Rethinking Model Selection in VLM Through the Lens of Gromov-Wasserstein Distance

Supplementary Material

8. Theoretical Analysis (Full)

Due to space limitations, we provide some technical analysis and the full proof of Theorem 1 here.

8.1. Definitions

Note the following definition is given in [31], which we put here for completeness.

Definition 4 (Approximate realizability[31])**.** *Define the 'approximate realizability' of a hypothesis class $\mathcal{G}$ on a paired dataset $D = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$ as*

$$\mathcal{R}(\mathcal{G}, D) = \min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \|g(\boldsymbol{x}) - \boldsymbol{y}\|. \tag{12}$$

8.2. Lemma 1

Lemma 1. *For any $(\boldsymbol{x}, \boldsymbol{y})$, $(\boldsymbol{x}', \boldsymbol{y}') \in \mathrm{supp}(\pi^*_\infty)$, we have:*

$$|d_{\mathcal{Y}}(g^*(\boldsymbol{x}), g^*(\boldsymbol{x}')) - d_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}')| \leq GW_\infty + 2\rho^*_{\pi_\infty}. \tag{13}$$

Proof. To prove the lemma, we need to prove

$$d_{\mathcal{Y}}(g^*(\boldsymbol{x}), g^*(\boldsymbol{x}')) - d_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}') \leq \mathrm{GW}_\infty + 2\rho^*_{\pi_\infty}, \quad \text{(case 1)}$$

$$d_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}') - d_{\mathcal{Y}}(g^*(\boldsymbol{x}), g^*(\boldsymbol{x}')) \leq \mathrm{GW}_\infty + 2\rho^*_{\pi_\infty}, \quad \text{(case 2)}$$

respectively.

Case 1. We start by bounding $d_{\mathcal{Y}}(g^*(\boldsymbol{x}), g^*(\boldsymbol{x}'))$. By triangle inequality, we have:

$$d_{\mathcal{Y}}(g^*(\boldsymbol{x}), g^*(\boldsymbol{x}')) \leq d_{\mathcal{Y}}(g^*(\boldsymbol{x}), \boldsymbol{y}) + d_{\mathcal{Y}}(\boldsymbol{y}, g^*(\boldsymbol{x}')).$$

Applying triangle inequality again on $d_{\mathcal{Y}}(g^*(\boldsymbol{x}'), \boldsymbol{y})$:

$$d_{\mathcal{Y}}(g^*(\boldsymbol{x}'), \boldsymbol{y}) \leq d_{\mathcal{Y}}(\boldsymbol{y}, \boldsymbol{y}') + d_{\mathcal{Y}}(g^*(\boldsymbol{x}'), \boldsymbol{y}').$$

Putting them back together:

$$\begin{aligned} &d_{\mathcal{Y}}(g^*(\boldsymbol{x}), g^*(\boldsymbol{x}')) \leq \\ &\qquad d_{\mathcal{Y}}(g^*(\boldsymbol{x}), \boldsymbol{y}) + d_{\mathcal{Y}}(\boldsymbol{y}, \boldsymbol{y}') + d_{\mathcal{Y}}(g^*(\boldsymbol{x}'), \boldsymbol{y}'). \end{aligned}$$

Subtracting $d_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}')$ from both sides, we have:

$$\begin{aligned} &d_{\mathcal{Y}}(g^*(\boldsymbol{x}), g^*(\boldsymbol{x}')) - d_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}') \\ &\leq d_{\mathcal{Y}}(g^*(\boldsymbol{x}), \boldsymbol{y}) + d_{\mathcal{Y}}(\boldsymbol{y}, \boldsymbol{y}') + d_{\mathcal{Y}}(g^*(\boldsymbol{x}'), \boldsymbol{y}') - d_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}'). \end{aligned}$$

Let π^*_∞ be the underlying optimal coupling for the GW_∞, then by Definition 2 we can substitute the GW_∞ into the inequality and obtain:

$$\begin{aligned} &d_{\mathcal{Y}}(g^*(\boldsymbol{x}), g^*(\boldsymbol{x}')) - d_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}') \\ &\qquad \leq \mathrm{GW}_\infty + d_{\mathcal{Y}}(g^*(\boldsymbol{x}), \boldsymbol{y}) + d_{\mathcal{Y}}(g^*(\boldsymbol{x}'), \boldsymbol{y}') \\ &\qquad \leq \mathrm{GW}_\infty + 2\rho^*_{\pi_\infty}, \quad \text{(Definition 3)} \end{aligned}$$

which proves case 1.

Case 2. Proof of case 2 is similar to case 1, we start by applying triangle inequality on $d_{\mathcal{Y}}(\boldsymbol{y}, \boldsymbol{y}')$ and obtain:

$$d_{\mathcal{Y}}(\boldsymbol{y}, \boldsymbol{y}') \leq d_{\mathcal{Y}}(\boldsymbol{y}, g^*(\boldsymbol{x})) + d_{\mathcal{Y}}(g^*(\boldsymbol{x}), \boldsymbol{y}').$$

Applying triangle inequality again on $d_{\mathcal{Y}}(g^*(\boldsymbol{x}), \boldsymbol{y}')$:

$$d_{\mathcal{Y}}(g^*(\boldsymbol{x}), \boldsymbol{y}') \leq d_{\mathcal{Y}}(g^*(\boldsymbol{x}), g^*(\boldsymbol{x}')) + d_{\mathcal{Y}}(g^*(\boldsymbol{x}'), \boldsymbol{y}').$$

Putting them back together:

$$\begin{aligned} &d_{\mathcal{Y}}(\boldsymbol{y}, \boldsymbol{y}') \leq \\ &\qquad d_{\mathcal{Y}}(\boldsymbol{y}, g^*(\boldsymbol{x})) + d_{\mathcal{Y}}(g^*(\boldsymbol{x}), g^*(\boldsymbol{x}')) + d_{\mathcal{Y}}(g^*(\boldsymbol{x}'), \boldsymbol{y}'). \end{aligned}$$

Apply Definition 2 similarly:

$$\begin{aligned} d_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}') \leq \mathrm{GW}_\infty &+ d_{\mathcal{Y}}(\boldsymbol{y}, g^*(\boldsymbol{x})) \\ &+ d_{\mathcal{Y}}(g^*(\boldsymbol{x}), g^*(\boldsymbol{x}')) + d_{\mathcal{Y}}(g^*(\boldsymbol{x}'), \boldsymbol{y}'). \end{aligned}$$

Subtracting $d_{\mathcal{Y}}(g^*(\boldsymbol{x}), g^*(\boldsymbol{x}'))$ from both sides, we have

$$\begin{aligned} &d_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}') - d_{\mathcal{Y}}(g^*(\boldsymbol{x}), g^*(\boldsymbol{x}')) \\ &\qquad \leq d_{\mathcal{Y}}(\boldsymbol{y}, g^*(\boldsymbol{x})) + d_{\mathcal{Y}}(g^*(\boldsymbol{x}'), \boldsymbol{y}') + \mathrm{GW}_\infty, \\ &\qquad \leq \mathrm{GW}_\infty + 2\rho^*_{\pi_\infty}, \end{aligned}$$

which completes the proof. □

8.3. Proof of Theorem 1

Proof. In Lemma 1, we have shown that the deviation of pairwise distance, caused by a mapping function g^* from space $\mathcal{X}$ to space $\mathcal{Y}$ can be bounded by the ∞-norm GW distance and the supremum of the mapping error of g^*_π. This can be intuitively translated to a bound on the Lipschitz-constant of the function. That is, by Lemma 1 we have

$$\frac{d_{\mathcal{Y}}(g^*(\boldsymbol{x}), g^*(\boldsymbol{x}'))}{d_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}')} \leq \frac{2\rho^*_{\pi_\infty} + \mathrm{GW}_\infty}{d_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}')} + 1.$$

By the definition of Lipschitz-continuity, $\sup_{x \neq x'} \frac{d_{\mathcal{Y}}(g^*(\boldsymbol{x}), g^*(\boldsymbol{x}'))}{d_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}')}$ is exactly the Lipschitz constant of g^*. Taking the supremum over all pairs $x \neq x'$ on both sides and noting that $d_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}') \geq r$ yields the upper bound on L_{g^*}:

$$\begin{aligned} L_{g^*} &= \sup_{x \neq x'} \frac{d_{\mathcal{Y}}(g^*(\boldsymbol{x}), g^*(\boldsymbol{x}'))}{d_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}')} \\ &\leq \sup_{x \neq x'} \left(1 + \frac{2\rho^*_{\pi_\infty} + \mathrm{GW}_\infty}{d_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}')}\right) \\ &= 1 + \frac{2\rho^*_{\pi_\infty} + \mathrm{GW}_\infty}{r}. \end{aligned}$$

□

9. Related Works

9.1. Performance Prediction for Model Selection.

Due to the prohibitive cost for full-training every single candidate large models, performance prediction, or transferability estimation, has become an increasing attentive area. However, most of the works in this domain focus on classification tasks only [5, 11, 18, 35, 43, 46, 47, 51, 56], and are hard to naively migrated to generative tasks. In addition, the problem being studied in this paper differs fundamentally from existing works in transferability estimation in the following sense: existing works mostly consider a *task-centric* perspective, that is, given the joint distribution of a target task, how do we evaluate the fitness of the model without actually fine-tuning them. Whereas the scope of this study is focus on the interplay of multi-modal representations, and how the choice of pre-train models influence it. Another recent work [55] considers the similar notion of assessing the quality of vision encoder by examining their clustering quality.

9.2. The Platonic Representation Hypothesis (full)

However, it is crucial to note that MutualNN has several drawbacks compared to GW, particularly in the context of model selection: (i) it suffers from sensitivity to an additional hyper-parameter (top-k neighbors). Since the task is training-free model selection, it is not feasible to optimize this hyper-parameter in hindsight based on post-VLM performance; (ii) MutualNN does not enjoy the theoretical properties of GW distance. While we can intuitively understand that MutualNN "approximates" the geometrical similarity between spaces, how to translate that into a formal guarantee, and how that influences the learnability of mappings across spaces remains dubious; (iii) akin to RSA, MutualNN enforces a "hard" correspondence between known image-text pairs. While such an assumption is reasonable (or even superior) when semantic information overlaps perfectly, it fails when modalities are complementary. If the semantic content of an image and text pair is distinct but related (non-overlapping), MutualNN incorrectly assumes the corresponding objects are affine and fails to capture the relationship. We hypothesize this is exactly why the original authors chose a relatively small k: they implicitly assume that for the most semantically aligned sub-regions, information overlaps rather than complements. In contrast, because GW learns the correspondence via the coupling matrix, it is inherently more robust to such complementary scenarios (as such correspondence will be down-weighted).

10. Experimental Details

We summarize the key training configurations of pre-training feature alignment (stage 1) and visual instruction tuning (stage 2) in Table 8.

10.1. Training Configurations

Configuration	Stage-1	Stage-2
learning rate	$2e-3$	$2e-5$
learning scheduler	cosine	cosine
warmup ratio	0.03	0.03
global batch size	256	128
training epoch	1	1
max sequence length	2048	2048

Table 8. LLaVa-1.5 training configuration.

10.2. Details of Model Pool

Provider	Source	Architecture	Size	ImgNet-1k
"Large" group: vision encoders size larger than ViT-L				
Laion	OpenCLIP [8]	`ViT-bigG-14`	2540M	80.09
BAAI	MLCD [2]	`ViT-bigG-14`	1842M	-
Laion	OpenCLIP [8]	`ViT-g-14`	1367M	78.47
Meta	DINO-v2 [36]	`giant`	1136M	-
Apple	DFN [12]	`ViT-H-14`	987M	84.37
Meta	CLIP [41]	`ViT-H-14`	986M	80.51
Laion	OpenCLIP [8]	`ViT-H-14`	986M	77.96
Google	SigLIP [53]	`SoViT-400m`	877M	83.08
"Small" group: vision encoders size smaller or equal to ViT-L				
Google	SigLIP [53]	`ViT-L-16`	652M	82.07
Apple	DFN [12]	`ViT-L-14`	428M	81.41
Laion	OpenCLIP [8]	`ViT-L-14`	428M	79.21
OpenAI	CLIP [41]	`ViT-L-14`	428M	76.56
Laion	OpenCLIP [8]	`ViT-L-14`	428M	75.25
Meta	DINO-v2	`Large`	304M	-
Google	SigLIP [53]	`ViT-B-16`	203M	78.49
OpenAI	CLIP [41]	`ViT-B-16`	150M	68.34
Meta	MetaCLIP [49]	`ViT-B-16`	149M	72.12
Meta	DINO-v2 [36]	`base`	87M	-

Table 9. Collection of vision-encoders in our experiments.

We summarized all vision encoders used in this study in Table 9, within each group, vision encoders are ordered by their size, models with similar size are ordered by accuracy. We try to make the collection as diverse as possible, covering models from different providers and different architectures. Any entry shown as "-" in Table 9 means zero-shot accuracy cannot be naively computed (DINO family), or there is no open reported zero-shot accuracy and paired text encoder is not found (MLCD).

10.3. Implementation Details of Baselines

RSA. For the implementation of Representational Similarity Analysis (RSA) [20], we use 1,000 randomly sampled image-text pairs to estimate the within-space pairwise similarity matrices. We employ cosine similarity as the comparison metric.

CCA. We implement Canonical Correlation Analysis (CCA) [33] using scikit-learn [38]. We increase the sample size to 5,000 pairs to ensure it exceeds the maximum feature dimension of either modality. We set the number of components to 10, as higher values yielded no significant changes in model ranking. The optimization is set to a maximum of 500 iterations with a tolerance of 10^{-6}, following scikit-learn defaults.

MutualNN. For MutualNN [16], we use 1,000 randomly sampled pairs and set the number of neighbors to $k = 10$, following the official implementation. To align with our GW distance setup, we treat the image modality as the source domain, calculating overlap based on the nearest neighbors with respect to the image features.