Visual Fourier Prompt Tuning

Runjia Zeng[1 *], Cheng Han[2 *], Qifan Wang[3], Chunshu Wu[4], Tong Geng[4], Lifu Huang[5], Ying Nian Wu[6] and Dongfang Liu[1†]

[1]Rochester Institute of Technology [2]University of Missouri - Kansas City
[3]Meta AI [4]University of Rochester
[5]Virginia Tech [6]University of California, Los Angeles

Content

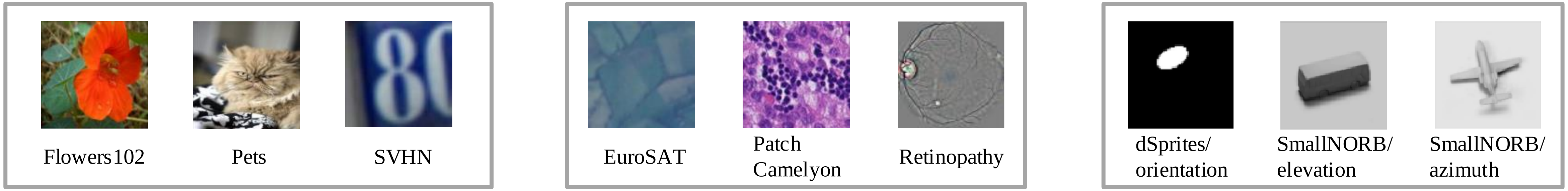
1. Introduction
2. Visual Fourier Prompt Tuning
3. Main Results
4. Study of Optimization
5. Study of Interpretability
6. Conclusion

Introduction

1. Introduction

- **Observation**

  A significant **performance degradation** occurs when there is a substantial **disparity** between the data used in pretraining and finetuning.



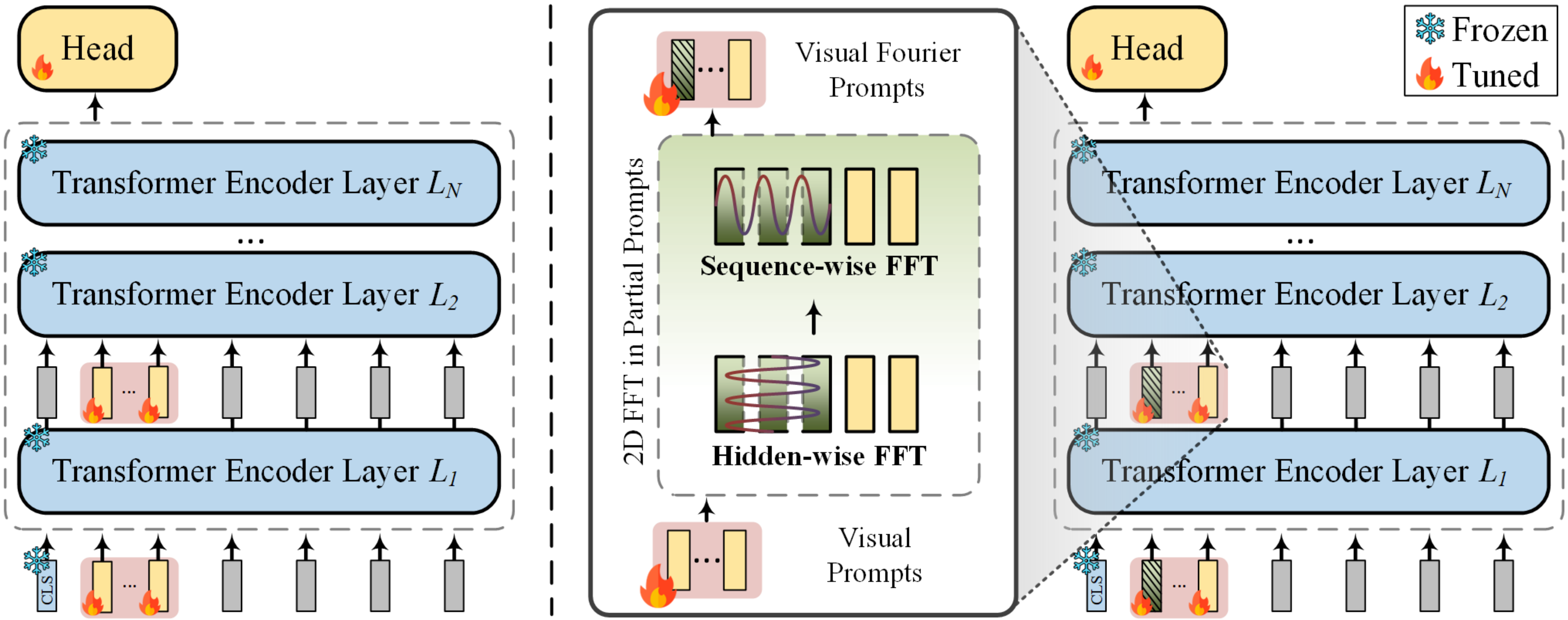
***Natural* <FID: 156.39>** ***Specialized* <FID: 245.69>** ***Structured* <FID: 234.96>**

- **Key Idea**

  Integrating **frequency domain** information into learnable prompt embeddings to elegantly assimilates data from both **spatial and frequency** domains.

Visual Fourier Prompt Tuning

2. Visual Fourier Prompt Tuning



(a) Visual Prompt Tuning (b) Fast Fourier Transform in Prompts (c) Visual Fourier Prompt Tuning

runjia.tech

Main Results

3. Main Results

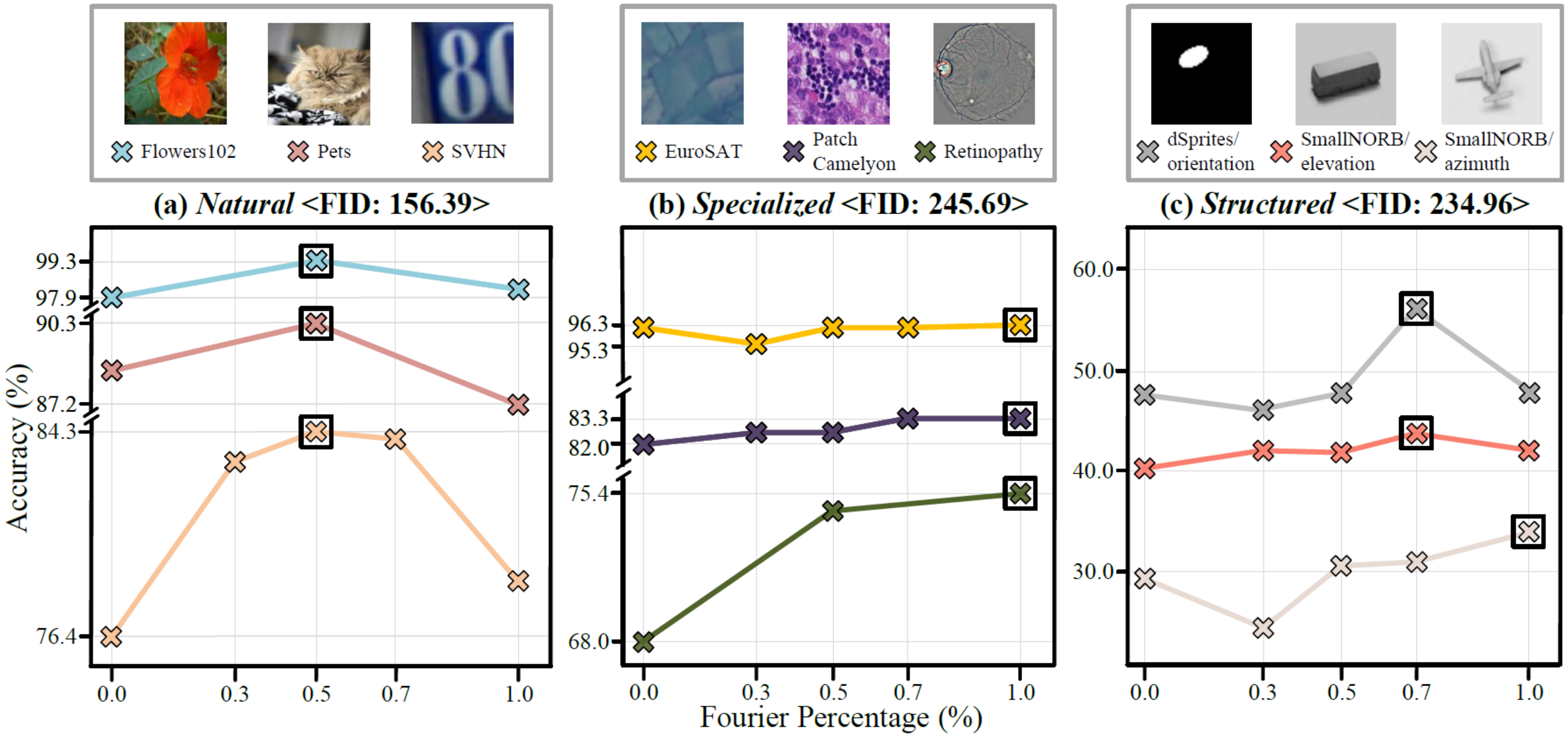
ViT-Base/16 [23] (85.8M)	Tuned/ Total	Scope Input	Scope Backbone	Extra params	FGVC [4] [5]	VTAB-1k [78] [19] *Natural* [7]	*Specialized* [4]	*Structured* [8]	Mean Total
Full [CVPR22][92]	100.00%		✓		88.54%	75.88%	83.36%	47.64%	65.57%
Linear [CVPR22][92]	0.08%				79.32% [0]	68.93% [1]	77.16% [1]	26.84% [0]	52.94%
Partial-1 [NeurIPS14][93]	8.34%				82.63% [0]	69.44% [2]	78.53% [0]	34.17% [0]	56.52%
MLP-3 [CVPR20][94]	1.44%			✓	79.80% [0]	67.80% [2]	72.83% [0]	30.62% [0]	53.21%
Sidetune [ECCV20][31]	10.08%		✓	✓	78.35% [0]	58.21% [0]	68.12% [0]	23.41% [0]	45.65%
Bias [NeurIPS17][30]	0.80%		✓		88.41% [3]	73.30% [3]	78.25% [0]	44.09% [2]	62.05%
Adapter [NeurIPS20][32]	1.02%		✓	✓	85.46% [1]	70.67% [4]	77.80% [0]	33.09% [0]	62.41%
LoRA [ICLR22][35]	—		✓	✓	89.46% [3]	78.26% [5]	83.78% [2]	56.20% [7]	72.25%
AdaptFormer [NeurIPS22][95]	—		✓	✓	—	<u>80.56%</u> [6]	84.88% [4]	<u>58.83%</u> [7]	<u>72.32%</u>
ARC_{att} [NeurIPS23][96]	—		✓	✓	89.12% [4]	80.41% [7]	**85.55%** [3]	58.38% [8]	72.32%
VPT-S [ECCV22][4]	0.16%	✓		✓	84.62% [1]	76.81% [4]	79.66% [0]	46.98% [4]	64.85%
VPT-D [ECCV22][4]	0.73%	✓		✓	89.11% [4]	78.48% [6]	82.43% [2]	54.98% [8]	69.43%
EXPRES [CVPR23][97]	—	✓		✓	—	79.69% [6]	84.03% [3]	54.99% [8]	70.20%
† E2VPT [ICCV23][5]	0.39%	✓	✓	✓	<u>89.22%</u> [4]	80.01% [6]	84.43% [3]	57.39% [8]	71.42%
▶ Ours	0.66%	✓		✓	**89.24%** [4] {4}	**81.35%** [6] {7}	<u>84.93%</u> [4] {4}	**60.19%** [8] {8}	**73.20%**

Swin-Base [24] (86.7M)	Tuned/ Total	VTAB-1k [78] [19] *Natural* [7]	*Specialized* [4]	*Structured* [8]
Full [ICLR23][98]	100.00%	79.10%	86.21%	59.65%
Linear [ICLR23][98]	0.06%	73.52% [5]	80.77% [0]	33.52% [0]
Partial-1 [NeurIPS14][93]	14.58%	73.11% [4]	81.70% [0]	34.96% [0]
MLP-3 [CVPR20][94]	2.42%	73.56% [5]	75.21% [0]	35.69% [0]
Bias [NeurIPS17][30]	0.29%	74.19% [2]	80.14% [0]	42.42% [0]
VPT [ECCV22][4]	0.25%	76.78% [6]	83.33% [0]	51.85% [0]
† E2VPT [ICCV23][5]	0.21%	<u>83.31%</u> [6]	<u>84.95%</u> [2]	<u>57.35%</u> [3]
▶ Ours	0.27%	**84.53%** [7] {5}	**86.15%** [2] {4}	**58.21%** [3] {6}

3. Main Results

Pretrained objectives / Methods	MAE [90] Tuned/ Total	MAE VTAB-1k [78] [19] *Natural* [7]	*Specialized* [4]	*Structured* [8]	MoCo v3 [26] Tuned/ Total	MoCo v3 VTAB-1k [78] [19] Natural [7]	Specialized [4]	Structured [8]
Full [CVPR22][92]	100.00%	59.31%	79.68%	53.82%	100.00%	71.95%	84.72%	51.98%
Linear [CVPR22][92]	0.04%	18.87% [0]	53.72% [0]	23.70% [0]	0.04%	67.46% [4]	81.08% [0]	30.33% [0]
Partial-1 [NeurIPS14][93]	8.30%	**58.44%** [5]	**78.28%** [1]	47.64%[1]	8.30%	72.31% [5]	84.58% [2]	47.89% [1]
Bias [NeurIPS17][30]	0.16%	54.55% [1]	75.68% [1]	**47.70%** [0]	0.16%	72.89% [3]	81.14% [0]	53.43% [4]
Adapter [NeurIPS20][32]	0.87%	54.90% [3]	75.19% [1]	38.98% [0]	1.12%	74.19% [4]	82.66% [1]	47.69% [2]
VPT-S [ECCV22][4]	0.05%	39.96% [1]	69.65% [0]	27.50% [0]	0.06%	67.34% [3]	82.26% [0]	37.55% [0]
VPT-D [ECCV22][4]	⋆ 0.31%	36.02% [0]	60.61% [1]	26.57% [0]	⋆ 0.22%	70.27% [4]	83.04% [0]	42.38% [0]
GPT [ICML23][101]	0.05%	47.61% [2]	76.86% [1]	36.80% [1]	0.06%	74.84% [4]	83.38% [1]	49.10% [3]
▶ Ours	0.38%	53.59% [6] {6}	77.75% [1] {3}	36.15% [1] {6}	0.22%	**77.47%** [5] {7}	**85.76%** [3] {4}	**58.74%** [6] {8}

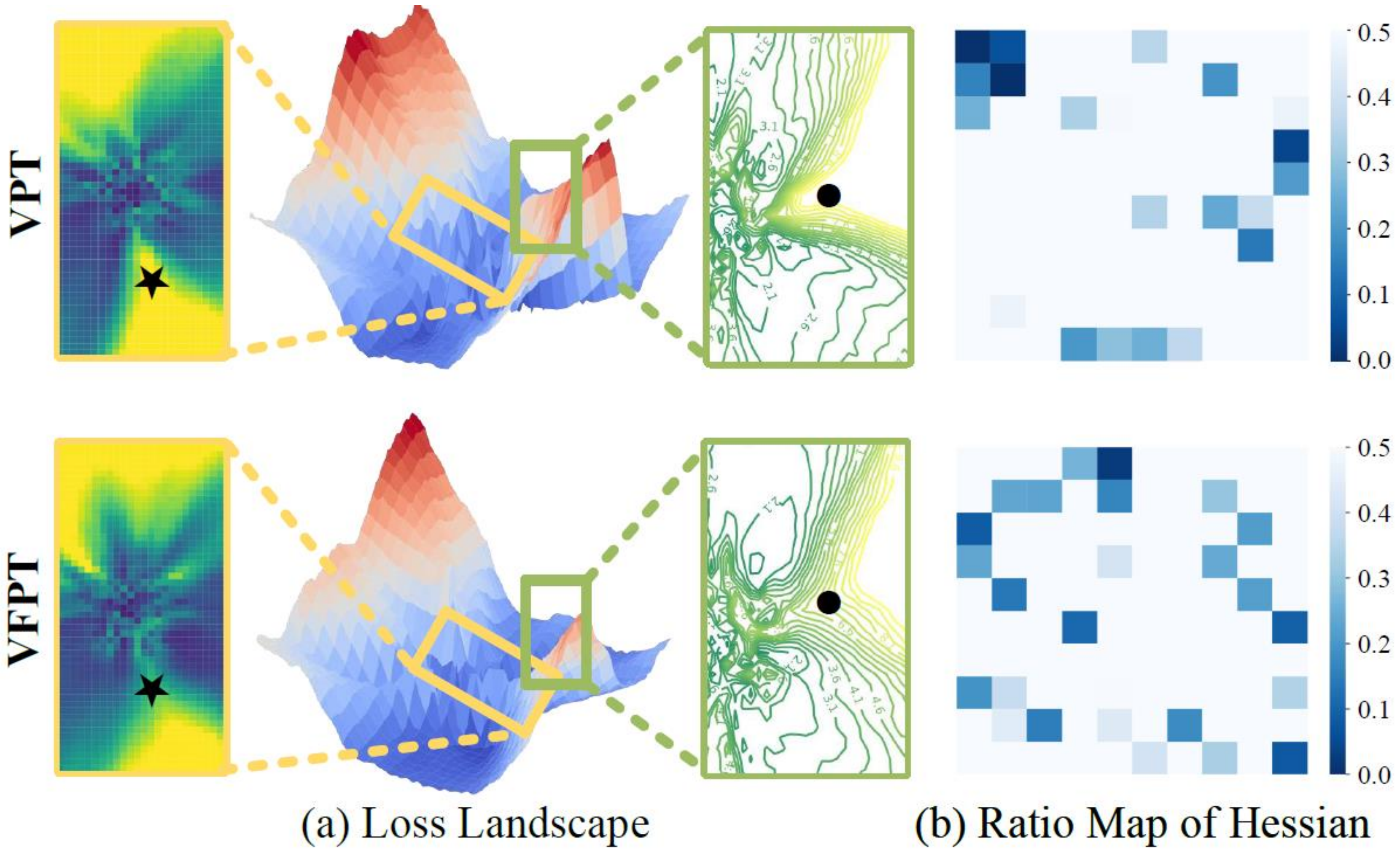
3. Main Results



runjia.tech

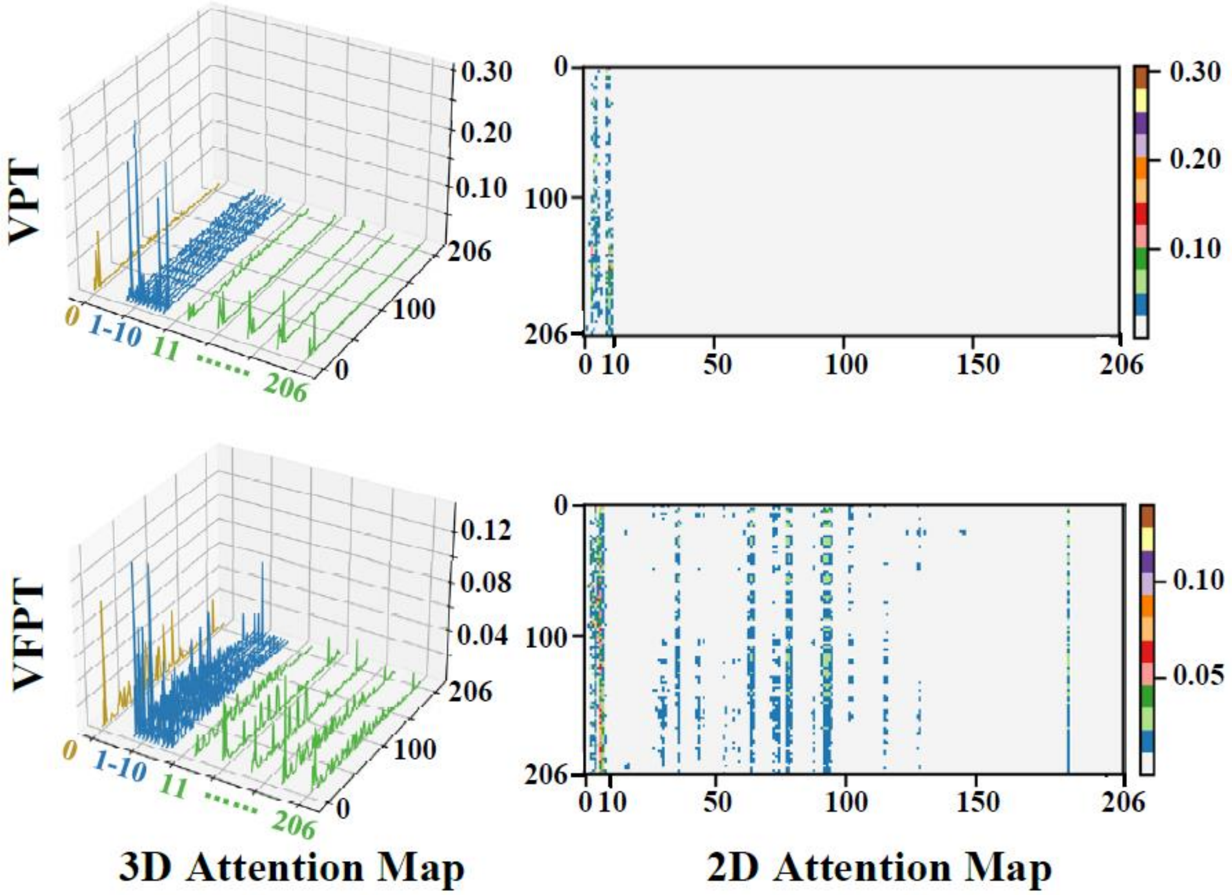
Study of Optimization

4. Study of Optimization

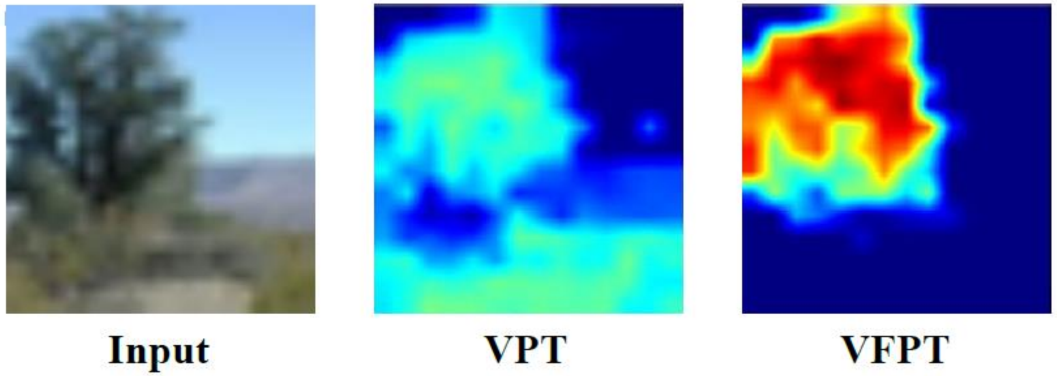


Study of Interpretability

5. Study of Interpretability



(a) Attention map



(b) GradCAM Heatmap

Conclusion

6. Conclusion

- **Simplicity**

  Integrating spatial and frequency domain information through an intuitive yet effective design.

- **Generality**

  Demonstrating generality across datasets with varying disparities while ensuring powerful performance.

- **Interpretability**

  Thoroughly investigating the associations between learnable prompts and frozen embeddings to elucidate our generality.

THANK YOU!