Model-Specific Task Similarity for VisionLanguage Model Selection via Layer Conductance

Anonymous ACL submission

Abstract

While open-sourced Vision-Language Models (VLMs) have proliferated, selecting the optimal pre-trained model for a specific downstream task remains challenging. Exhaustive evaluation is often infeasible due to computational constraints and data limitations in few-shot scenarios. Existing selection methods fail to fully address this: they either rely on data-intensive proxies or use symmetric textual descriptors that neglect the inherently directional and model-specific nature of transferability. To address this problem, we propose a framework that grounds model selection in the internal functional dynamics of the visual encoder. Our approach represents each task via layer-wise conductance and derives a target-conditioned block-importance distribution through entropy-regularized alignment. Building on this, we introduce Directional Conductance Divergence (DCD), an asymmetric metric that quantifies how effectively a source task covers the targets salient functional blocks. This allows for predicting target model rankings by aggregating source-task ranks without labeled target-task evaluation. Experimental results on 48 VLMs across 21 datasets demonstrate that our method outperforms state-of-the-art baselines, achieving a 14.7% improvement in NDCG@5 over SWAB.

1 Introduction

VLMs pretrained on large-scale image–text pairs have become core engines for tasks such as visual understanding and embodied intelligence, particularly in zero-shot transfer scenarios (Radford et al., 2021; Ma et al., 2024; Zhang et al., 2024). With the rapid emergence of open-source VLMs spanning diverse architectures and pretraining strategies, selecting the most suitable pretrained model for a target task has become increasingly critical. Exhaustive evaluation is often infeasible, since target domains usually provide only a few or even no test data samples at all. Therefore, transferability estimation becomes the mainstream approach to crack this problem (Ding et al., 2024; Xue et al., 2024).

Transferability estimation is a central and long-standing problem in the transfer learning area (Ding et al., 2024; Xue et al., 2024). Though many methods were proposed, they fall short in the VLMs selection problem. More specifically, data-driven proxy metrics are unstable in few-shot regimes (You et al., 2021; Nguyen et al., 2020). Alternatively, prior task-centric formulations like LOVM (Zohar et al., 2023) and SWAB (Yi et al., 2024) estimate transferability by the embeddings of textual task descriptions. Being target domain data free is one major merit of these methods, but it also leads to the performance bottleneck. This is because task similarity is inherently model-specific (Achille et al., 2019; Mehra et al., 2024). This results in the *data* vs. *performance* dilemma.

This paper develops an alternative approach to address the aforementioned dilemma. Our idea is inspired by practical scenarios that a few unlabeled target images are usually available in the target task. Though these unlabeled target images offer relatively weak visual signals, it turns out that they can elicit strong signals for estimating model-specific task similarity if done properly. To leverage them, we ground task relatedness in model-specific internal dynamics. Concretely, we represent each task by the layer-wise contribution profile of the model's visual encoder, computed via layer conductance over coarse-grained blocks (Dhamdhere et al., 2019). This yields a lightweight, label-free, and inherently model-specific task representation, directly reflecting which functional blocks the target domain relies on.

Building on the above view, we derive a target-conditioned block-importance distribution via entropy-regularized alignment (Cuturi, 2013)

and propose Directional Conductance Divergence (DCD), an asymmetric metric that captures whether source-task mechanisms cover target-critical blocks. Note the asymmetric property is desirable, since transfer is inherently directional and model-specific: tasks that appear similar under one model's representations may diverge under another (Achille et al., 2019). We then predict target-task rankings by aggregating source-task model rankings with DCD-induced similarity weights, avoiding labeled target-task evaluation while relying on amortizable source-side computation. Under limited sampling, our approach achieves roughly a 15% improvement in NDCG@5 and τ@5 over SWAB.

The main contributions of this work are threefold. (I) We introduce a model-specific task representation based on layer conductance of the visual encoder, which requires only a few unlabeled images from the target task. (II) We propose Directional Conductance Divergence (DCD), a target-conditioned asymmetric transferability metric that measures whether source tasks match the target's critical functional blocks. (III) We develop a similarity-weighted rank aggregation scheme to predict target-task model rankings without direct target-task evaluation, and validate its effectiveness across diverse VLMs and datasets.

2 Related Work

Vision-Language Models. VLMs, trained via contrastive learning on large-scale image–text pairs, have become foundational to cross-modal representation learning (Radford et al., 2021). Prominent examples include CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), FLAVA (Singh et al., 2022), Florence (Yuan et al., 2021), and CoCa (Yu et al., 2022). Despite these advances, the performance of VLMs varies considerably across tasks and domains and often degrades under distribution shift (Zohar et al., 2023). Meanwhile, the rapid proliferation of open-source models, where each trained on distinct architectures and corpora, has given rise to a heterogeneous "VLMs Zoo" (Schuhmann et al., 2022; Ilharco et al., 2021; Wolf et al., 2020). In this landscape, strong benchmark results do not guarantee downstream effectiveness, shifting the central challenge from training a single universal model to selecting the most suitable VLM for a given target task. This work investigates label-free and reliable VLM selection from the zoo without exhaustive labeled per-task evaluation.

Model Selection. A key challenge in deploying pre-trained models lies in estimating their transferability. Existing methods address it via proxy metrics such as H-Score (Bao et al., 2019), LogME (You et al., 2021) and LEEP (Nguyen et al., 2020), or through representation similarity such as Task2Vec (Achille et al., 2019) and Model Spider (Zhang et al., 2023). Unsupervised approaches further attempt to predict performance using only unlabeled data signals (Sun et al., 2021). However, these methods uniformly demand sufficient target-task data. This demand ill-suits VLMs, which are often deployed in zero-shot or few-shot scenarios (Zohar et al., 2023; Yi et al., 2024). Moreover, most prior work is developed for unimodal architectures and rarely accounts for VLMs' multimodal nature.

Recent paradigms decouple VLM evaluation from labeled image. LOVM (Zohar et al., 2023) ranks VLMs using only textual task descriptions, while SWAB (Yi et al., 2024) advances this via an optimal transport framework that bridges modality and capability gaps across tasks. Parallel efforts like Model Label Learning (MLL) (Tan et al., 2025) further propose pre-assigning capability labels to VLMs to enable efficient selection and ensemble-based reuse for downstream adaptation. These methods pushed the frontier of state of the art, but they also have limitations worth further investigation: they ignore the practical prevalence of few-shot settings where a small set of target images is available. In this regime, data-heavy methods yield unstable estimates, whereas text-only approaches waste informative visual cues already present in the available samples.

VLMs performance hinges on both visual representation quality and vision–language alignment (Radford et al., 2021; Yang et al., 2024). When limited target-domain images are available, leveraging these samples to reveal performance differences among candidate models becomes essential for reliable selectiona setting that remains largely unexplored. Furthermore, prior task-centric formulations (LOVM (Zohar et al., 2023) and SWAB (Yi et al., 2024)) embed tasks into text space to estimate transferability. This formulation overlooks the fact that task similarity is inherently model-specific (Achille et al., 2019; Mehra et al., 2024): tasks that appear similar under one models representations may diverge substantially un-

der another, due to differences in architectural biases, pre-training objectives, and learned features. Recent empirical analysis (Sachdeva et al., 2025) reinforces this by mapping the complex, non-uniform patterns of positive and negative transfer across different VLM architectures. Ignoring this model-specific nature of task relatedness therefore leads to inaccurate transferability estimation, particularly across heterogeneous VLM collections. Motivated by these observations, we propose a VLMs selection framework tailored to few-shot and weakly supervised settings, which leverages limited visual evidence while explicitly modeling task relatedness in a model-specific manner.

Layer Conductance. Layer conductance is an attribution method that quantifies the contribution of intermediate-layer activations to model outputs (Dhamdhere et al., 2019; Sundararajan et al., 2017). It extends Integrated Gradients from input features to hidden representations, enabling principled attribution while satisfying axiomatic properties such as completeness. We instantiate our framework with layer conductance because it provides block-level attribution signals for internal functional dynamics. Our framework is not inherently restricted to layer conductance, and alternative attribution or representation-analysis methods may provide complementary task representations.

3 Unlabeled Images Elicit Similarity Signals

This section motivates and validates the core premise of our approach that a small set of unlabeled target images can reveal reliable, model-specific signals of task relatedness, which align more closely with transferability than text-only semantic proxies.

Task Representations via Layer Conductance. We consider the setting where a few unlabeled images are available. A classical view in computer vision treats deep neural networks as cascaded filter banks that learn hierarchical representations through successive transformations (Zeiler and Fergus, 2014; Bau et al., 2017). Under this perspective, different layers capture distinct aspects of task-relevant information. If a task consistently relies on specific blocks (e.g., early texture-oriented stages versus later semantic stages), the distribution of layer contributions provides a compact summary of the task as perceived by the model. This intuition resonates with the task

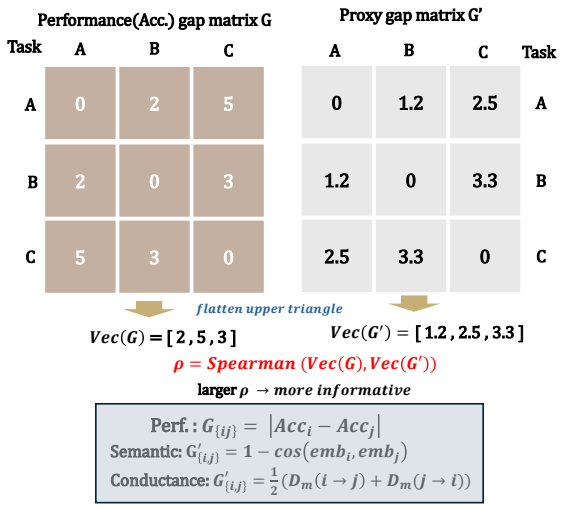


Figure 1: **How we measure proxy reliability.** We construct a performance gap matrix G with entries $G_{ij} = |\mathrm{Acc}_i - \mathrm{Acc}_j|$, and a proxy gap matrix G' from either semantic distances or conductance-induced distances.

embedding paradigm of Task2Vec (Achille et al., 2019), which represents tasks through statistics of internal learning signals. We extend this to a model-specific formulation: rather than constructing a universal task graph, we characterize each task by its layer-wise dependency structure within a given VLM. Specifically, we use layer conductance (Dhamdhere et al., 2019; Sundararajan et al., 2017) to extract a block-wise contribution vector from the visual encoder. We use this distribution as a model-specific summary of how a task engages internal visual blocks, providing a basis for estimating task relatedness from the model's perspective.

Alignment with Performance. We first ask: does a similarity proxy preserve the true structure of transferability? If two tasks are highly related from a model's perspective, the performance gap between them should be small (Wang et al., 2019). A useful proxy should therefore assign smaller distances to task pairs with smaller performance gaps. To probe this, we select five representative VLMs spanning ResNet and ViT backbones with diverse pretraining sources and compute a task–task proxy matrix using: (i) our conductance representation derived from a 25-image unlabeled set, and (ii) a semantic baseline derived from task text embeddings. Figure 1 illustrates these matrices and the corresponding performance gaps, with further details deferred to Appendix B. While absolute correlations are modest, as expected under extreme su-

Model	ρ_{cond} (Ours)	ρ_{sem}	Δ
RN101-OpenAI	**0.128**	-0.127	**+0.255**
RN50x4-OpenAI	**0.131**	-0.095	**+0.227**
ViT-B/32-OpenAI	**0.011**	-0.051	+0.062
RN50-CC12M	**0.023**	0.002	+0.020
ViT-B/16-DFN2B	**0.031**	0.026	+0.006

Table 1: **Alignment with Performance Gap.** Our method (ρ_{cond}) consistently improves over semantic proxies (ρ_{sem}), especially for ResNets where semantic correlation can be negative.

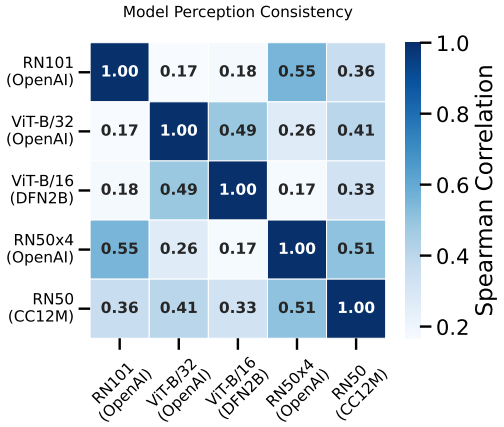


Figure 2: **Model-to-Model Correlation Matrix in Task Perception.** High correlations within architectural families and lower correlations across them suggest that conductance-induced task perception is model-dependent, while also indicating that architecture and pretraining can influence the induced task geometry.

pervision constraints, the conductance proxy consistently aligns more closely with ground truth than the semantic baseline.

Table 1 quantifies this effect via Spearman correlation between the proxy-derived rankings and the performance gap ordering. Notably, for ResNet-based models (RN101-OpenAI and RN50x4-OpenAI), the semantic proxy can even become negatively correlated with ground truth, suggesting that text-only descriptors may invert the transferability signal under certain architectures. In contrast, the conductance proxy remains positively correlated across all five models, confirming its reliability as a selection signal.

The Model-Specific Nature of Task Perception. A second hypothesis central to our method is that task relatedness is model-specific: two tasks perceived as similar by one model may be distinct to another due to differences in architecture, pretraining objectives, and learned features (Achille et al., 2019; Mehra et al., 2024). If true, the task perception induced by conductance should vary across models. To test this, we compare conductance-induced task distance vectors. Figure 2 reports the resulting model-to-model correlation matrix. We observe clear architectural clustering: models within the same backbone family exhibit stronger agreement in relative task distances, while cross-architecture agreement drops substantially. This pattern suggests that conductance-induced task perception is model-dependent, while also indicating that architecture and pretraining can influence the induced task geometry. Rather than treating this clustering as sufficient evidence of task relatedness, we use it as motivation for a model-specific formulation in Section 4, where target-conditioned importance and directional divergence are computed individually for each candidate model.

4 Model Selection Via Unlabeled Image

Inspired by the insights of Section 3, this section designs a method to select appropriate VLMs for downstream tasks utilizing a few unlabeled images from the target task.

4.1 Preliminary

The Model. Let $\mathcal{M}$ be a finite set of candidate VLMs and $\mathcal{T}$ a set of downstream tasks. For a model $m \in \mathcal{M}$, we decompose its visual encoder into d_m coarse-grained functional blocks (e.g., stages in ResNet or transformer blocks in ViT), indexed by $i \in \{1, \ldots, d_m\}$. The number of blocks d_m may vary across models. Each task $\tau \in \mathcal{T}$, is associated with a small unlabeled image set $\mathcal{X}_\tau = \{x_j\}_{j=1}^{N_\tau}$ sampled i.i.d. from its data distribution P_τ. In the leave-one-out protocol, the held-out target task τ provides N_{tgt} unlabeled images, while each source task $\sigma \in \mathcal{T} \setminus \tau$ provides N_{src} images for constructing task representations (Eq. 4).

Layer Conductance Vector. To quantify block-level contributions to the scalar probe, we instantiate our framework with *layer conductance* (Dhamdhere et al., 2019). Given an input image $x \sim P_\tau$ for task $\tau \in \mathcal{T}$, we define the layer conductance vector $g_m(x) \in \mathbb{R}^{d_m}$ as the concatenation of block-wise attribution scores:

$$g_m(x) = \left[g_m^{(1)}(x), \ldots, g_m^{(d_m)}(x) \right]^\top . \quad (1)$$

Layer conductance requires a scalar objective. Let $f_m(x) \in \mathbb{R}^p$ denote the visual embedding produced by model m. In the unlabeled setting, we use the ℓ_2 norm of the image embedding as a simple label-free scalar probe for inducing conductance profiles:

$$F_m(x) \triangleq \|f_m(x)\|_2. \tag{2}$$

This choice is architecture-agnostic and aggregates information across embedding dimensions without requiring labels or class semantics. In practice, Layer conductance produces an attribution tensor for each block output; we summarize it by the mean absolute attribution over activation dimensions to obtain a scalar score $g_m^{(i)}(x) \in \mathbb{R}_{\geq 0}$. The resulting $g_m(x)$ maps x into a model-specific layer contribution space. More details are in Appendix A. Note that F_m is used only as an unsupervised scalar probe to extract relative block-importance patterns, rather than as a direct proxy for downstream accuracy or transferability.

Task Representation. Since a task is characterized by its underlying distribution rather than individual samples, we define the task representation $v_{m,\tau}$ as the expectation of the conductance vector over P_τ:

$$v_{m,\tau} \triangleq \mathbb{E}_{x\sim P_\tau}[g_m(x)] \in \mathbb{R}^{d_m}. \tag{3}$$

Since P_τ is inaccessible in practice, we approximate $v_{m,\tau}$ by an empirical estimate $\hat{v}_{m,\tau}$ computed from a finite unlabeled sample set $\mathcal{X}_\tau = \{x_j\}_{j=1}^N$ drawn i.i.d. from P_τ:

$$\hat{v}_{m,\tau} \triangleq \frac{1}{N}\sum_{j=1}^{N} g_m(x_j). \tag{4}$$

N denotes the number of available unlabeled images for task τ, which equals N_{src} for source tasks and N_{tgt} for the held-out target task.

This vector captures the layer-wise dependency structure of task τ on the internal blocks of model m, and serves as the foundation for the subsequent similarity analysis. For notational convenience, we use $v_{m,\tau}^{(i)}$ to denote the i-th component of $v_{m,\tau}$.

4.2 Target-Conditioned Importance Distribution

Given the task representation $v_{m,\tau}$, we aim to identify which blocks are most critical for the target task τ. Treating all blocks uniformly neglects the hierarchical structure of deep networks, where different blocks capture features of varying semantic complexity. We therefore seek an importance distribution $\alpha_{m,\tau} \in \Delta^{d_m-1}$ over the d_m blocks that emphasizes the most relevant blocks while preserving robustness.

Normalization. To eliminate scale variations across models and ensure numerical stability, we define the normalized representation $u_{m,\tau} \in \mathbb{R}^{d_m}$ as:

$$u_{m,\tau} \triangleq \frac{v_{m,\tau}}{\max(\|v_{m,\tau}\|_2, \epsilon)}, \tag{5}$$

where $\epsilon > 0$ is a small constant for numerical stability. Each component $u_{m,\tau}^{(i)}$ reflects the relative contribution of block i to task τ.

Derivation via Entropy-Regularized Alignment. We require the importance distribution $\alpha_{m,\tau}$ to satisfy two desiderata: (I) higher weights for blocks with larger normalized conductance values (alignment), and (II) sufficient dispersion to avoid over-reliance on any single block (robustness).

Following these principles, we define $\alpha_{m,\tau}$ as the solution to an entropy-regularized alignment objective. Specifically, we seek $\alpha \in \Delta^{d_m-1}$ that maximizes a trade-off between alignment with $u_{m,\tau}$ and Shannon entropy $H(\alpha) = -\sum_i \alpha_i \log \alpha_i$:

$$\alpha_{m,\tau} = \underset{\alpha\in\Delta^{d_m-1}}{\arg\max}\left(\langle \alpha, u_{m,\tau}\rangle + \frac{1}{\eta}H(\alpha)\right), \tag{6}$$

where $\eta > 0$ is a temperature parameter controlling distribution sharpness.

This objective admits a closed-form solution given by the softmax function, which can be easily verified (see Appendix C.2 for a detailed derivation):

$$\alpha_{m,\tau}^{(i)} = \frac{\exp\left(\eta\, u_{m,\tau}^{(i)}\right)}{\sum_{j=1}^{d_m}\exp\left(\eta\, u_{m,\tau}^{(j)}\right)}. \tag{7}$$

The resulting distribution depends solely on the target task representation $v_{m,\tau}$ and re-weights blocks according to task demands. The parameter η modulates the degree of concentration: large η focuses mass on dominant blocks, while small η yields a smoother weighting.

4.3 Directional Conductance Divergence

We define a metric to quantify the transferability from a source task σ to the target task τ. Unlike conventional symmetric metrics (e.g., Euclidean or Cosine distance), effective transfer learning hinges on directional coverage (Zamir et al., 2019): a source task is valuable primarily if it encompasses the internal mechanisms critical to the target.

To capture this asymmetry, we propose Directional Conductance Divergence (DCD). We first define the element-wise relative deviation of the source task σ with respect to the target τ at block i:

$$\delta_m^{(i)}(\tau, \sigma) \triangleq \frac{\left| v_{m,\tau}^{(i)} - v_{m,\sigma}^{(i)} \right|}{\left| v_{m,\tau}^{(i)} \right| + \epsilon}, \tag{8}$$

where $\epsilon > 0$ is a small constant for numerical stability. The denominator normalizes by the target activation magnitude, so the divergence is interpreted as a relative error measured in the target's context.

We then define DCD as the expected relative deviation under the target's importance distribution:

$$\begin{aligned} D_m(\tau \to \sigma) &\triangleq \sum_{i=1}^{d_m} \alpha_{m,\tau}^{(i)} \cdot \delta_m^{(i)}(\tau, \sigma) \\ &= \mathbb{E}_{i \sim \alpha_{m,\tau}} \left[\delta_m^{(i)}(\tau, \sigma) \right] \end{aligned}$$

Intuitively, $D_m(\tau \to \sigma)$ measures how well the source task σ covers the target-critical mechanisms emphasized by $\alpha_{m,\tau}$: smaller values indicate better coverage.

From Divergence to Similarity The DCD $D_m(\tau \to \sigma)$ is a nonnegative, target-conditioned divergence. To aggregate evidence across multiple candidate sources, we convert DCD values into a normalized similarity distribution over the source-task set.

Let $\mathcal{S}_\tau \subseteq \mathcal{T} \setminus \{\tau\}$ denote the set of available source tasks for target τ. For a fixed model m, we define a softmin-based similarity score:

$$s_m(\sigma \mid \tau) \triangleq e^{-\gamma (D_m(\tau \to \sigma) - \min_{\sigma' \in \mathcal{S}_\tau} D_m(\tau \to \sigma'))} \tag{9}$$

where $\gamma > 0$ is a temperature parameter modulating the sharpness of the distribution. Subtracting the minimum DCD provides a numerically stable baseline shift without changing the relative ordering across σ.

We normalize the similarity scores to obtain a probability distribution over source tasks:

$$p_m(\sigma \mid \tau) \triangleq \frac{s_m(\sigma \mid \tau)}{\sum_{\sigma' \in \mathcal{S}_\tau} s_m(\sigma' \mid \tau)} \tag{10}$$

Thus, $p_m(\sigma \mid \tau)$ assigns larger mass to source tasks with smaller DCD, and the distribution becomes more concentrated as γ increases.

Ranking via Similarity Aggregation Given the DCD-induced weights $\{p_m(\sigma \mid \tau)\}_{\sigma \in \mathcal{S}_\tau}$, we estimate how model m would rank on the target task τ.

Assume that for each model m and source task $\sigma \in \mathcal{S}_\tau$, we have access to its performance-based rank $R_m(\sigma)$ among all candidate models (smaller is better). That is, $R_m(\sigma)$ is obtained by sorting models by their ground-truth performance on σ and taking the resulting rank value. We then predict the target-task rank of model m via a similarity-weighted average of source-task ranks:

$$\widehat{R}_m(\tau) \triangleq \sum_{\sigma \in \mathcal{S}_\tau} p_m(\sigma \mid \tau) \cdot R_m(\sigma) \tag{11}$$

Finally, given a candidate model set $\mathcal{M}$, we obtain the estimated ranking for target task τ by sorting models in ascending order of $\widehat{R}_m(\tau)$ (smaller is better):

$$\widehat{\pi}_\tau \triangleq \operatorname{argsort}_{m \in \mathcal{M}} \widehat{R}_m(\tau) \tag{12}$$

This ranking serves as the basis for model selection and is evaluated against ground-truth rankings in our experiments. The theoretical motivation for asymmetry and the full derivations are provided in Appendix C.

5 Experiments

5.1 Experimental Setup

Datasets and Model Zoo. Building upon the protocols established in the LOVM benchmark (Zohar et al., 2023) and the expansion strategies adopted by SWAB (Yi et al., 2024), we establish our experimental setup using their publicly available datasets and model repositories. To facilitate a more comprehensive evaluation, we augment the original model zoo with additional representative architectures, including EVA01-g-14 (Sun et al., 2023) and nllb-clip-base (Visheratin, 2023), to expand the collection to a total of 48 models. This extensive model zoo encapsulates a broad spectrum of architectural designs,

pre-training corpora, and training paradigms. Detailed information on the datasets and models is provided in Appendix D.

Evaluation Protocol. To ensure a rigorous and equitable comparison, we adhere to the leave-one-out evaluation protocol employed by ModelGPT and SWAB (Yi et al., 2024; Zohar et al., 2023). Specifically, we iteratively designate one of the 21 datasets as the *held-out* target domain, while the remaining datasets constitute the source pool. All datasets are obtained from official repositories and evaluated using standard test splits.

To establish ground-truth rankings, we leverage the text templates provided by LOVM (Zohar et al., 2023) to construct zero-shot classifiers and measure the Top-1 accuracy of each VLM on the respective test splits. To evaluate robustness under low-shot sampling, we repeat each experiment over 10 independent random draws of unlabeled images. In each run, we resample the target and source images according to the specified N_{tgt} and N_{src}, recompute conductance-based task representations, and obtain the final ranking. We report the mean and standard deviation across runs.

Metrics. Following the baseline evaluation protocol (Zohar et al., 2023; Yi et al., 2024), we report top-k diagnostics: NDCG@5, Kendall's τ@5, NDCG@3, and NDCG@7. NDCG@5 is more discriminative than Top-5 Recall used in prior studies, as it captures graded relevance and relative ordering rather than set overlap only. We report metric definitions in Appendix G.3 and standard NDCG@5/Recall@1 in Appendix F.1. Additional evaluation results, including hyperparameter sensitivity, sampling efficiency, model clustering, and probe variation, are reported in Appendix F.

5.2 Main Performance Comparison

Baselines. We benchmark our method against several representative baselines from the model selection literature. Specifically, we consider two ranking-based approaches: INB, which ranks VLMs based on their ImageNet performance, and AvgRank, which calculates the mean rank across the remaining 20 source datasets under the leave-one-out protocol. We also include GroupRank, a coarse domain/task grouping baseline. For each target dataset, GroupRank selects source datasets that share the same domain/task group as annotated in Table 5, and predicts each model's target rank by averaging its ground-truth ranks over these grouped source datasets. Furthermore, we compare our approach with two learning-based baselines: ModelGPT (Zohar et al., 2023) and SWAB (Yi et al., 2024), the current state-of-the-art method.

Results. As shown in Table 2, we make the following observations. (I) Under the baseline evaluation protocol, our method achieves the strongest overall performance among the compared methods, with the NDCG@5 of 0.707 and the highest τ@5 of 0.365. (II) The performance gap between our method and SWAB reveals the limitation of relying solely on semantic similarity for model selection. Meanwhile, the gap between GroupRank and DCD suggests that manually defined domain/task groups are informative but insufficient to explain the performance gains. By explicitly accounting for model-specific conductance patterns, our attribution-based approach more effectively captures target-conditioned model–task alignment. (III). The larger gain in NDCG@5 relative to τ@5 indicates that our method is particularly effective at identifying and ranking the best-matched model at the top position, which is crucial for practical deployment.Taken together, these results suggest that effective VLM selection requires a model-specific characterization of task relevance. By grounding similarity in model-aware attribution signals, our approach achieves more accurate model-task alignment with only a few unlabeled target samples and reusable source-task representations.

5.3 Ablation Study

Baselines. We ablate two key components of DCD: the similarity formulation and the block-weighting strategy. For similarity formulation, we replace DCD with two symmetric alternatives, Cosine similarity and Soft-KL/JSD, while keeping the conductance representation and rank aggregation pipeline unchanged. For block weighting, we compare our target-conditioned weighting against Uniform weighting and Target-norm weighting under the same directional relative-deviation distance. These variants isolate whether the gains come from directional similarity modeling, target-conditioned block importance, or their combination. Detailed definitions of all ablation variants are provided in Appendix E.

Metric	INB	AvgRank	ModelGPT	SWAB	GroupRank	Ours
NDCG@5 ↑	0.539	0.524	0.544	0.616	0.609	**0.707**
τ@5 ↑	0.268	0.254	0.267	0.318	0.270	**0.365**
Sum ↑	0.807	0.778	0.811	0.934	0.879	**1.072**
Other metrics:						
NDCG@3 ↑	0.384	0.386	0.386	0.393	0.366	**0.402**
NDCG@7 ↑	0.651	0.699	0.702	0.691	0.741	**0.752**

Table 2: **Main results on VLM selection.** NDCG@5 and τ@5 are averaged over 10 independent random low-shot sampling runs for stochastic methods. Our method is evaluated in the lightweight regime with N_{tgt}=1 target image and N_{src}=25 source images per task. The corresponding distribution over 10 runs is provided in Appendix G.2. Best in **bold**, second-best underlined.

Variant	NDCG@5 ↑	τ@5 ↑	Sum ↑
Similarity formulation			
Cosine (Sym.)	0.639	0.301	0.940
Soft-KL/JSD (Sym.)	0.640	0.302	0.942
Weighting strategy			
Uniform weighting	0.625	0.254	0.879
Target-norm weighting	0.670	0.243	0.913
DCD (Ours)	**0.707**	**0.365**	**1.072**

Table 3: **Ablation study.** We evaluate two groups of variants: similarity formulation and block-weighting strategy. Cosine and Soft-KL/JSD replace DCD with symmetric similarity measures, while Uniform weighting and Target-norm weighting replace the target-conditioned block weighting used in DCD. DCD combines directional relative deviation with target-conditioned block weighting. Detailed variant definitions are provided in Appendix E.

Results. Table 3 reports the results of both ablation groups. (I) In the similarity formulation group, DCD consistently outperforms both symmetric alternatives. Compared with the stronger symmetric baseline, DCD improves NDCG@5 from 0.640 to 0.707 and τ@5 from 0.302 to 0.365, confirming the importance of directional source-to-target coverage. (II) Cosine similarity and Soft-KL/JSD perform nearly identically, suggesting that once the similarity is forced to be symmetric, the choice between geometric and distributional comparison has limited impact. (III) In the block-weighting group, Uniform weighting already provides a non-trivial signal, while Target-norm weighting improves NDCG@5 from 0.625 to 0.670. The full DCD formulation achieves the best overall performance, especially on τ@5, indicating that target-conditioned block importance and directional relative deviation are both important for robust model ranking.

6 Conclusion

We introduced a model-specific framework for few-shot vision-language model selection that characterizes tasks through layer-wise contributions in the visual encoder and quantifies transferability using Directional Conductance Divergence (DCD), an asymmetric metric measuring target-critical feature coverage. Empirical evaluation across 48 VLMs and 21 datasets shows significant improvements over existing baselines, achieving NDCG@5 of 0.707 and τ@5 of 0.365, representing gains of approximately 15% over SWAB, while theoretical analysis supports the necessity of asymmetry under coverage semantics.

Future work can improve both the generality and practicality of our framework. We will investigate architecture-aware strategies to align functional blocks across heterogeneous backbones, and extend attribution signals to vision-language interaction for alignment-sensitive tasks. We also plan to strengthen performance under extreme sampling through adaptive anchor selection and uncertainty-aware aggregation. Finally, incorporating robustness and fairness criteria would enable more reliable deployment across diverse scenarios.

Limitations

First, our evaluation mainly relies on top-k ranking diagnostics, including intersection-based NDCG@k and Kendall's τ@k. While these metrics are useful for comparing the relative ordering of commonly selected top models, they do not exhaustively capture ranking errors over the full model zoo. We therefore report standard full-ranking metrics in Appendix F.1 and treat the intersection-based metrics as complementary diagnostics. Moreover, our current instantiation uses

layer conductance and the embedding norm as a label-free scalar probe; although this design avoids labels and text prompts and is suitable for low-supervision settings, alternative attribution methods, scalar probes, or vision–language alignment objectives may further improve task representation.

Second, conductance representations are inevitably affected by model architecture and pre-training. Although the GroupRank baseline and model-clustering analysis suggest that coarse task/domain groups or architecture identity alone cannot fully explain the observed gains, fully disentangling task-specific functional structure from architecture-induced effects remains future work. In addition, conductance extraction is more computationally expensive than a standard forward-only pass. Source-side representations can be precomputed and reused, but target-side attribution over a large model zoo remains the main practical bottleneck, motivating more efficient attribution extraction and architecture-aware caching strategies.

References

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C. Fowlkes, Stefano Soatto, and Pietro Perona. 2019. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6429–6438.

Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas J. Guibas. 2019. An information-theoretic approach to transferability in task transfer learning. In *Proceedings of the IEEE International Conference on Image Processing*, pages 2309–2313.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3319–3327.

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.

Adam Coates, Andrew Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 215–223.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300.

Kedar Dhamdhere, Mukund Sundararajan, and Qiqi Yan. 2019. How important is a neuron. In *International Conference on Learning Representations*.

Yuhe Ding, Bo Jiang, Aijing Yu, Aihua Zheng, and Jian Liang. 2024. Which model to transfer? A survey on transferability estimation. *Preprint*, arXiv:2402.15231.

Emma Dugas, Will Cukierski, and 1 others. 2015. Diabetic retinopathy detection. Kaggle Competition.

Mark Everingham, Luc van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2010. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research*, 32(11):1231–1237.

Ian J. Goodfellow, Dumitru Erhan, Pierre-Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, and 1 others. 2013. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. Zenodo.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning*, pages 4904–4916.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, 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.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and 1 others. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 554–561.

Alex Krizhevsky and Geoffrey Hinton. 2009. Learning multiple layers of features from tiny images.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 2002. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. 2024. A survey on vision-language-action models for embodied AI. *Preprint*, arXiv:2405.14093.

Akshay Mehra, Yunbei Zhang, and Jihun Hamm. 2024. Understanding the transferability of representations via task-relatedness. In *Advances in Neural Information Processing Systems*.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y. Ng, and 1 others. 2011. Reading digits in natural images with unsupervised feature learning. NIPS Workshop.

Cuong V. Nguyen, Tal Hassner, Matthias W. Seeger, and Cédric Archambeau. 2020. LEEP: A new measure to evaluate transferability of learned representations. In *Proceedings of the International Conference on Machine Learning*, pages 7294–7305.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. 2012. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505.

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 International Conference on Machine Learning*, pages 8748–8763.

Bhuvan Sachdeva, Karan Uppal, Abhinav Java, and Vineeth N. Balasubramanian. 2025. Understanding task transfer in vision-language models. *Preprint*, arXiv:2511.18787.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15617–15629.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. 2011. The german traffic sign recognition benchmark: A multi-class classification competition. In *Proceedings of the International Joint Conference on Neural Networks*, pages 1453–1460.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. EVA-CLIP: Improved training techniques for CLIP at scale. *Preprint*, arXiv:2303.15389.

Xiaoxiao Sun, Yunzhong Hou, Weijian Deng, Hongdong Li, and Liang Zheng. 2021. Ranking models in unlabeled new environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11741–11751.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning*, pages 3319–3328.

Hao-Zhe Tan, Zhi Zhou, Yufeng Li, and Lan-Zhe Guo. 2025. Vision-language model selection and reuse for downstream adaptation. In *Proceedings of the International Conference on Machine Learning*.

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.

Alexander A. Visheratin. 2023. NLLB-CLIP: Train performant multilingual image retrieval model on a budget. *Preprint*, arXiv:2309.01859.

Boyu Wang, Jorge A. Mendez, Mingbo Cai, and Eric Eaton. 2019. Transfer learning via minimizing the performance gap between domains. In *Advances in Neural Information Processing Systems*, pages 10644–10654.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clément Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3485–3492.

Yihao Xue, Rui Yang, Xiaohan Chen, Weibo Liu, Zidong Wang, and Xiaohui Liu. 2024. A review on transferability estimation in deep transfer learning. *IEEE Transactions on Artificial Intelligence*, 5(12):5894–5914.

Shijia Yang, Bohan Zhai, Quanzeng You, Jianbo Yuan, Hongxia Yang, and Chenfeng Xu. 2024. Law of vision representation in MLLMs. *Preprint*, arXiv:2408.16357.

Chao Yi, Yuhang He, De-Chuan Zhan, and Han-Jia Ye. 2024. Bridge the modality and capability gaps in vision-language model selection. In *Advances in Neural Information Processing Systems*.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. 2021. LogME: Practical assessment of pre-trained models for transfer learning. In *Proceedings of the International Conference on Machine Learning*, pages 12133–12143.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. CoCa: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*.

Lu Yuan and 1 others. 2021. Florence: A new foundation model for computer vision. *Preprint*, arXiv:2111.11432.

Amir Zamir, Alexander Sax, William B. Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. 2019. Taskonomy: Disentangling task transfer learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 6241–6245.

Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, and 1 others. 2019. A large-scale study of representation learning with the visual task adaptation benchmark. *Preprint*, arXiv:1910.04867.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644.

Yi-Kai Zhang, Ting-Ji Huang, Yao-Xiang Ding, De-Chuan Zhan, and Han-Jia Ye. 2023. Model spider: Learning to rank pre-trained models efficiently. In *Advances in Neural Information Processing Systems*.

Orr Zohar, Shih-Cheng Huang, Kuan-Chieh Wang, and Serena Yeung. 2023. LOVM: Language-only vision model selection. In *Advances in Neural Information Processing Systems*.

A Layer Conductance's Detail

Intuition. Layer conductance quantifies the contribution of individual neurons to the final prediction by measuring the flow of attribution through them. A neuron with high conductance contributes more strongly to the scalar output along the attribution path, suggesting a larger role in the corresponding forward computation.

Background and notation. Let $F : \mathbb{R}^n \to \mathbb{R}$ denote the scalar model output used for attribution, with input $x \in \mathbb{R}^n$ and baseline $x' \in \mathbb{R}^n$. We consider the straight-line path $\gamma(\alpha) = x' + \alpha(x - x')$ for $\alpha \in [0, 1]$.

Scalar probe used in this paper. For a VLM m, let $f_m(x) \in \mathbb{R}^p$ be its image embedding. We define

$$F_m(x) \triangleq \|f_m(x)\|_2, \tag{13}$$

and compute layer conductance with respect to $F_m(x)$. This serves as a simple label-free scalar probe that does not require labels, text prompts, or class semantics, and aggregates attribution across embedding dimensions. Unless otherwise specified, we use the default settings in Captum (Kokhlikyan et al., 2020).

Integrated Gradients. Integrated Gradients (IG) attributes the prediction to input features. For the i-th feature,

$$\mathrm{IG}_i(x) := (x_i - x_i') \int_0^1 \frac{\partial F(\gamma(\alpha))}{\partial x_i}\, d\alpha. \tag{14}$$

Layer conductance formulation. For a hidden neuron y (scalar activation), layer conductance extends IG by decomposing attribution flow through

y. The conductance of y with respect to input feature i is

$$\mathrm{Cond}_i^y(x) := (x_i - x_i') \int_0^1 \frac{\partial F(\gamma(\alpha))}{\partial y} \frac{\partial y}{\partial x_i} \, d\alpha. \tag{15}$$

The total conductance of y aggregates contributions from all inputs:

$$\mathrm{Cond}^y(x) := \sum_i (x_i - x_i') \int_0^1 \frac{\partial F(\gamma(\alpha))}{\partial y} \frac{\partial y}{\partial x_i} \, d\alpha. \tag{16}$$

For a layer $h(x) \in \mathbb{R}^m$ with neurons $\{y_j\}_{j=1}^m$, layer conductance returns $\{\mathrm{Cond}^{y_j}(x)\}_{j=1}^m$. It satisfies completeness: $\sum_{j=1}^m \mathrm{Cond}^{y_j}(x) = F(x) - F(x')$.

Efficient computation. Direct evaluation of Eq. 16 requires computing $\frac{\partial y}{\partial x_i}$ for all input dimensions, which is costly. By recognizing the equivalence between total conductance and path integrated gradients on hidden activations, we rewrite

$$\mathrm{Cond}^y(x) = \int_0^1 \frac{\partial F(\gamma(\alpha))}{\partial \gamma_y(\alpha)} \frac{\partial \gamma_y(\alpha)}{\partial \alpha} \, d\alpha, \tag{17}$$

where $\gamma_y(\alpha) = y(\gamma(\alpha))$ is the activation of y along the path. We approximate the integral via Riemann sum. Let $x^{(k)} = x' + \frac{k}{n}(x - x')$ and $y^{(k)} = y(x^{(k)})$ for $k = 0, \ldots, n$. Then

$$\mathrm{Cond}^y(x) \approx \sum_{k=1}^n \frac{\partial F(x^{(k)})}{\partial y^{(k)}} \left(y^{(k)} - y^{(k-1)} \right). \tag{18}$$

This formulation avoids computing input-level gradients and scales efficiently to high-dimensional settings.

Block score aggregation. Captum returns a tensor-valued conductance attribution for the chosen block output (Kokhlikyan et al., 2020). We convert it to a scalar block score by taking the mean absolute attribution over all tensor entries:

$$g_m^{(i)}(x) \triangleq \mathrm{Mean}\left(\left| \mathrm{Cond}_m^{(i)}(x) \right| \right), \tag{19}$$

where $\mathrm{Cond}_m^{(i)}(x)$ denotes the conductance attribution tensor for block i. We then form $g_m(x) = [g_m^{(1)}(x), \ldots, g_m^{(d_m)}(x)]^\top$, and compute the task representation by averaging $g_m(x)$ over unlabeled samples.

Implementation. We use PyTorch Captum (captum.attr.LayerConductance), which implements the efficient computation in Eq. 18. For detailed derivations and theoretical properties, see (Dhamdhere et al., 2019; Kokhlikyan et al., 2020).

B Details regarding Section 3

We analyze five representative VLMs (ResNet and ViT backbones; diverse pre-training sources) on the same $|\mathcal{T}|$=21 downstream tasks. For each task τ, we sample N_{src}=25 unlabeled images and compute conductance-based task representations $\hat{v}_{m,\tau}$ following Section 4.1. We use η=2.5 and ϵ=10^{-8} throughout.

Performance gap matrix (Figure 1, left). For each model m, let $a_m(\tau)$ be its benchmark score on task τ (e.g., accuracy; for non-accuracy metrics we use the task's primary scalar score). We define the ground-truth task-pair performance gap matrix

$$G_m(\tau, \sigma) \triangleq \left| a_m(\tau) - a_m(\sigma) \right|. \tag{20}$$

Proxy gap matrix (Figure 1, right). *Semantic:* we embed each task τ using model m's text encoder (averaging normalized text features over class-name prompts `'A photo of {c}.'`) to obtain $e_m(\tau)$, and set

$$G_m^{\text{sem}}(\tau, \sigma) \triangleq 1 - \cos\big(e_m(\tau), e_m(\sigma)\big). \tag{21}$$

Conductance: we compute the target-conditioned importance $\alpha_{m,\tau}$ (Section 4.2) and directional divergence (Section 4.3)

$$D_m(\tau \to \sigma) \triangleq \sum_{i=1}^{d_m} \alpha_{m,\tau}^{(i)} \cdot \frac{\left| \hat{v}_{m,\tau}^{(i)} - \hat{v}_{m,\sigma}^{(i)} \right|}{\left| \hat{v}_{m,\tau}^{(i)} \right| + \epsilon}, \tag{22}$$

then symmetrize it to match the undirected proxy gap in Figure 1:

$$G_m^{\text{cond}}(\tau, \sigma) \triangleq \tfrac{1}{2}\left(D_m(\tau \to \sigma) + D_m(\sigma \to \tau) \right). \tag{23}$$

Proxy reliability (Table 1). Let $\mathrm{Vec}(\cdot)$ flatten the upper-triangular entries excluding the diagonal (Figure 1). For each model m, we compute

$$\begin{aligned} \rho_{\text{cond}}(m) &= \mathrm{Spearman}\big(\mathrm{Vec}(G_m), \mathrm{Vec}(G_m^{\text{cond}})\big), \\ \rho_{\text{sem}}(m) &= \mathrm{Spearman}\big(\mathrm{Vec}(G_m), \mathrm{Vec}(G_m^{\text{sem}})\big). \end{aligned} \tag{24}$$

Model-specific task perception (Figure 2). For any two models (m_1, m_2), we measure agreement in conductance-induced task distances by

$$\mathrm{Corr}(m_1, m_2) = \mathrm{Spearman}\big(\mathrm{Vec}(G_{m_1}^{\text{cond}}), \mathrm{Vec}(G_{m_2}^{\text{cond}})\big). \tag{25}$$

and stack $\mathrm{Corr}(m_1, m_2)$ into the model–model correlation matrix.

C Theoretical Analysis and Derivations

Our method adopts model-specific task representations and a directional transfer distance, contrasting with prior symmetric approaches (Zohar et al., 2023; Yi et al., 2024). This appendix provides the theoretical motivation for asymmetry under a target-conditioned coverage view of transfer, derives the entropy-regularized target-conditioned importance distribution, and proves that Directional Conductance Divergence (DCD) is a smooth relaxation of an idealized set-restricted formulation.

C.1 Coverage Semantics and Asymmetry

Definition 1 (Salient Set)**.** *Let $\mathbf{u}_{m,\tau} \in \mathbb{R}^{d_m}$ be the normalized representation of task τ on model m. Let $\rho_k(\mathbf{u}_{m,\tau})$ denote the k-th largest value in $\mathbf{u}_{m,\tau}$. The* salient set *of task τ is defined as*

$$\mathcal{S}_{m,\tau}^{(k)} \triangleq \left\{ i \in \{1, \dots, d_m\} \;\middle|\; u_{m,\tau}^{(i)} \geq \rho_k(\mathbf{u}_{m,\tau}) \right\}. \tag{26}$$

This set represents the top-k functional blocks on which task τ relies most heavily; for an appropriate choice of k, it serves as a proxy for the task-critical blocks.

Coverage Semantics for Model Selection. We argue that transfer distance from a source task σ to a target task τ should follow a target-conditioned *coverage semantics*. In our setting, coverage does not merely mean that the source activates a superset of target-critical blocks. Instead, because DCD compares conductance magnitudes, a source is considered useful when it matches or closely approximates the target on the target-critical functional blocks. Discrepancies outside the target salient set should not be heavily penalized, since they correspond to source-specific mechanisms that are not essential for the target.

This target-conditioned view naturally induces asymmetry: a source task may sufficiently cover the blocks required by the target, while the target may fail to cover additional blocks that are salient for the source. We formalize this principle via two assumptions on a directional distance function $d(\cdot \to \cdot)$.

Assumption 1 (Target-Sufficiency)**.** *If the source task σ matches the target task τ on all indices in the target salient set in terms of normalized conductance magnitude, the distance should be zero regardless of discrepancies outside the target salient set:*

$$\forall i \in \mathcal{S}_{m,\tau}^{(k)}, \quad u_{m,\sigma}^{(i)} = u_{m,\tau}^{(i)} \implies d(\tau \to \sigma) = 0.$$

Assumption 2 (Salient-Discriminativity)**.** *Conversely, if the direction is reversed and the target task τ fails to match the source task σ on some block that is salient for σ, then the reverse distance must be strictly positive:*

$$\exists j \in \mathcal{S}_{m,\sigma}^{(k)} \text{ s.t. } u_{m,\tau}^{(j)} \neq u_{m,\sigma}^{(j)} \implies d(\sigma \to \tau) > 0.$$

Proposition 1 (Impossibility of Symmetry)**.** *Let $d(\cdot \to \cdot)$ satisfy Assumptions 1 and 2. Suppose $\mathcal{S}_{m,\tau}^{(k)} \subsetneq \mathcal{S}_{m,\sigma}^{(k)}$, and the representations match on $\mathcal{S}_{m,\tau}^{(k)}$ but differ on at least one residual index in $\mathcal{S}_{m,\sigma}^{(k)} \setminus \mathcal{S}_{m,\tau}^{(k)}$. Then d is necessarily asymmetric.*

Proof. Evaluate the distance in both directions:

- **Forward ($\tau \to \sigma$).** Since σ matches τ on all indices in $\mathcal{S}_{m,\tau}^{(k)}$, Assumption 1 yields $d(\tau \to \sigma) = 0$.
- **Reverse ($\sigma \to \tau$).** Since $\mathcal{S}_{m,\tau}^{(k)} \subsetneq \mathcal{S}_{m,\sigma}^{(k)}$, there exists at least one index $j \in \mathcal{S}_{m,\sigma}^{(k)} \setminus \mathcal{S}_{m,\tau}^{(k)}$ where the representations differ. By Assumption 2, $d(\sigma \to \tau) > 0$.

Hence $d(\tau \to \sigma) \neq d(\sigma \to \tau)$, so d is asymmetric. □

Remark 1. Proposition 1 reveals a limitation of symmetric metrics, such as cosine similarity and Euclidean distance, under target-conditioned coverage semantics. A symmetric metric cannot distinguish whether the unmatched blocks are irrelevant source-specific mechanisms or target-critical mechanisms. As a result, symmetric metrics may over-penalize comprehensive source tasks that contain additional mechanisms not required by the target.

C.2 Entropy-Regularized Alignment

In this section, we provide the detailed derivation for the closed-form solution of the importance distribution $\alpha_{m,\tau}$. For notational simplicity and without loss of generality, we drop the subscripts (m, τ) and denote the alignment vector as $u \in \mathbb{R}^d$ and the decision variable as $\alpha \in \Delta^{d-1}$.

Problem Formulation. Consider the entropy-regularized alignment problem:

$$\max_{\alpha \in \Delta^{d-1}} \left\{ \langle \alpha, u \rangle + \frac{1}{\eta} H(\alpha) \right\}, \tag{27}$$

where $\Delta^{d-1} = \{\alpha \in \mathbb{R}^d : \alpha \geq 0,\ \mathbf{1}^\top \alpha = 1\}$ denotes the $(d-1)$-dimensional probability simplex, $H(\alpha) = -\sum_i \alpha_i \log \alpha_i$ is the Shannon entropy, and $\eta > 0$ is a temperature parameter.

Closed-Form Solution.

Theorem 1. *The optimization problem in (27) admits a unique global maximizer α^*, given by:*

$$\begin{aligned} \alpha^* &= \mathrm{softmax}(\eta u), \\ \alpha_i^* &= \frac{\exp(\eta u_i)}{\sum_{j=1}^d \exp(\eta u_j)}. \end{aligned} \tag{28}$$

Proof. The proof proceeds in three steps.

Step 1: Strict Concavity and Interior Solution. The objective $\mathcal{J}(\alpha)$ is strictly concave over $\mathrm{int}(\Delta^{d-1})$, being the sum of a linear function and the strictly concave entropy $H(\alpha)$. This guarantees uniqueness of the maximizer. Moreover, since

$$\frac{\partial H(\alpha)}{\partial \alpha_i} = -(\log \alpha_i + 1) \to +\infty \quad \text{as } \alpha_i \to 0^+,$$

the maximizer cannot occur on the boundary of the simplex. Hence the optimal solution lies strictly in the simplex interior, i.e., $\alpha_i^* > 0$ for all i, and the non-negativity constraints are inactive.

Step 2: KKT Conditions. We form the Lagrangian with multiplier λ for the equality constraint:

$$\mathcal{L}(\alpha, \lambda) = \langle \alpha, u \rangle - \frac{1}{\eta} \sum_i \alpha_i \log \alpha_i + \lambda(\mathbf{1}^\top \alpha - 1). \tag{29}$$

The stationarity condition $\nabla_\alpha \mathcal{L} = 0$ yields:

$$u_i - \frac{1}{\eta}(\log \alpha_i + 1) + \lambda = 0, \quad \forall i. \tag{30}$$

Rearranging (30) gives

$$\log \alpha_i = \eta(u_i + \lambda) - 1 \quad \Longrightarrow \quad \alpha_i = e^{\eta\lambda - 1} \cdot e^{\eta u_i}. \tag{31}$$

Denoting $C \triangleq e^{\eta\lambda - 1}$, we have $\alpha_i = C \cdot e^{\eta u_i}$.

Step 3: Normalization. Imposing $\mathbf{1}^\top \alpha = 1$:

$$C \sum_{j=1}^d e^{\eta u_j} = 1 \quad \Longrightarrow \quad C = \left(\sum_{j=1}^d e^{\eta u_j} \right)^{-1}. \tag{32}$$

Substituting back yields (28). Since $\exp(\eta u_i) > 0$ for all i, the solution satisfies $\alpha_i^* > 0$, confirming feasibility. □

C.3 DCD as a Continuous Relaxation

The target-conditioned coverage semantics above can be instantiated by a set-restricted divergence that focuses solely on the target's salient blocks. Recall that $\mathcal{S}_{m,\tau}^{(k)}$ denotes the salient set of task τ on model m (Definition 1). Define the tail mass of the importance distribution as

$$t_{m,\tau}^{(k)}(\eta) \triangleq \sum_{i \notin \mathcal{S}_{m,\tau}^{(k)}} \alpha_{m,\tau}^{(i)}. \tag{33}$$

The hard importance distribution supported exclusively on $\mathcal{S}_{m,\tau}^{(k)}$ is then

$$\bar{\alpha}_{m,\tau}^{(i)}(\eta) \triangleq \begin{cases} \alpha_{m,\tau}^{(i)} / (1 - t_{m,\tau}^{(k)}(\eta)), & i \in \mathcal{S}_{m,\tau}^{(k)}, \\ 0, & \text{otherwise}, \end{cases} \tag{34}$$

which yields the *set-restricted divergence*

$$d_m^{(k)}(\tau \to \sigma; \eta) \triangleq \sum_{i=1}^{d_m} \bar{\alpha}_{m,\tau}^{(i)}(\eta)\, \delta_m^{(i)}(\tau, \sigma). \tag{35}$$

Here, $d_m^{(k)}$ focuses only on target-salient blocks and ignores discrepancies outside $\mathcal{S}_{m,\tau}^{(k)}$. It therefore exactly captures the idealized target-sufficiency principle in Assumption 1. However, it requires a discrete choice of k, for which no universally principled criterion exists. We therefore adopt DCD $D_m(\tau \to \sigma)$ as a smooth relaxation, replacing the hard indicator $\mathbf{1}[i \in \mathcal{S}_{m,\tau}^{(k)}]$ with softmax weights $\alpha_{m,\tau}$. We quantify this relaxation via the tail mass $t_{m,\tau}^{(k)}(\eta)$.

Lemma 1 (Tail Mass Bound)**.** *Let $u_{(1)} \geq \cdots \geq u_{(d_m)}$ be the sorted entries of $\mathbf{u}_{m,\tau}$. Assume $\Delta_k \triangleq u_{(k)} - u_{(k+1)} > 0$. Then the tail mass satisfies*

$$t_{m,\tau}^{(k)}(\eta) \leq \frac{d_m - k}{k} \exp(-\eta \Delta_k). \tag{36}$$

Proof. Let $\mathcal{S} \triangleq \mathcal{S}_{m,\tau}^{(k)}$ and denote $\alpha^{(i)} \triangleq \alpha_{m,\tau}^{(i)}$ for brevity. Since $\Delta_k > 0$, the k-th and $(k+1)$-th

largest entries are separated, and hence $|\mathcal{S}| = k$. By definition of softmax,

$$\sum_{i\notin\mathcal{S}} \alpha^{(i)} = \sum_{i\notin\mathcal{S}} \frac{\exp(\eta u_{m,\tau}^{(i)})}{\sum_{j=1}^{d_m} \exp(\eta u_{m,\tau}^{(j)})} = \frac{\sum_{i\notin\mathcal{S}} \exp(\eta u_{m,\tau}^{(i)})}{\sum_{j=1}^{d_m} \exp(\eta u_{m,\tau}^{(j)})}. \tag{37}$$

Lower bound the denominator by keeping only the contribution of indices in $\mathcal{S}$:

$$\sum_{j=1}^{d_m} \exp(\eta u_{m,\tau}^{(j)}) \geq \sum_{j\in\mathcal{S}} \exp(\eta u_{m,\tau}^{(j)}). \tag{38}$$

Hence

$$\sum_{i\notin\mathcal{S}} \alpha^{(i)} \leq \frac{\sum_{i\notin\mathcal{S}} \exp(\eta u_{m,\tau}^{(i)})}{\sum_{j\in\mathcal{S}} \exp(\eta u_{m,\tau}^{(j)})}. \tag{39}$$

For any $i \notin \mathcal{S}$, $u_{m,\tau}^{(i)} \leq u_{(k+1)}$, while for any $j \in \mathcal{S}$, $u_{m,\tau}^{(j)} \geq u_{(k)}$. Therefore,

$$\begin{aligned} \sum_{i\notin\mathcal{S}} \exp(\eta u_{m,\tau}^{(i)}) &\leq (d_m - k)\exp(\eta u_{(k+1)}), \\ \sum_{j\in\mathcal{S}} \exp(\eta u_{m,\tau}^{(j)}) &\geq k\exp(\eta u_{(k)}). \end{aligned} \tag{40}$$

Combining the above inequalities yields

$$\begin{aligned} \sum_{i\notin\mathcal{S}} \alpha^{(i)} &\leq \frac{(d_m - k)\exp(\eta u_{(k+1)})}{k\exp(\eta u_{(k)})} \\ &= \frac{d_m - k}{k}\exp\big(-\eta(u_{(k)} - u_{(k+1)})\big) \\ &= \frac{d_m - k}{k}\exp(-\eta\Delta_k). \end{aligned} \tag{41}$$

This concludes the proof of Lemma 1. □

Proposition 2 (DCD as a Set-Restricted Relaxation)**.** *Assume that conductance representations are uniformly bounded, i.e., there exists $V_{\max} > 0$ such that $|v_{m,\tau}^{(i)}| \leq V_{\max}$ for all m, τ, i. Since the denominator in $\delta_m^{(i)}(\tau,\sigma)$ is lower bounded by ϵ, there exists $B = 2V_{\max}/\epsilon$ such that $\delta_m^{(i)}(\tau,\sigma) \leq B$ for all i. For any source task σ, the following holds:*

$$D_m(\tau \to \sigma) = \Big(1 - t_{m,\tau}^{(k)}(\eta)\Big)\, d_m^{(k)}(\tau \to \sigma;\eta) + r_m^{(k)}(\tau,\sigma;\eta), \tag{42}$$

where

$$r_m^{(k)}(\tau,\sigma;\eta) \triangleq \sum_{i\notin\mathcal{S}_{m,\tau}^{(k)}} \alpha_{m,\tau}^{(i)}\, \delta_m^{(i)}(\tau,\sigma). \tag{43}$$

Furthermore,

$$\begin{aligned} 0 \leq r_m^{(k)}(\tau,\sigma;\eta) &\leq B\, t_{m,\tau}^{(k)}(\eta), \\ \big|D_m(\tau \to \sigma) - d_m^{(k)}(\tau \to \sigma;\eta)\big| &\leq 2B\, t_{m,\tau}^{(k)}(\eta). \end{aligned} \tag{44}$$

Consequently, combining with Lemma 1, $D_m(\tau \to \sigma)$ approximates the set-restricted divergence $d_m^{(k)}(\tau \to \sigma;\eta)$ with an $O(\exp(-\eta\Delta_k))$ error.

Proof. Let $\mathcal{S} \triangleq \mathcal{S}_{m,\tau}^{(k)}$, $\alpha^{(i)} \triangleq \alpha_{m,\tau}^{(i)}$, and $\delta^{(i)} \triangleq \delta_m^{(i)}(\tau,\sigma)$ for brevity. Define the tail mass $t \triangleq \sum_{i\notin\mathcal{S}} \alpha^{(i)}$, so that $\sum_{i\in\mathcal{S}} \alpha^{(i)} = 1 - t$. By definition,

$$\begin{aligned} D_m(\tau \to \sigma) &= \sum_{i=1}^{d_m} \alpha^{(i)}\delta^{(i)} \\ &= \sum_{i\in\mathcal{S}} \alpha^{(i)}\delta^{(i)} + \sum_{i\notin\mathcal{S}} \alpha^{(i)}\delta^{(i)}. \end{aligned} \tag{45}$$

Define the hard distribution $\bar{\alpha}^{(i)}$ supported on $\mathcal{S}$ by

$$\bar{\alpha}^{(i)} = \begin{cases} \alpha^{(i)}/(1-t), & i \in \mathcal{S}, \\ 0, & i \notin \mathcal{S}. \end{cases} \tag{46}$$

Then the set-restricted divergence is

$$\begin{aligned} d_m^{(k)}(\tau \to \sigma;\eta) &= \sum_{i=1}^{d_m} \bar{\alpha}^{(i)}\delta^{(i)} \\ &= \sum_{i\in\mathcal{S}} \frac{\alpha^{(i)}}{1-t}\delta^{(i)}. \end{aligned} \tag{47}$$

Thus,

$$\begin{aligned} \sum_{i\in\mathcal{S}} \alpha^{(i)}\delta^{(i)} &= (1-t)\sum_{i\in\mathcal{S}} \frac{\alpha^{(i)}}{1-t}\delta^{(i)} \\ &= (1-t)\, d_m^{(k)}(\tau \to \sigma;\eta). \end{aligned} \tag{48}$$

Let

$$r_m^{(k)}(\tau,\sigma;\eta) \triangleq \sum_{i\notin\mathcal{S}} \alpha^{(i)}\delta^{(i)}. \tag{49}$$

This gives the claimed decomposition:

$$\begin{aligned} D_m(\tau \to \sigma) = (1-t)\, d_m^{(k)}(\tau \to \sigma;\eta) \\ + r_m^{(k)}(\tau,\sigma;\eta). \end{aligned} \tag{50}$$

Since each $\alpha^{(i)} \geq 0$ and each $\delta^{(i)} \geq 0$, we have $r_m^{(k)}(\tau, \sigma; \eta) \geq 0$. Under the boundedness condition $\delta^{(i)} \leq B$ for all i,

$$\begin{aligned} r_m^{(k)}(\tau, \sigma; \eta) &= \sum_{i \notin \mathcal{S}} \alpha^{(i)} \delta^{(i)} \\ &\leq \sum_{i \notin \mathcal{S}} \alpha^{(i)} \cdot B \\ &= Bt. \end{aligned} \tag{51}$$

This proves $0 \leq r_m^{(k)}(\tau, \sigma; \eta) \leq Bt$. Moreover, since $d_m^{(k)}(\tau \to \sigma; \eta)$ is an expectation of $\delta^{(i)}$ under a probability distribution supported on $\mathcal{S}$, it also satisfies $0 \leq d_m^{(k)}(\tau \to \sigma; \eta) \leq B$. Therefore,

$$\begin{aligned} &D_m(\tau \to \sigma) - d_m^{(k)}(\tau \to \sigma; \eta) \\ &\quad = -t\, d_m^{(k)}(\tau \to \sigma; \eta) + r_m^{(k)}(\tau, \sigma; \eta). \end{aligned} \tag{52}$$

Taking absolute values gives

$$\begin{aligned} &\left| D_m(\tau \to \sigma) - d_m^{(k)}(\tau \to \sigma; \eta) \right| \\ &\quad \leq tB + Bt = 2Bt. \end{aligned} \tag{53}$$

Finally, substituting the bound on t from Lemma 1 yields the stated $O(\exp(-\eta \Delta_k))$ approximation rate. □

Corollary 1. *Assume $\Delta_k > 0$ and the boundedness condition in Proposition 2. If the source matches the target on the target salient set, i.e.,*

$$\delta_m^{(i)}(\tau, \sigma) = 0 \quad \textit{for all } i \in \mathcal{S}_{m,\tau}^{(k)},$$

then

$$\begin{aligned} D_m(\tau \to \sigma) &= r_m^{(k)}(\tau, \sigma; \eta) \\ &\leq B\, t_{m,\tau}^{(k)}(\eta) \\ &\leq B \cdot \frac{d_m - k}{k} \exp(-\eta \Delta_k). \end{aligned} \tag{54}$$

Hence $D_m(\tau \to \sigma) \to 0$ when η or Δ_k is sufficiently large. Thus, DCD recovers the discrete Target-Sufficiency semantics in Assumption 1 up to an exponentially small relaxation error, while remaining intrinsically asymmetric as required by Proposition 1, through its target-conditioned weighting $\alpha_{m,\tau}$.

D Dataset and Model Details

D.1 Model Zoo

We curated a diverse model zoo by extending the collections in LOVM (Zohar et al., 2023) and SWAB (Yi et al., 2024). Incorporating additional architectures like EVA01-g-14 (Sun et al., 2023) and nllb-clip-base (Visheratin, 2023), our final benchmark comprises 48 models, as detailed in Table 4.

D.2 Datasets

To evaluate model selection under diverse real-world conditions, we employ datasets spanning various image domains and tasks, as summarized in Table 5.

E Additional Details of Ablation Variants

Similarity formulation variants. For the similarity formulation ablation, all variants use the same conductance-based task representation and rank aggregation pipeline, but differ in how source–target task similarity is computed. Cosine similarity measures the normalized inner product between task representation vectors. Soft-KL/JSD first applies softmax normalization to task representations and then computes the Jensen–Shannon divergence as a symmetric distributional distance. Both variants replace the directional DCD similarity with symmetric alternatives.

Block-weighting variants. For the block-weighting ablation, all variants use the same directional relative-deviation distance:

$$D_m(\tau \to \sigma) = \sum_i \alpha_{m,\tau}^{(i)} \frac{|v_{m,\tau}^{(i)} - v_{m,\sigma}^{(i)}|}{|v_{m,\tau}^{(i)}| + \epsilon}. \tag{55}$$

They differ only in the choice of the block-importance weights $\alpha_{m,\tau}$. Uniform weighting assigns equal importance to all blocks, i.e., $\alpha_{m,\tau}^{(i)} = 1/d_m$. Target-norm weighting sets $\alpha_{m,\tau}^{(i)} \propto |v_{m,\tau}^{(i)}|$, directly using the magnitude of target conductance as block importance. DCD uses the entropy-regularized target-conditioned distribution $\alpha_{m,\tau}^{(i)} = \mathrm{softmax}(\eta u_{m,\tau}^{(i)})$, where $u_{m,\tau}$ is the normalized target representation.

F Additional Evaluation Results

F.1 Standard Ranking Metrics

In addition to the metrics used in the main experiments, Table 6 reports standard ranking metrics over the full candidate model set. NDCG@5 (no intersection) evaluates the predicted top-5 list without restricting evaluation to the intersection

Model (Family)	Pretrain Tag	Params(M)	FLOPs(B)	ImageNet Acc.(%)
RN50	openai	102.01	18.18	59.82
RN50	cc12m	102.01	18.18	35.91
RN101	openai	119.69	25.50	62.28
RN101	yfcc15m	119.69	25.50	34.07
RN101-quickgelu	openai	119.69	25.50	62.28
RN101-quickgelu	yfcc15m	119.69	25.50	34.87
RN50x4	openai	178.30	51.82	66.27
RN50x64	openai	623.26	552.65	73.91
ViT-B-32	openai	151.28	14.78	63.32
ViT-B-32	laion2b_e16	151.28	14.78	65.65
ViT-B-32	datacomp_xl_s13b_b90k	151.28	14.78	69.17
ViT-B-32	commonpool_m_clip_s128m_b4k	151.28	14.78	27.25
ViT-B-32-256	datacomp_s34b_b86k	151.29	17.46	72.81
ViT-B-32-quickgelu	laion400m_e31	151.28	14.78	62.94
ViT-B-32-quickgelu	metaclip_fullcc	151.28	14.78	67.66
ViT-B-16	openai	149.62	41.09	68.34
ViT-B-16	laion2b_s34b_b88k	149.62	41.09	70.23
ViT-B-16	datacomp_l_s1b_b8k	149.62	41.09	63.10
ViT-B-16	commonpool_l_laion_s1b_b8k	149.62	41.09	55.26
ViT-B-16	dfn2b	149.62	41.09	76.24
ViT-B-16-quickgelu	metaclip_fullcc	149.62	41.09	72.12
ViT-B-16-plus-240	laion400m_e31	208.38	64.03	69.04
ViT-L-14	openai	427.62	175.33	75.54
ViT-L-14	laion400m_e31	427.62	175.33	72.71
ViT-L-14	datacomp_xl_s13b_b90k	427.62	175.33	79.21
ViT-L-14	commonpool_xl_clip_s13b_b90k	427.62	175.33	76.37
ViT-L-14-quickgelu	metaclip_fullcc	427.62	175.33	79.17
ViT-L-14-quickgelu	dfn2b	427.62	175.33	81.41
ViT-L-14-336	openai	427.94	395.22	76.56
ViT-H-14	laion2b_s32b_b79k	986.11	381.68	77.96
ViT-H-14-quickgelu	metaclip_fullcc	986.11	381.68	80.51
ViT-H-14-378-quickgelu	dfn5b	986.71	1054.05	84.37
ViT-g-14	laion2b_s12b_b42k	1366.68	581.15	76.63
ViT-bigG-14	laion2b_s39b_b160k	2539.57	1065.36	80.09
roberta-ViT-B-32	laion2b_s12b_b32k	212.72	105.87	61.71
xlm-roberta-base-ViT-B-32	laion5b_s13b_b90k	366.12	105.87	62.36
convnext_base_w	laion2b_s13b_b82k	179.39	49.38	70.78
convnext_base_w_320	laion_aesthetic_s13b_b82k	179.39	71.94	71.67
convnext_large_d	laion2b_s26b_b102k_augreg	351.77	107.50	75.91
convnext_large_d_320	laion2b_s29b_b131k_ft	351.77	157.98	76.60
convnext_xxlarge	laion2b_s34b_b82k_augreg_soup	1200.58	443.03	79.47
coca_ViT-B-32	laion2b_s13b_b90k	253.56	33.34	63.31
coca_ViT-L-14	laion2b_s13b_b90k	638.45	214.52	75.61
EVA01-g-14	laion400m_s11b_b41k	1136.44	547.36	78.52
EVA02-B-16	merged2b_s8b_b131k	149.69	41.09	74.72
EVA02-L-14-336	merged2b_s6b_b61k	428.08	395.16	80.39
nllb-clip-base	v1	501.89	369.60	24.32
nllb-clip-base-siglip	v1	507.47	472.91	39.09

Table 4: Overview of the model zoo used in our experiments. The table reports the model name, pre-training tag, parameter count, FLOPs, and ImageNet classification accuracy. Most data are sourced from (Tan et al., 2025).

between predicted and ground-truth top-5 models. Recall@1 measures whether the top-ranked predicted model matches the ground-truth best model.

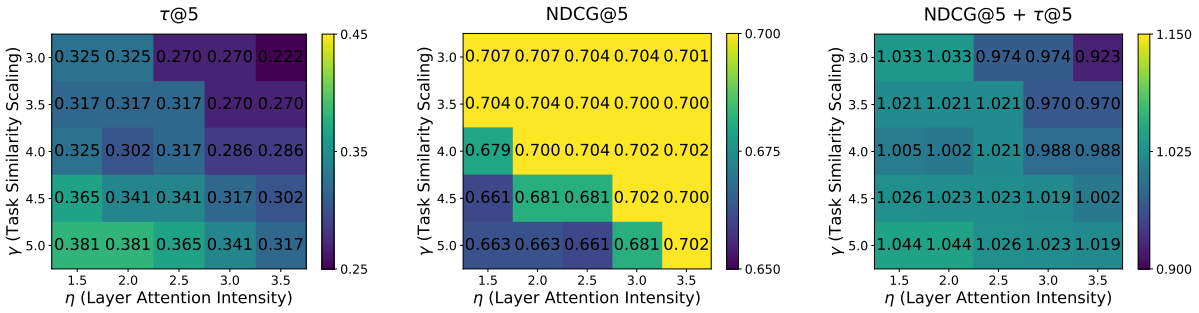
F.2 Hyperparameter Sensitivity

Our method incorporates two key hyperparameters: η (layer attention intensity), which regulates the softmax temperature to weight network blocks based on their conductance, and γ, which scales task similarity to modulate discriminative strength when aggregating rankings from source tasks. We evaluate the sensitivity of our approach by varying $\eta \in \{1.5, 2.0, 2.5, 3.0, 3.5\}$ and $\gamma \in \{3.0, 3.5, 4.0, 4.5, 5.0\}$. The results are visualized as heatmaps in Figure 3.

The empirical results demonstrate that our method is highly robust to hyperparameter variations. Specifically, varying η from 1.5 to 3.5 yields only marginal fluctuations in NDCG@5, suggesting that task representations remain discriminative provided that informative layers are sufficiently prioritized. Similarly, performance remains stable across the tested range of γ, with τ@5 ranging from 0.222 to 0.381. The configuration of $\eta = 2.0$ and $\gamma = 5.0$ attains the peak τ@5 (0.365), striking an effective balance between layer-wise noise suppression and task discrimination. These findings confirm that our method maintains strong performance across a broad parameter space without the need for exhaustive per-task tuning.

Dataset	Domain	Task
Stanford Cars (Krause et al., 2013)	car picture	image classification
CIFAR100 (Krizhevsky and Hinton, 2009)	natural picture	image classification
CLEVR-D (Johnson et al., 2017)	natural picture	object distance estimation
Country211 (Radford et al., 2021)	natural picture	geo-localization
Retinopathy (Dugas et al., 2015)	retina scan	image classification
DMLab (Zhai et al., 2019)	natural picture	object distance estimation
DTD (Cimpoi et al., 2014)	texture picture	image classification
EuroSAT (Helber et al., 2019)	satellite images	image classification
FER2013 (Goodfellow et al., 2013)	facial picture	facial expression classification
GTSRB (Stallkamp et al., 2011)	traffic picture	image classification
KITTI (Geiger et al., 2013)	nature picture	object distance estimation
MNIST (LeCun et al., 2002)	digit picture	image classification
Flowers102 (Nilsback and Zisserman, 2008)	flower picture	image classification
Oxford Pet (Parkhi et al., 2012)	pet photograph	image classification
PCam (Veeling et al., 2018)	medical picture	image classification
RenderedSST2 (Radford et al., 2021)	text picture	optical character recognition
RESISC45 (Cheng et al., 2017)	satellite picture	land cover classification
STL10 (Coates et al., 2011)	natural picture	image classification
SUN397 (Xiao et al., 2010)	natural picture	scene classification
SVHN (Netzer et al., 2011)	natural picture	OCR
VOC2007 (Everingham et al., 2010)	natural picture	image and pose classification

Table 5: Datasets used in our evaluation, with the corresponding domain type and task. Most data are sourced from (Zohar et al., 2023).



(a) τ@5 (higher is better) (b) NDCG@5 (higher is better) (c) Sum = NDCG@5 + τ@5

Figure 3: **Sensitivity to η and γ.** Heatmaps of selection performance when varying η (softmax temperature for target-conditioned block importance) and γ (softmin temperature for DCD-based source weighting). All values are averaged over 10 independent random low-shot sampling runs. Performance is stable across a wide range.

Method	NDCG@5 (no intersection) ↑	Recall@1 ↑
ModelGPT	0.826	0.333
SWAB	0.828	0.381
Ours	**0.833**	**0.476**

Table 6: **Standard ranking metrics.** NDCG@5 (no intersection) evaluates the predicted top-5 list over the full candidate model set, while Recall@1 evaluates whether the top-ranked predicted model matches the ground-truth best model.

F.3 Efficiency of Task Sampling

We separately vary the number of unlabeled images for source-task and target-task characterization. For source-task sampling, we fix $N_{\text{tgt}} = 1$ and vary N_{src} from 1 to 100, averaging over 5 independent random low-shot sampling runs. As shown in Figure 4(a), performance improves rapidly when N_{src} increases to 25 and then saturates around 0.71 NDCG@5 and 0.38 τ@5. This suggests that a compact source library with about 25 anchor images per task is sufficient and reusable across target tasks.

We further vary N_{tgt} while keeping the source library fixed. As shown in Figure 4(b), performance stays within a small fluctuation range across target sample sizes, with NDCG@5 close to 0.70 and competitive τ@5 even in the single-shot setting. Thus, very few target images already provide informative model-specific conductance sig-

nals, and increasing N_{tgt} brings limited additional gain. Small fluctuations at larger N_{tgt} should not be over-interpreted, as part of the variation may come from sampling noise.

F.4 Model Clustering in Conductance induced Task Space

To examine whether conductance-induced task similarity merely reflects architecture identity, we cluster all 48 models using HDBSCAN on their task-similarity structures. For each model, we compute conductance-based task representations with 25 unlabeled samples per task, derive its pairwise similarity matrix over 21 tasks, flatten the matrix, and apply softmax normalization before clustering.

Table 7 shows that the induced clusters partly follow architecture and pretraining patterns, but are not determined by architecture alone. For example, several clusters mix different families, including ConvNeXt with EVA, NLLB-CLIP with ViT, and text-variant models with ViT. This suggests that the conductance-induced task space captures model-dependent task perception beyond coarse architecture identity, while still being influenced by architecture and pretraining.

F.5 Variation of the Scalar Probe

To check whether the scalar probe $F_m(x) = \|f_m(x)\|_2$ is overly normalized, we report its statistics across representative models and tasks in Table 8. For each model–task pair, we sample 50 images and compute the mean, standard deviation, coefficient of variation (CV), minimum, and maximum. The values vary across models and tasks, indicating that the probe is not a degenerate constant. We use $F_m(x)$ only to induce conductance profiles, not as a direct proxy for downstream accuracy or transferability.

G Miscellaneous Details

G.1 Computational Cost

Conductance extraction is more expensive than a standard forward-only pass, and our method is not designed for raw latency superiority. Its practical advantage lies in source-side reusability and low-supervision target-side selection. The cost decomposes into an offline stage, where source-task conductance representations are computed once and reused across target tasks, and an online stage, where only a few unlabeled target samples are used to compute target conductance before lightweight rank aggregation.

In our implementation, extracting target conductance across 48 VLMs with one target sample takes approximately 11.2 hours. Increasing the target sample size to 150 takes approximately 180 hours, which is sub-linear in the sample count due to fixed overhead such as model loading. By comparison, standard forward evaluation on 5 labeled target images across all candidate models takes approximately 2.7 hours, but requires target labels and does not benefit from source-side amortization.

G.2 Distribution over 10 Runs for Main Results

Figure 5 shows the boxplots of the main results over 10 random runs.

G.3 Evaluation Metrics

For each task τ, we compare estimated ranking $\widehat{\pi}_\tau$ (from $\widehat{R}_m(\tau)$, Sec. 4.3) against ground-truth π^*_τ (from $R_m(\tau)$) on the Top-k intersection

$$\begin{aligned}\mathcal{I}^k_\tau \triangleq \{m \mid \hat{r}_\tau(m) \le k\} \\ \cap \{m \mid r^*_\tau(m) \le k\},\end{aligned} \tag{56}$$

where $\hat{r}_\tau(m)$ and $r^*_\tau(m)$ denote predicted and ground-truth ranks. We report NDCG@k, Kendall's τ@k (SciPy implementation with tie-handling), and their sum; both metrics are set to 0 when $|\mathcal{I}^k_\tau| < 2$.

NDCG@k. Define relevance $\text{rel}_\tau(m) \triangleq \text{rank}_\downarrow(r^*_\tau(m))$. Ordering $\mathcal{I}^k_\tau$ by $\hat{r}_\tau(m)$ as $(m_{(1)}, \ldots, m_{(n)})$,

$$\text{NDCG@k}(\tau) \triangleq \frac{\sum_{i=1}^{n} \frac{2^{\text{rel}_\tau(m_{(i)})}-1}{\log_2(i+1)}}{\text{IDCG@k}(\tau)}. \tag{57}$$

where $n = |\mathcal{I}^k_\tau|$ and IDCG@k(τ) is the ideal DCG@k.

τ@k.

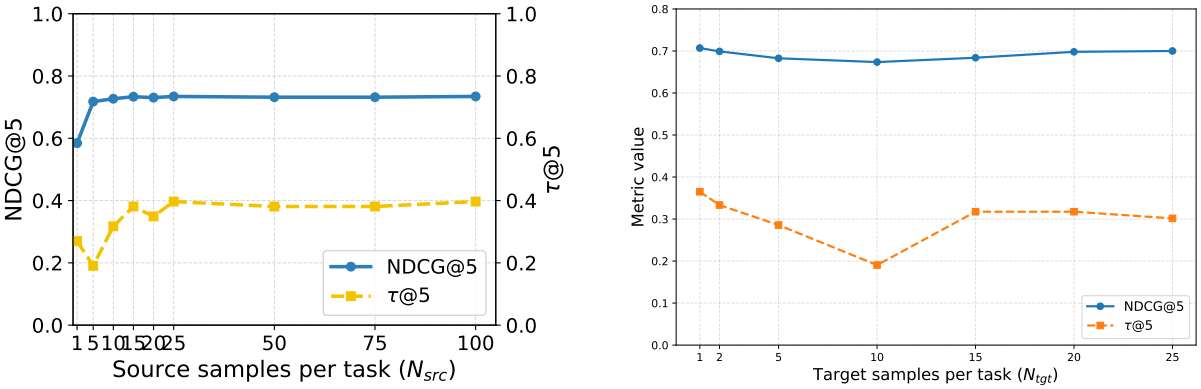
$$\tau@k(\tau) \triangleq \frac{C - D}{\binom{n}{2}}, \tag{58}$$

where C and D count concordant and discordant pairs between predicted and ground-truth ranks on $\mathcal{I}^k_\tau$.

Sum. $\text{Score}(\tau) \triangleq \text{NDCG@5}(\tau) + \tau@5(\tau)$.

Table 7: **Model clustering in conductance-induced task space.** HDBSCAN clustering is performed on softmax-normalized task-similarity vectors derived from layer conductance. Cluster -1 denotes noise points. Clusters marked with $\star$ contain multiple architecture families.

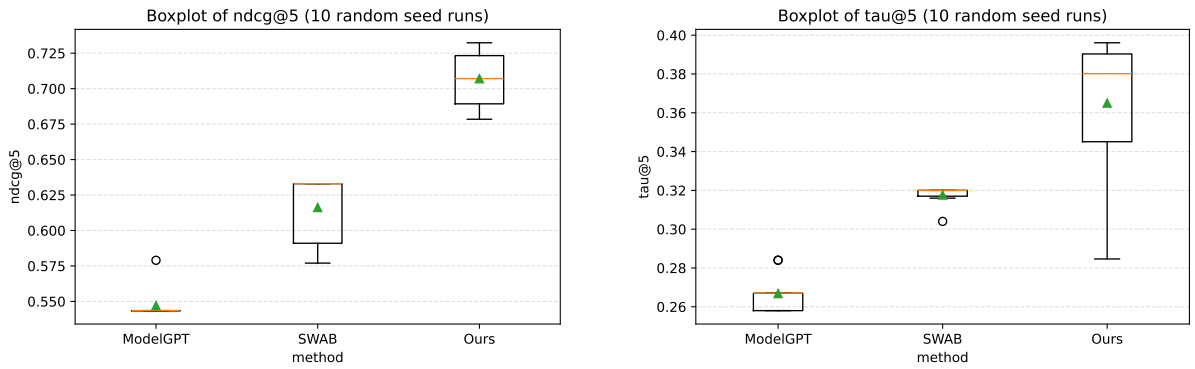
Cluster	Family	Architecture	Model
$-1^{\dagger}$	CoCa	CoCa	coca_vit_b_32_laion2b_s13b_b90k
	CoCa	CoCa	coca_vit_l_14_laion2b_s13b_b90k
	EVA	EVA	eva02_b_16_merged2b_s8b_b131k
	EVA	EVA	eva02_l_14_336_merged2b_s6b_b61k
	ResNet	ResNet-50x	rn50x64_openai
	ViT	ViT-B/16	vit_b_16_dfn2b
	ViT	ViT-B/16	vit_b_16_laion2b_s34b_b88k
	ViT	ViT-B/16	vit_b_16_plus_240_laion400m_e31
	ViT	ViT-B/16	vit_b_16_quickgelu_metaclip_fullcc
	ViT	ViT-B/32	vit_b_32_256_datacomp_s34b_b86k
	ViT	ViT-B/32	vit_b_32_commonpool_m_clip_s128m...
	ViT	ViT-B/32	vit_b_32_quickgelu_metaclip_fullcc
	ViT	ViT-bigG	vit_bigg_14_laion2b_s39b_b160k
	ViT	ViT-H	vit_h_14_quickgelu_metaclip_fullcc
	ViT	ViT-L	vit_l_14_quickgelu_metaclip_fullcc
0	ResNet	ResNet-101	rn101_quickgelu_yfcc15m
	ResNet	ResNet-101	rn101_yfcc15m
	ResNet	ResNet-50	rn50_cc12m
1	ResNet	ResNet-101	rn101_openai
	ResNet	ResNet-101	rn101_quickgelu_openai
	ResNet	ResNet-50	rn50_openai
	ResNet	ResNet-50x	rn50x4_openai
2	ConvNeXt	ConvNeXt	convnext_base_w_320_laion_aesthe...
	ConvNeXt	ConvNeXt	convnext_base_w_laion2b_s13b_b82k
$3\star$	ConvNeXt	ConvNeXt	convnext_large_d_320_laion2b_s29...
	ConvNeXt	ConvNeXt	convnext_large_d_laion2b_s26b_b1...
	ConvNeXt	ConvNeXt	convnext_xxlarge_laion2b_s34b_b8...
	EVA	EVA	eva01_g_14_laion400m_s11b_b41k
$4\star$	NLLB-CLIP	NLLB-CLIP	nllb_clip_base_siglip_v1
	ViT	ViT-B/16	vit_b_16_openai
	ViT	ViT-H	vit_h_14_378_quickgelu_dfn5b
	ViT	ViT-L	vit_l_14_336_openai
	ViT	ViT-L	vit_l_14_openai
	ViT	ViT-L	vit_l_14_quickgelu_dfn2b
5	ViT	ViT-B/16	vit_b_16_commonpool_l_laion_s1b_b8k
	ViT	ViT-B/16	vit_b_16_datacomp_l_s1b_b8k
$6\star$	Text-variant	XLM-RoBERTa	xlm_roberta_base_vit_b_32_laion5...
	ViT	ViT-B/32	vit_b_32_datacomp_xl_s13b_b90k
	ViT	ViT-B/32	vit_b_32_openai
	ViT	ViT-G	vit_g_14_laion2b_s12b_b42k
	ViT	ViT-H	vit_h_14_laion2b_s32b_b79k
	ViT	ViT-L	vit_l_14_commonpool_xl_clip_s13b...
	ViT	ViT-L	vit_l_14_datacomp_xl_s13b_b90k
	ViT	ViT-L	vit_l_14_laion400m_e31
$7\star$	Text-variant	RoBERTa	roberta_vit_b_32_laion2b_s12b_b32k
	ViT	ViT-B/32	vit_b_32_laion2b_e16
$8\star$	NLLB-CLIP	NLLB-CLIP	nllb_clip_base_v1
	ViT	ViT-B/32	vit_b_32_quickgelu_laion400m_e31



(a) Source samples per task (N_{src}).

(b) Target samples per task (N_{tgt}).

Figure 4: **Efficiency of Task Sampling.** Source-task sampling saturates around $N_{\text{src}} = 25$, while target-task sampling remains stable with small fluctuations across different N_{tgt} values.



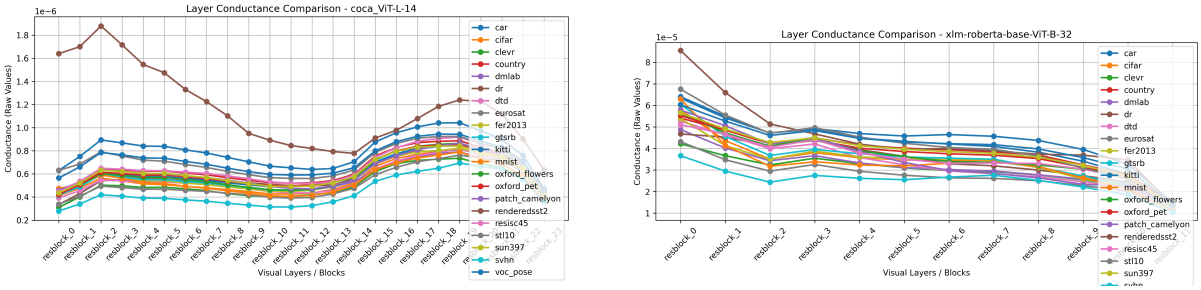
(a) NDCG@5 over 10 runs.

(b) Kendall's τ@5 over 10 runs.

Figure 5: Boxplots of main results over 10 random sampling runs. Each box summarizes the distribution across runs for three methods (ModelGPT, SWAB, and ours).

G.4 Ground-Truth Rankings and Conductance Examples

To characterize the dataset-dependent nature of model performance, we summarize the ground-truth ranking of all models on each dataset.Figure 7 shows that the induced model order is not universal. We further visualize layer conductance patterns aggregated over all datasets for two representative models (Figure 6), including coca-ViT-L-14 (laion2b_s13b_b90k) and xlm-roberta-base-ViT-B-32 (laion5b_s13b_b90k).



(a) coca-ViT-L-14 (laion2b_s13b_b90k).

(b) xlm-roberta-base-ViT-B-32 (laion5b_s13b_b90k).

Figure 6: Layer conductance patterns aggregated over all datasets for two representative models.

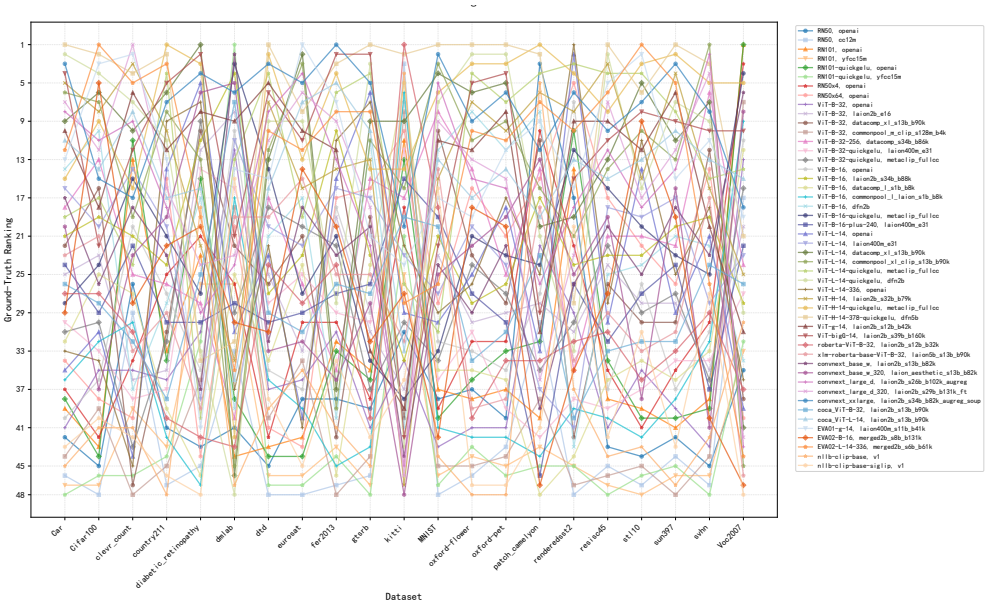


Figure 7: Ground-truth model ranking across all datasets. The ranking order varies across datasets.

Table 8: **Variation of the scalar probe across models and tasks.** We report statistics of $F_m(x) = \|f_m(x)\|_2$ across 5 representative models and 6 tasks, with 50 images per model–task pair. CV denotes the coefficient of variation, i.e., σ/μ.

Model	Task	Mean	Std	CV	Min	Max
ViT-B-16	CIFAR-100	11.28	0.2904	0.0257	10.34	11.81
	MNIST	11.51	0.2220	0.0193	11.07	11.96
	STL-10	11.11	0.4540	0.0409	10.18	12.10
	SVHN	11.47	0.4701	0.0410	10.52	12.57
	GTSRB	11.40	0.4302	0.0377	10.48	12.31
	DTD	11.41	0.3385	0.0297	10.75	12.09
ViT-L-14	CIFAR-100	20.11	0.9469	0.0471	17.62	21.94
	MNIST	21.56	0.4689	0.0218	20.50	22.95
	STL-10	19.97	0.8796	0.0441	18.45	21.94
	SVHN	20.87	0.5883	0.0282	19.51	22.79
	GTSRB	19.43	0.8268	0.0425	17.56	20.92
	DTD	19.91	0.6529	0.0328	18.38	21.26
RN50	CIFAR-100	2.15	0.1762	0.0819	1.83	2.55
	MNIST	2.10	0.1030	0.0490	1.89	2.33
	STL-10	2.07	0.2104	0.1017	1.63	2.63
	SVHN	2.17	0.1123	0.0517	1.92	2.43
	GTSRB	1.95	0.1443	0.0738	1.65	2.23
	DTD	1.89	0.2267	0.1197	1.62	2.79
convnext_base_w	CIFAR-100	7.93	0.5370	0.0677	6.88	9.51
	MNIST	7.51	0.1971	0.0263	7.07	7.98
	STL-10	8.53	0.4890	0.0573	7.53	9.49
	SVHN	7.25	0.3494	0.0482	6.69	8.26
	GTSRB	7.59	0.5462	0.0720	6.52	8.73
	DTD	9.31	0.8790	0.0944	7.81	11.50
EVA02-B-16	CIFAR-100	0.12	0.0079	0.0682	0.11	0.15
	MNIST	0.16	0.0129	0.0796	0.14	0.19
	STL-10	0.11	0.0049	0.0452	0.10	0.12
	SVHN	0.14	0.0123	0.0877	0.11	0.17
	GTSRB	0.11	0.0076	0.0671	0.10	0.13
	DTD	0.13	0.0096	0.0730	0.11	0.16