The Inter-Intra Modal Measure: A Predictive Lens on Fine-Tuning Outcomes in Vision-Language Models

Supplementary Material

6. Theorem 1 Proof

Proof. **Step 1. Representing IIMM as an Expectation**

The standard i.i.d. asumption on image embeddings yields $P_{XX} = P_X \otimes P_X$ and $Q_{XX} = Q_X \otimes Q_X$. By definition, we have

$$\mathrm{IIMM}(P) = \frac{1}{2}\Big(\underbrace{\mathbb{E}_{(x,y')\sim P_{XY}}[x^\top y']}_{A(P)} + \underbrace{\mathbb{E}_{(x,x')\sim P_{XX}}[x^\top x']}_{B(P)}\Big).$$

Similarly, with the perturbation,

$$\mathrm{IIMM}(Q) = \frac{1}{2}\Big(\mathbb{E}_{(x,y')\sim Q_{XY}}[x^\top y'] + \mathbb{E}_{(x,x')\sim Q_{XX}}[x^\top x']\Big).$$

Step 2. Lipschitz Continuity of the Inner Product

Define the function

$$f(x,y) = x^\top y.$$

For any two pairs (x,y) and (x',y') in $S^{d-1}\times S^{d-1}$, we have

$$\begin{aligned}\big|f(x,y) - f(x',y')\big| &= \big|x^\top y - x'^\top y'\big| \\ &\le \big|x^\top y - x'^\top y\big| + \big|x'^\top y - x'^\top y'\big| \\ &\le \|x-x'\|_2\,\|y\|_2 + \|x'\|_2\,\|y-y'\|_2 \\ &\le \|x-x'\|_2 + \|y-y'\|_2,\end{aligned}$$

because $\|x\|_2 = \|y\|_2 = \|x'\|_2 = 1$. Hence, f is 1–Lipschitz with respect to the metric

$$d\big((x,y),(x',y')\big) = \|x-x'\|_2 + \|y-y'\|_2.$$

Step 3. Application of Kantorovich–Rubinstein Duality

By the Kantorovich–Rubinstein duality, for any 1–Lipschitz function f and any two probability measures μ and ν, we have

$$\Big|\mathbb{E}_{z\sim\mu}[f(z)] - \mathbb{E}_{z\sim\nu}[f(z)]\Big| \le W_1(\mu,\nu).$$

Apply this result to the inter–modal term using $f(x,y) = x^\top y$. Thus,

$$\begin{aligned}&\Big|\mathbb{E}_{(x,y')\sim Q_{XY}}[x^\top y'] - \mathbb{E}_{(x,y')\sim P_{XY}}[x^\top y']\Big| \\ &\le W_1(P_{XY},Q_{XY}) \le \delta_A.\end{aligned}$$

Similarly, for the intra–modal term, define

$$g(x,x') = x^\top x'.$$

Using the same Lipschitz argument (with the metric $d((x,x'),(y,y')) = \|x-y\|_2 + \|x'-y'\|_2$) we obtain

$$\begin{aligned}&\Big|\mathbb{E}_{(x,x')\sim Q_{XX}}[x^\top x'] - \mathbb{E}_{(x,x')\sim P_{XX}}[x^\top x']\Big| \\ &\le W_1(P_{XX},Q_{XX}) \le 2W_1(P_X,Q_X) \le 2\delta_B.\end{aligned}$$

Step 4. Combining the Two Contributions

By the triangle inequality,

$$\begin{aligned}&\Big|\mathrm{IIMM}(Q) - \mathrm{IIMM}(P)\Big| \\ &= \frac{1}{2}\Big|\big[A(Q)+B(Q)\big] - \big[A(P)+B(P)\big]\Big| \\ &\le \frac{1}{2}\Big(\big|A(Q)-A(P)\big| + \big|B(Q)-B(P)\big|\Big) \\ &\le \frac{1}{2}\big(\delta_A + 2\delta_B\big).\end{aligned}$$

Step 5. Finite–Sample Concentration

In practice, one estimates the expectations in $A(P)$ and $B(P)$ from N independent samples. Since the inner product $x^\top y$ is bounded in $[-1,1]$, standard concentration inequalities (e.g., Hoeffding's inequality) imply that, with probability at least $1-\eta$, the empirical estimates $\widehat{A}(P)$ and $\widehat{B}(P)$ satisfy

$$\Big|\widehat{A}(P) - A(P)\Big| \le \epsilon_N \quad \text{and} \quad \Big|\widehat{B}(P) - B(P)\Big| \le \epsilon_N,$$

where

$$\epsilon_N = O\Big(\sqrt{\frac{\log(1/\eta)}{N}}\Big).$$

Hence, with the corresponding estimates $\widehat{\mathrm{IIMM}}(P)$ and $\widehat{\mathrm{IIMM}}(Q)$, we have

$$\Big|\widehat{\mathrm{IIMM}}(Q) - \widehat{\mathrm{IIMM}}(P)\Big| \le \frac{\delta_A + 2\delta_B}{2} + 2\epsilon_N,$$

where the constant 2 in front of ϵ_N can be absorbed in the big-O notation.

This completes the proof. □

7. Combining inter- and intra- measures

To determine how best to combine the inter- and intra- measures, we used the Pearson correlation, r_p, and its 95% CI of the measure determined by different convex combinations of the inter- and intra- measure and the gain over zero-shot error. We see in Table 7 and Table 8 that we do not have

enough power to make strong conclusions about the best coefficient value. We do see a pattern of smaller variance in the correlation estimate when combining the inter- and intra- terms compared to using either term in isolation.

We chose a coefficient of 0.5 for simplicity in further analysis.

8. Additional Results

To ensure completeness and reproducibility, Table 5 and Table 6 present the per-task zero-shot accuracy, fine-tuned accuracy, raw accuracy gain, and gain over zero-shot error for each base model and PEFT method, respectively. Kendall's tau scores, detailed in Table 1 and Table 2 of the main text, were calculated by first computing the transferability measures of interest for each dataset's zero-shot embeddings. These transferability measures, combined with the per-dataset gains over zero-shot error and raw accuracy gains, form paired observations across datasets. Kendall's tau was then computed using:

$$\tau = \frac{P - Q}{\sqrt{(P + Q + T)(P + Q + U)}}$$

Here, P and Q represent the number of concordant and discordant pairs, respectively. The formula considers all possible pairs of datasets, comparing their relative ordering in each list. Ties are corrected for by T and U, which denote the number of tied pairs in the first and second list, respectively.

Table 5. Performance metrics per model and dataset. ZS Acc: Zero-Shot Accuracy, FT Acc: Fine-Tuned Accuracy, Raw Gain: FT Acc - ZS Acc, Gain over ZS: Normalized Gain over Zero-Shot Error.

Model	Dataset	ZS Acc (%)	FT Acc (%)	Raw Gain (%)	Gain over ZS
CLIP	Cars	58.87	78.26	19.39	0.47
	DTD	42.07	74.31	32.23	0.56
	SVHN	27.27	96.00	68.73	0.95
	EuroSAT	44.07	97.85	53.78	0.96
	CIFAR100	61.71	85.67	23.96	0.63
	GTSRB	33.65	96.72	63.07	0.95
	MNIST	50.47	99.23	48.76	0.98
	RESISC45	56.56	92.10	35.54	0.82
	SUN397	61.34	75.95	14.61	0.38
	STL10	97.36	98.42	1.06	0.40
CoCa	Cars	83.85	90.10	6.26	0.39
	DTD	51.91	79.26	27.34	0.57
	SVHN	54.74	96.90	42.16	0.93
	EuroSAT	42.85	98.26	55.41	0.97
	CIFAR100	74.12	88.07	13.95	0.54
	GTSRB	42.39	98.30	55.91	0.97
	MNIST	69.04	99.45	30.41	0.98
	RESISC45	60.13	94.60	34.48	0.86
	SUN397	66.24	74.40	8.17	0.24
	STL10	96.24	98.06	1.83	0.49
EVA-02	Cars	78.80	88.55	9.75	0.46
	DTD	50.96	81.22	30.27	0.62
	SVHN	24.92	97.26	72.34	0.96
	EuroSAT	68.04	98.33	30.30	0.95
	CIFAR100	87.64	92.90	5.26	0.43
	GTSRB	46.64	97.78	51.14	0.96
	MNIST	44.21	99.63	55.42	0.99
	RESISC45	65.57	95.22	29.65	0.86
	SUN397	70.71	78.02	7.31	0.25
	STL10	99.48	99.51	0.04	0.07
SigLIP	Cars	90.80	94.64	3.84	0.42
	DTD	62.61	84.63	22.02	0.59
	SVHN	55.87	96.95	41.07	0.93
	EuroSAT	43.63	98.63	55.00	0.98
	CIFAR100	70.91	90.28	19.37	0.67
	GTSRB	52.84	98.85	46.01	0.98
	MNIST	83.52	99.62	16.10	0.98
	RESISC45	60.56	95.89	35.33	0.90
	SUN397	70.23	79.31	9.08	0.31
	STL10	98.19	99.18	0.99	0.54

Table 6. Performance metrics per PEFT method and dataset. ZS Acc: Zero-Shot Accuracy, FT Acc: Fine-Tuned Accuracy, Raw Gain: FT Acc - ZS Acc, Gain over ZS: Normalized Gain over Zero-Shot Error.

Model	Dataset	ZS Acc (%)	FT Acc (%)	Raw Gain (%)	Gain over ZS
Attention-Weight Tuning	Cars	58.87	81.26	22.39	0.54
	DTD	42.07	80.16	38.09	0.66
	SVHN	27.27	96.57	69.29	0.95
	EuroSAT	44.07	98.52	54.44	0.97
	CIFAR100	61.71	88.16	26.45	0.69
	GTSRB	33.65	98.15	64.50	0.97
	MNIST	50.47	99.56	49.09	0.99
	RESISC45	56.56	94.62	38.06	0.88
	SUN397	61.34	76.31	14.97	0.39
	STL10	97.36	98.44	1.08	0.41
BitFit	Cars	58.87	74.16	15.28	0.37
	DTD	42.07	70.43	28.35	0.49
	SVHN	27.27	95.15	67.88	0.93
	EuroSAT	44.07	98.04	53.96	0.96
	CIFAR100	61.71	83.90	22.19	0.58
	GTSRB	33.65	95.30	61.65	0.93
	MNIST	50.47	99.15	48.68	0.98
	RESISC45	56.56	90.37	33.81	0.78
	SUN397	61.34	73.68	12.34	0.32
	STL10	97.36	98.31	0.95	0.36
LoRA	Cars	58.87	68.23	9.35	0.23
	DTD	42.07	65.90	23.83	0.41
	SVHN	27.27	80.34	53.07	0.73
	EuroSAT	44.07	94.63	50.56	0.90
	CIFAR100	61.71	80.49	18.78	0.49
	GTSRB	33.65	91.39	57.74	0.87
	MNIST	50.47	97.89	47.42	0.96
	RESISC45	56.56	85.87	29.32	0.67
	SUN397	61.34	67.99	6.65	0.17
	STL10	97.36	95.03	-2.34	-0.89
CLIP-Adapter	Cars	58.87	62.45	3.58	0.09
	DTD	42.07	48.24	6.17	0.11
	SVHN	27.27	49.30	22.03	0.30
	EuroSAT	44.07	88.52	44.44	0.79
	CIFAR100	61.71	72.13	10.42	0.27
	GTSRB	33.65	67.32	33.67	0.51
	MNIST	50.47	94.26	43.79	0.88
	RESISC45	56.56	80.21	23.65	0.54
	SUN397	61.34	67.46	6.12	0.16
	STL10	97.36	95.03	-2.34	-0.89

Table 7. Pearson correlation and 95% confidence intervals of zero-shot measure and gain over zero-shot error for different values of α in convex combinations of the inter and intra measures, αintra $+ (1 - \alpha)$inter. Data from four base models trained over 9 tasks.

Model	CLIP		CoCa		EVA-02		SigLIP	
α	r_p	95% CI	r_p	95% CI	r_p	95% CI	r_p	95% CI
0.00	0.79	(0.26, 0.95)	0.91	(0.62, 0.98)	0.85	(0.43, 0.97)	0.89	(0.55, 0.98)
0.10	0.92	(0.66, 0.98)	0.95	(0.77, 0.99)	0.91	(0.62, 0.98)	0.94	(0.73, 0.99)
0.20	0.97	(0.86, 0.99)	0.96	(0.82, 0.99)	0.94	(0.73, 0.99)	0.96	(0.82, 0.99)
0.30	0.98	(0.9, 1.0)	0.97	(0.86, 0.99)	0.95	(0.77, 0.99)	0.97	(0.86, 0.99)
0.40	0.97	(0.86, 0.99)	0.97	(0.86, 0.99)	0.95	(0.77, 0.99)	0.96	(0.82, 0.99)
0.50	0.96	(0.82, 0.99)	0.96	(0.82, 0.99)	0.95	(0.77, 0.99)	0.96	(0.82, 0.99)
0.60	0.95	(0.77, 0.99)	0.96	(0.82, 0.99)	0.94	(0.73, 0.99)	0.96	(0.82, 0.99)
0.70	0.93	(0.7, 0.99)	0.95	(0.77, 0.99)	0.94	(0.73, 0.99)	0.95	(0.77, 0.99)
0.80	0.92	(0.66, 0.98)	0.95	(0.77, 0.99)	0.93	(0.7, 0.99)	0.95	(0.77, 0.99)
0.90	0.91	(0.62, 0.98)	0.94	(0.73, 0.99)	0.93	(0.7, 0.99)	0.94	(0.73, 0.99)
1.00	0.90	(0.59, 0.98)	0.94	(0.73, 0.99)	0.93	(0.7, 0.99)	0.94	(0.73, 0.99)

Table 8. Pearson correlation and 95% confidence intervals of zero-shot measure and gain over zero-shot error for different values of α in convex combinations of the inter and intra measures, αintra $+ (1 - \alpha)$inter. Data from PEFT methods trained over 9 tasks.

Method	Attention-WT		BitFit		LoRA		Adapter	
α	r_p	95% CI	r_p	95% CI	r_p	95% CI	r_p	95% CI
0.00	0.81	(0.32, 0.96)	0.78	(0.24, 0.95)	0.80	(0.29, 0.96)	0.61	(-0.09, 0.91)
0.10	0.93	(0.7, 0.99)	0.92	(0.66, 0.98)	0.92	(0.66, 0.98)	0.72	(0.11, 0.94)
0.20	0.98	(0.9, 1.0)	0.97	(0.86, 0.99)	0.97	(0.86, 0.99)	0.76	(0.19, 0.95)
0.30	0.98	(0.9, 1.0)	0.98	(0.9, 1.0)	0.98	(0.9, 1.0)	0.77	(0.22, 0.95)
0.40	0.97	(0.86, 0.99)	0.97	(0.86, 0.99)	0.97	(0.86, 0.99)	0.77	(0.22, 0.95)
0.50	0.96	(0.82, 0.99)	0.96	(0.82, 0.99)	0.95	(0.77, 0.99)	0.76	(0.19, 0.95)
0.60	0.94	(0.73, 0.99)	0.94	(0.73, 0.99)	0.94	(0.73, 0.99)	0.75	(0.17, 0.94)
0.70	0.93	(0.7, 0.99)	0.93	(0.7, 0.99)	0.93	(0.7, 0.99)	0.74	(0.15, 0.94)
0.80	0.92	(0.66, 0.98)	0.92	(0.66, 0.98)	0.91	(0.62, 0.98)	0.73	(0.13, 0.94)
0.90	0.90	(0.59, 0.98)	0.91	(0.62, 0.98)	0.90	(0.59, 0.98)	0.73	(0.13, 0.94)
1.00	0.89	(0.55, 0.98)	0.90	(0.59, 0.98)	0.89	(0.55, 0.98)	0.72	(0.11, 0.94)

Table 9. Linear fit of IIMM with different intra-modal measures to gain over zero-shot error following fine-tuning by pre-trained model (alpha = 0.5).*

Model	CoCa		EVA-02		CLIP		SigLIP	
Intra-Modal Measure	p-value	r_s	p-value	r_s	p-value	r_s	p-value	r_s
Intra-Images Distance	$\mathbf{< 10^{-3}}$	**0.95**	$\mathbf{< 10^{-3}}$	**0.93**	**0.001**	**0.91**	**0.002**	**0.88**
H-Score	0.020	0.75	0.077	0.62	0.002	0.88	0.020	0.75
TransRate	0.004	0.85	0.007	0.82	0.005	0.83	0.004	0.85
GBC	0.002	0.88	0.050	0.67	0.042	0.68	0.016	0.77