

ZIP:

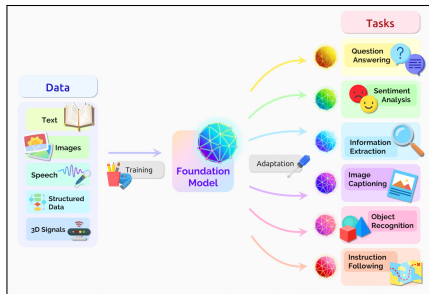
An Efficient Zeroth-order Prompt Tuning for Black-Box Vision-Language Models

Seonghwan Park¹ Jaehyeon Jeong¹ Yongjun Kim¹ Jaeho Lee^{1,2} Namhoon Lee^{1,2}

¹POSTECH

²Yonsei University

Foundation models are bringing about a paradigm shift in artificial intelligence.



Foundation model (Bommasani et al. 2021)

- ▶ a large machine learning model trained on a vast quantity of data at scale so that it can be adapted to a wide range of downstream tasks (e.g., BERT (Devlin et al. 2019), GPT (Brown et al. 2020), T5 (Raffel et al. 2020), LLaMA (Touvron et al. 2023), CLIP (Radford et al. 2021), DALL-E (Ramesh et al. 2021))

Image from NVIDIA Blog (<https://blogs.nvidia.com/blog/what-are-foundation-models/>)

Some foundation models are provided black-box APIs.



- ▶ Fine-tuning these models for specific downstream tasks can create more performance refinement (Liu et al. 2022).
- ▶ However, fine-tuning these models is not only computationally expensive, but also requires full access to model specifications.
- ▶ High-performing foundation models are provided only as a software-as-a-service (e.g., ChatGPT (OpenAI 2023) and Gemini (Google 2023)) without model details due to commercial interests and security concerns.

Black-box prompt tuning enables fine-tuning without access to the model details.

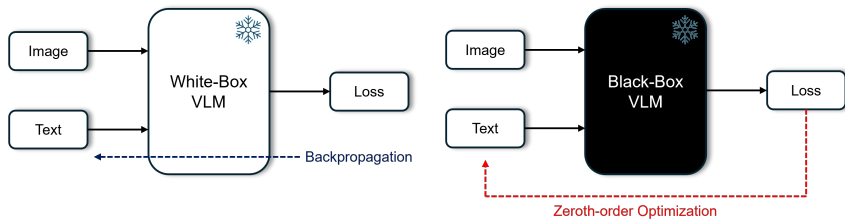


Figure: White-box optimization (left) and derivative-free optimization (right)

- ▶ Recent works have suggested to fine-tune such black-box models via so-called black-box prompt tuning (BBPT) (Sun et al. 2022; Oh et al. 2023; Yu et al. 2023).
- ▶ Black-box prompt tuning combines prompt tuning (Lester et al. 2021) with derivative-free optimization such as zeroth-order optimization (ZOO) (Ghadimi and Lan 2013), offering a viable solution.

Major challenge to BBPT

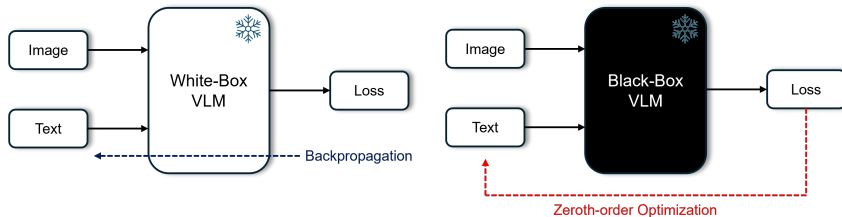


Figure: White-box optimization (left) and derivative-free optimization (right)

- ▶ Many existing BBPT approaches **require excessive model evaluations (i.e., queries)**, often spanning several tens of thousand times (Tsai et al. 2020; Oh et al. 2023).
- ▶ They result in significant performance drop when they are given a limited query budget.
- ▶ This is critical in many practical scenarios where large models are provided in the form of prediction APIs, and users can only make use of it with a limited budget.

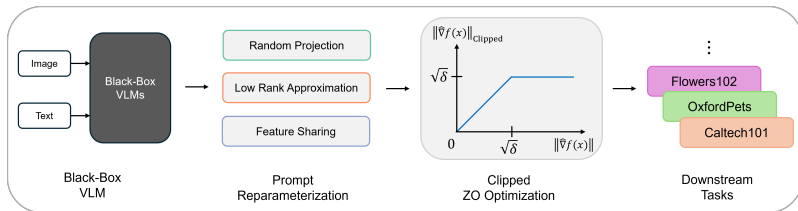
Limitations of ZO-SGD

$$\text{SPSA: } \hat{\nabla} f(\theta; \mathcal{B}) = \frac{1}{N} \sum_{i=1}^N \frac{f(\theta + cz_i; \mathcal{B}) - f(\theta - cz_i; \mathcal{B})}{2c} (z_i)^{-1}$$

$$\text{ZO-SGD: } \theta_{t+1} = \theta_t - \eta_t \hat{\nabla} f(\theta_t; \mathcal{B}_t)$$

- ▶ ZO-SGD is effective for BBPT but can suffer from high variance and slow convergence, especially for high-dimensional problems (Spall 1992; Ghadimi and Lan 2013; Duchi et al. 2015).
- ▶ High query requirements make BBPT infeasible in constrained environments like API-based models.
- ▶ Our key idea is to reduce the dimensionality of θ and the variance of $\hat{\nabla} f$.

ZIP: Zeroth-order Intrinsic-dimensional Prompt-tuning



Two key techniques address ZOO challenges:

1. Prompt Reparameterization

- ▶ Dimensionality Reduction: Utilize random projections and low-rank approximations to minimize the parameter space.
- ▶ Feature Sharing: Mitigate expressiveness loss caused by dimensionality reduction.

2. Clipped ZO Optimization

- ▶ Gradient Clipping: Stabilize ZOO by reducing variance during training.
- ▶ Intrinsic-Dimensional Clipping: Automatically determine clipping thresholds based on problem dimensionality.

Dimensionality reduction

Random Projection

- ▶ Projects learnable parameters of each context tokens $\theta_i \in \mathbb{R}^p$ onto a lower-dimensional space $v_i \in \mathbb{R}^q$:

$$\theta_i = \theta_{0,i} + \mathbf{M}_i v_i$$

- ▶ \mathbf{M}_i : random projection matrix.
- ▶ The total number of parameters are reduced from $d = p \times m$ to $d' = q \times m$ with $d' \ll d$.

Dimensionality reduction

Low-Rank Reparameterization:

- ▶ Apply additional reparameterization with a low-rank approximation as below:

$$\mathbf{W} = [v'_1 | v'_2 | \cdots | v'_m] = \mathbf{U} \text{diag}(\mathbf{s}) \mathbf{V}^T$$

- ▶ $\mathbf{U} \in \mathbb{R}^{q \times r}$, $\mathbf{s} \in \mathbb{R}^r$, $\mathbf{V} \in \mathbb{R}^{r \times m}$: trainable parameters.
- ▶ The total number of parameters are reduced from $d' = q \times m$ to $d'' = r \times (q + m + 1)$.

Feature sharing

While reducing the number of parameters can accelerate training speed with zeroth-order, it may also reduce the model expressive power.

Feature Sharing:

- ▶ Incorporate a vector $\mathbf{u} \in \mathbb{R}^q$ within \mathbf{W} , which can serve as a common base across the partitioned vectors as follows:

$$\begin{aligned}\Xi &= \mathbf{W} + \mathbf{u} \otimes \mathbf{1} \\ &= \mathbf{U} \text{diag}(\mathbf{s}) \mathbf{V}^T + \mathbf{u} \otimes \mathbf{1} \\ &= [\mathbf{w}_1 | \mathbf{w}_2 | \cdots | \mathbf{w}_m]\end{aligned}$$

- ▶ Ξ : final trainable parameter matrix.
- ▶ The total number of parameters are slightly increased from $d'' = r \times (q + m + 1)$ to $\delta = r \times (q + m + 1) + q$.
- ▶ Then, the updated parameters for context tokens are computed as:

$$\theta_i = \theta_{0,i} + \mathbf{M}_i \mathbf{w}_i.$$

Intrinsic-dimensional clipping

We obtain the final trainable parameter matrix Ξ in which there are $\delta = r \times (q + m + 1) + q$ parameters in total. Now the problem reduces to:

$$\min_{\Xi} f(\Xi, \omega; \mathcal{D}).$$

- ▶ One can consider employing ZO-SGD (Ghadimi and Lan 2013) to solve this problem, but it can still cause slow convergence in practice due to its large variance.

Intrinsic-dimensional clipping

To address this issue, we propose a simple yet robust zeroth-order method based on **intrinsic-dimensional clipping** defined as follows:

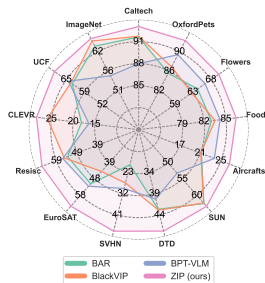
$$\Xi_{t+1} = \Xi_t - \eta_t \alpha_t \widehat{\nabla} f(\Xi_t, \omega; \mathcal{B}),$$

where α_t is a scaling factor defined as follows:

$$\alpha_t = \min \left(\frac{\sqrt{\delta}}{\sqrt{\sum_{i=1}^{\delta} \widehat{\nabla} f(\Xi_t, \omega; \mathcal{B})_i^2}}, 1 \right)$$

- ▶ Clips the zeroth-order stochastic gradient estimates $\widehat{\nabla} f$ if its norm exceeds $\sqrt{\delta}$ as a threshold.
- ▶ No need to manually select the clipping threshold (which is prone to be suboptimal) or perform an expensive hyperparameter search.
- ▶ Detailed proof is provided to validate the approach.

Result 1: ZIP performs the best on diverse vision-language tasks



Method	#Params	Caltech101	OxfordPets	Flowers102	Food101	FGVCAircraft	SUN397	DTD	SVHN	EuroSAT	Resisc45	CLEVR	UCF101	ImageNet	Average
Manual Prompt	0k	93.2	89.1	70.8	<u>85.9</u>	<u>24.8</u>	<u>62.6</u>	<u>44.1</u>	19.2	48.4	<u>57.2</u>	15.2	<u>67.5</u>	66.7	<u>57.3</u>
BAR	37.6k	92.5	85.6	65.0	83.0	21.6	62.4	42.9	19.8	51.6	53.9	18.1	63.5	64.0	55.7
BLACKVIP	9.9k	92.6	86.9	63.5	83.5	21.5	62.3	43.1	27.5	44.4	55.5	<u>25.9</u>	64.0	65.5	56.6
BPTVLM	4.0k	88.6	<u>89.4</u>	66.9	84.2	24.0	53.2	40.6	<u>29.8</u>	<u>53.0</u>	56.2	16.4	64.8	55.5	55.6
ZIP	0.4k	94.0	92.3	<u>70.4</u>	86.4	26.8	63.3	47.8	47.8	64.6	66.1	28.4	69.8	<u>66.2</u>	63.4

- ZIP consistently outperforms state-of-the-art BBPT methods across 11 out of 13 datasets.
- On average, ZIP achieves accuracy gains of +7.7%, +6.8%, and +7.8% over BAR, BlackVIP, and BPTVLM, respectively.
- ZIP shows remarkable performance in digit recognition, surpassing the next best method by +18.0% on the SVHN dataset.

Result 2: ZIP performs well with less function calls

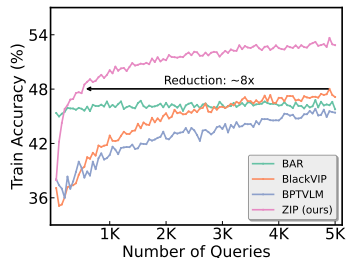
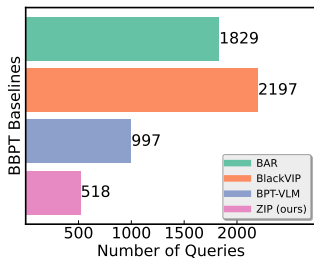


Figure: Query efficiency across 13 vision-language tasks

Figure: Training progress across 13 vision-language tasks

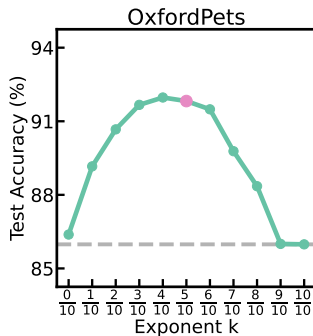
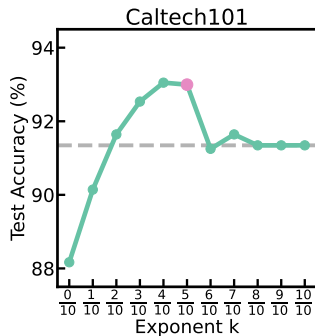
- ▶ ZIP significantly reduces the number of function calls required to reach target accuracy levels.
- ▶ ZIP achieves over eight times faster training compared to competing methods, demonstrating remarkable efficiency.

Result 3: ZIP generalizes well on unseen data

Method	Set	Caltech101	OxfordPets	Flowers102	Food101	FGVCAircraft	SUN397	DTD	SVHN	EuroSAT	Resisc45	CLEVR	UCF101	ImageNet	Average
BAR	Base	96.5	87.3	67.5	87.5	25.6	<u>69.2</u>	51.2	23.5	60.2	66.1	27.5	66.4	69.6	61.4
BLACKVIP		<u>96.6</u>	87.7	<u>67.9</u>	87.6	25.8	69.0	51.8	26.4	66.4	69.9	38.9	67.0	<u>70.3</u>	63.5
BPTVLM		93.2	<u>90.6</u>	66.9	<u>88.7</u>	<u>29.1</u>	65.3	<u>53.2</u>	<u>45.4</u>	<u>70.3</u>	<u>72.0</u>	<u>41.5</u>	<u>68.3</u>	66.3	<u>65.4</u>
ZIP		97.0	94.9	72.1	89.9	29.7	70.3	61.7	52.9	84.0	81.6	50.1	75.1	72.1	71.6
BAR	New	94.5	94.9	73.2	88.9	29.2	74.6	55.8	27.3	72.1	<u>62.3</u>	27.1	73.3	64.9	<u>64.5</u>
BLACKVIP		93.2	90.9	74.5	<u>89.4</u>	30.9	<u>73.9</u>	<u>55.4</u>	21.8	48.8	61.2	<u>28.0</u>	<u>72.6</u>	66.8	62.1
BPTVLM		92.7	<u>95.8</u>	72.7	85.4	<u>32.3</u>	64.8	45.3	<u>40.1</u>	47.0	61.3	28.4	65.0	55.2	60.5
ZIP		<u>94.3</u>	97.0	<u>73.4</u>	90.0	32.9	71.5	51.0	45.8	<u>64.4</u>	65.2	26.8	69.5	<u>65.6</u>	65.2
BAR	Harmonic	<u>95.5</u>	90.9	70.2	88.2	27.3	71.8	53.4	25.3	<u>65.6</u>	64.1	27.3	67.7	67.2	<u>62.9</u>
BLACKVIP		94.9	89.3	<u>71.0</u>	88.5	28.1	<u>71.4</u>	<u>53.5</u>	23.9	56.3	65.3	32.6	<u>69.7</u>	<u>68.5</u>	62.8
BPTVLM		92.9	<u>93.1</u>	69.7	87.0	<u>30.6</u>	65.0	48.9	<u>42.6</u>	56.3	<u>66.2</u>	<u>33.7</u>	66.6	60.2	62.9
ZIP		95.6	95.9	72.7	89.9	31.2	70.9	55.8	49.1	72.9	72.5	34.9	72.2	68.7	67.9

- ZIP consistently outperforms all BBPT baselines, achieving the highest base, new, and harmonic mean scores.
- These results are attributed to reduced model capacity and rough gradient estimates, which mitigate overfitting to noisy outliers, enhancing generalization.

Ablations: Intrinsic dimensional clipping



- ▶ Automatically determined $\sqrt{\delta}$ (where δ is a problem dimensionality) clipping thresholds deliver near-optimal performance on Caltech101 and OxfordPets datasets.
- ▶ The gray dashed line (*i.e.*, no clipping) highlights the benefits of intrinsic dimensional clipping.
- ▶ Enables zeroth-order optimization without hyperparameter tuning, simplifying the process.

Conclusion

ZIP

1. Dimensionality Reduction

- ▶ Combines random projection and low-rank approximation to address query inefficiency.

2. Feature Sharing

- ▶ Improves model expressiveness while ensuring parameter efficiency.

3. Intrinsic-Dimensional Clipping

- ▶ Stabilizes training and accelerates optimization using a $\sqrt{\delta}$ threshold, eliminating manual hyperparameter tuning.










Key Results

- ▶ Outperforms state-of-the-art BBPT baselines in accuracy across 13+ tasks.
- ▶ Demonstrates strong robustness to unseen data.
- ▶ Achieves up to 48% improvement in query efficiency under practical constraints compared to leading BBPT methods.










Poster Session

- ▶ Thu 24 Apr 3:00 PM SGT - Poster Session 2

References I

-  Bommasani, Rishi et al. (2021). "On the opportunities and risks of foundation models". In: *arXiv preprint arXiv:2108.07258*.
-  Brown, Tom et al. (2020). "Language models are few-shot learners". In: *NeurIPS*.
-  Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL*.
-  Duchi, John C et al. (2015). "Optimal rates for zero-order convex optimization: The power of two function evaluations". In: *IEEE Transactions on Information Theory*.
-  Ghadimi, Saeed and Guanhui Lan (2013). "Stochastic first-and zeroth-order methods for nonconvex stochastic programming". In: *SIAM Journal on Optimization*.
-  Google (2023). "Gemini: a family of highly capable multimodal models". In: *arXiv preprint arXiv:2312.11805*.
-  Lester, Brian, Rami Al-Rfou, and Noah Constant (2021). "The Power of Scale for Parameter-Efficient Prompt Tuning". In: *EMNLP*.
-  Liu, Haokun et al. (2022). "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning". In: *NeurIPS*.
-  Oh, Changdae et al. (2023). "BlackVIP: Black-Box Visual Prompting for Robust Transfer Learning". In: *CVPR*.

References II

-  OpenAI (2023). *ChatGPT*. <https://openai.com/blog/chatgpt/>.
-  Radford, Alec et al. (2021). “Learning transferable visual models from natural language supervision”. In: *ICML*.
-  Raffel, Colin et al. (2020). “Exploring the limits of transfer learning with a unified text-to-text transformer”. In: *Journal of machine learning research* 21.140, pp. 1–67.
-  Ramesh, Aditya et al. (2021). “Zero-shot text-to-image generation”. In: *ICML*.
-  Spall, James C (1992). “Multivariate stochastic approximation using a simultaneous perturbation gradient approximation”. In: *IEEE transactions on automatic control*.
-  Sun, Tianxiang et al. (2022). “Black-box tuning for language-model-as-a-service”. In: *ICML*.
-  Touvron, Hugo et al. (2023). “Llama: Open and efficient foundation language models”. In: *arXiv preprint arXiv:2302.13971*.
-  Tsai, Yun-Yun, Pin-Yu Chen, and Tsung-Yi Ho (2020). “Transfer learning without knowing: Reprogramming black-box machine learning models with scarce data and limited resources”. In: *ICML*.
-  Yu, Lang et al. (2023). “Black-box prompt tuning for vision-language model as a service”. In: *IJCAI*.