., Prostate-E), likely due to misleading similarity caused by distribution shifts. In contrast, our sparse ensemble method is more robust, as it integrates multiple ref-

Modality	Metric	Pre-trained	SFT	Zero-shot			
				Avg	Sim	Lin	Ours
CT Head	ROUGE-L	0.201	0.2477	0.2124	0.2161	0.2140	**0.2249**
	BERTScore F1	0.8387	0.8499	0.8397	0.8412	0.8405	**0.8433**
CT Abdomen	ROUGE-L	0.1264	0.1387	0.1369	0.1374	**0.1393**	0.1379
	BERTScore F1	0.8039	0.8060	0.8068	0.8071	0.8076	**0.8077**
MR	ROUGE-L	0.1831	0.2153	0.1867	0.1914	**0.1949**	0.1896
	BERTScore F1	0.8365	0.8418	0.8378	0.8369	0.8380	**0.8385**
X-ray	ROUGE-L	0.1681	0.2159	0.1776	0.1794	0.1830	**0.2084**
	BERTScore F1	0.8494	0.8580	0.8521	0.8515	0.8522	**0.8584**

Table 3: Performance comparison on the medical report impression generation task. All methods share the same k-NN retrieval step over task prototypes; only the combination weights differ. Avg: uniform averaging (Chronopoulou et al. 2023); Sim: softmax similarity-weighted combination (Zhao et al. 2024b); Lin: convex least-squares projection; Ours: ℓ_1-regularized sparse projection (ours). The best zero-shot result for each metric is shown in bold.

	CT (head)	MR	XR
Sim	0.34	0.33	0.33
Lin	0.80	0.18	0.02
Ours	0.82	0.18	0.00

Table 4: Comparison of weight distributions in our similarity-based, linear combination and regularized linear combination methods for CT abdomen medical report impression task.

	A	B	C	D	E	F
DICE	90.5%	86.4%	54.6%	90.0%	0.1%	91.0%

Table 5: Performance of Prostate dataset only the single nearest LoRA. Results are inconsistent across tasks.

erences while suppressing irrelevant ones. This highlights the importance of geometry-aware composition over naïve nearest-neighbor selection.

Enhancing supervised LoRA via cross-task composition. Although our method is designed for zero-shot scenarios where no in-domain LoRA is available, it can also enhance performance when a supervised LoRA exists. By including the supervised adapter in the candidate pool during composition, we evaluate whether cross-task fusion offers benefits. As shown in Table 6, our method assigns the highest weight to the in-domain LoRA (C) while also incorporating support from others (F), and assigning negative weights to less relevant ones (A). This yields a slight improvement in Dice score (90.8% vs. 90.5%), illustrating that our approach can flexibly refine existing adapters by leveraging complementary knowledge—without requiring additional retraining.

Efficiency of Retrieval-based Composition. Finally, we compare the training cost between our method and conventional supervised fine-tuning. For example, training a

	A	B	C	D	E	F
weight	-0.21	-0.07	1.10	0.05	0.03	0.11

Table 6: Linear combination weights for Prostate-C, including its own supervised LoRA. The method selectively blends complementary knowledge.

LoRA on medical report data takes approximately 30 minutes on 8 NVIDIA H100 GPUs. In contrast, retrieving relevant adapters and solving our sparse projection takes under 3 minutes per task—less than 10% of the time. For large-scale systems or privacy-sensitive domains, this efficiency makes retrieval-based composition a practical and scalable alternative to fine-tuning.

Discussion and Conclusion

We introduce a geometry-aware framework for adapter composition by modeling adapter reuse as a sparse reconstruction problem in latent task space. This approach enables interpretable, plug-and-play zero-shot transfer and achieves strong results across multiple domains.

Unlike prior methods based on averaging or similarity-weighted fusion, our sparse projection strategy offers finer control over adapter selection and greater robustness in out-of-distribution settings. Ablation studies show our method can also enhance supervised adapters without retraining.

However, the effectiveness of our approach depends on the diversity and coverage of Adapter-VecDB, which in turn relies on contributions from the broader open-source community. In domains with limited adapters, this remains a constraint. Scaling to larger libraries will also require faster retrieval and more efficient optimization, such as through sparse indexing or prototype clustering. Future directions include jointly learning task prototypes, improving scalability, and extending to multi-modal settings. As adapter ecosystems grow, geometry-aware composition offers a practical

and privacy-preserving alternative to traditional fine-tuning.

Acknowledgments

Quanzheng Li's research is supported in part by the National Institutes of Health under Grant R01HL159183.

References

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2008. Convex multi-task feature learning. *Machine learning*, 73(3): 243–272.

Carass, A.; Roy, S.; Gherman, A.; Reinhold, J. C.; Jesson, A.; Arbel, T.; Maier, O.; Handels, H.; Ghafoorian, M.; Platel, B.; et al. 2020. Evaluating white matter lesion segmentations with refined Sørensen-Dice analysis. *Scientific reports*, 10(1): 8242.

Chronopoulou, A.; Peters, M. E.; Fraser, A.; and Dodge, J. 2023. Adaptersoup: Weight averaging to improve generalization of pretrained language models. *arXiv preprint arXiv:2302.07027*.

Donoho, D. L. 2006. Compressed sensing. *IEEE Transactions on information theory*, 52(4): 1289–1306.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Elhamifar, E.; and Vidal, R. 2013. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11): 2765–2781.

Fan, C.; Lu, Z.; Liu, S.; Gu, C.; Qu, X.; Wei, W.; and Cheng, Y. 2025. Make LoRA Great Again: Boosting LoRA with Adaptive Singular Values and Mixture-of-Experts Optimization Alignment. *arXiv preprint arXiv:2502.16894*.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; and Mikolov, T. 2013. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26.

Gatidis, S.; and Kuestner, T. 2022. A whole-body FDG-PET/CT dataset with manually annotated tumor lesions (FDG-PET-CT-Lesions). Dataset.

Halbheer, M.; Mühlematter, D. J.; Becker, A.; Narnhofer, D.; Aasen, H.; Schindler, K.; and Turkoglu, M. O. 2024. LoRA-Ensemble: Efficient Uncertainty Modelling for Self-attention Networks. *arXiv preprint arXiv:2405.14438*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, C.; Liu, Q.; Lin, B. Y.; Pang, T.; Du, C.; and Lin, M. 2023. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*.

Jiang, M.; Yang, H.; Cheng, C.; and Dou, Q. 2023. IOP-FL: Inside-outside personalization for federated medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(7): 2106–2117.

Kampffmeyer, M.; Chen, Y.; Liang, X.; Wang, H.; Zhang, Y.; and Xing, E. P. 2019. Rethinking knowledge graph propagation for zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11487–11496.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Li, W.; Qu, C.; Chen, X.; Bassi, P. R.; Shi, Y.; Lai, Y.; Yu, Q.; Xue, H.; Chen, Y.; Lin, X.; et al. 2024. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, 97: 103285.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

Liu, Q.; Dou, Q.; and Heng, P. A. 2020. Shape-aware Meta-learning for Generalizing Prostate MRI Segmentation to Unseen Domains. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Luo, M.; Wong, J.; Trabucco, B.; Huang, Y.; Gonzalez, J. E.; Salakhutdinov, R.; Stoica, I.; et al. 2025. Stylus: Automatic adapter selection for diffusion models. *Advances in Neural Information Processing Systems*, 37: 32888–32915.

Luo, X.; Li, Z.; Zhang, S.; Liao, W.; and Wang, G. 2024. Rethinking Abdominal Organ Segmentation (RAOS) in the clinical scenario: A robustness evaluation benchmark with challenging cases. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 531–541. Springer.

Ma, J.; Zhang, Y.; Gu, S.; Zhu, C.; Ge, C.; Zhang, Y.; An, X.; Wang, C.; Wang, Q.; Liu, X.; et al. 2021. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6695–6714.

Ma, X.; Gong, Y.; He, P.; Zhao, H.; and Duan, N. 2023. Query rewriting for retrieval-augmented large language models. *arXiv preprint arXiv:2305.14283*.

Ma, Y.; Liang, Z.; Dai, H.; Chen, B.; Gao, D.; Ran, Z.; Zihan, W.; Jin, L.; Jiang, W.; Zhang, G.; et al. 2024. Modula: Mixture of domain-specific and universal lora for multi-task learning. *arXiv preprint arXiv:2412.07405*.

Mairal, J.; Bach, F.; Ponce, J.; and Sapiro, G. 2010. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(1).

Meng, R.; Song, S.; Jin, P.; Teng, L.; Wang, Y.; Sun, Y.; Chen, L.; Oh, Y.; Li, X.; Li, Q.; et al. 2025. MAST-Pro: Dynamic Mixture-of-Experts for Adaptive Segmentation of Pan-Tumors with Knowledge-Driven Prompts. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 310–320. Springer.

Muqeeth, M.; Liu, H.; and Raffel, C. 2023. Soft merging of experts with adaptive routing. *arXiv preprint arXiv:2306.03745*.

Ostapenko, O.; Su, Z.; Ponti, E. M.; Charlin, L.; Roux, N. L.; Pereira, M.; Caccia, L.; and Sordoni, A. 2024. Towards modular llms by building and reusing a library of loras. *arXiv preprint arXiv:2405.11157*.

Parvez, M. R.; Chi, J.; Ahmad, W. U.; Tian, Y.; and Chang, K.-W. 2022. Retrieval enhanced data augmentation for question answering on privacy policies. *arXiv preprint arXiv:2204.08952*.

Peng, W.; Li, G.; Jiang, Y.; Wang, Z.; Ou, D.; Zeng, X.; Xu, D.; Xu, T.; and Chen, E. 2024. Large language model based long-tail query rewriting in taobao search. In *Companion Proceedings of the ACM on Web Conference 2024*, 20–28.

Pfeiffer, J.; Kamath, A.; Rücklé, A.; Cho, K.; and Gurevych, I. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *16th Conference of the European Chapter of the Associationfor Computational Linguistics, EACL 2021*, 487–503. Association for Computational Linguistics (ACL).

Prabhakar, A.; Li, Y.; Narasimhan, K.; Kakade, S.; Malach, E.; and Jelassi, S. 2024. Lora soups: Merging loras for practical skill composition tasks. *arXiv preprint arXiv:2410.13025*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rezayi, S.; Liu, Z.; Wu, Z.; Dhakal, C.; Ge, B.; Dai, H.; Mai, G.; Liu, N.; Zhen, C.; Liu, T.; et al. 2024. Exploring new frontiers in agricultural nlp: Investigating the potential of large language models for food applications. *IEEE Transactions on Big Data*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Roweis, S. T.; and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500): 2323–2326.

Seo, M.; Baek, J.; Thorne, J.; and Hwang, S. J. 2024. Retrieval-augmented data augmentation for low-resource domain tasks. *arXiv preprint arXiv:2402.13482*.

Shi, Y.; Shu, P.; Liu, Z.; Wu, Z.; Li, Q.; and Li, X. 2024. MGH Radiology Llama: A Llama 3 70B Model for Radiology. *arXiv preprint arXiv:2408.11848*.

Shu, P.; Zhao, H.; Jiang, H.; Li, Y.; Xu, S.; Pan, Y.; Wu, Z.; Liu, Z.; Lu, G.; Guan, L.; et al. 2024. LLMs for Coding and Robotics Education. *arXiv preprint arXiv:2402.06116*.

Shu, Y.; Kou, Z.; Cao, Z.; Wang, J.; and Long, M. 2021. Zoo-tuning: Adaptive transfer from a zoo of models. In *International Conference on Machine Learning*, 9626–9637. PMLR.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, X.; Ye, Y.; and Gupta, A. 2018. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6857–6866.

Wang, Y.; Agarwal, S.; Mukherjee, S.; Liu, X.; Gao, J.; Awadallah, A. H.; and Gao, J. 2022. AdaMix: Mixture-of-adaptations for parameter-efficient model tuning. *arXiv preprint arXiv:2205.12410*.

Wortsman, M.; Ilharco, G.; Gadre, S. Y.; Roelofs, R.; Gontijo-Lopes, R.; Morcos, A. S.; Namkoong, H.; Farhadi, A.; Carmon, Y.; Kornblith, S.; et al. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, 23965–23998. PMLR.

Yang, Z.; Lin, X.; He, Q.; Huang, Z.; Liu, Z.; Jiang, H.; Shu, P.; Wu, Z.; Li, Y.; Law, S.; et al. 2024. Examining the Commitments and Difficulties Inherent in Multimodal Foundation Models for Street View Imagery. *arXiv preprint arXiv:2408.12821*.

Zhai, Y.; Zhang, H.; Lei, Y.; Yu, Y.; Xu, K.; Feng, D.; Ding, B.; and Wang, H. 2023. Uncertainty-penalized reinforcement learning from human feedback with diverse reward lora ensembles. *arXiv preprint arXiv:2401.00243*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhao, H.; Fu, J.; and He, Z. 2023. Prototype-based hyperadapter for sample-efficient multi-task tuning. *arXiv preprint arXiv:2310.11670*.

Zhao, H.; Liu, Z.; Wu, Z.; Li, Y.; Yang, T.; Shu, P.; Xu, S.; Dai, H.; Zhao, L.; Mai, G.; et al. 2024a. Revolutionizing finance with llms: An overview of applications and insights. *arXiv preprint arXiv:2401.11641*.

Zhao, Z.; Gan, L.; Wang, G.; Zhou, W.; Yang, H.; Kuang, K.; and Wu, F. 2024b. Loraretriever: Input-aware lora retrieval and composition for mixed tasks in the wild. *arXiv preprint arXiv:2402.09997*.