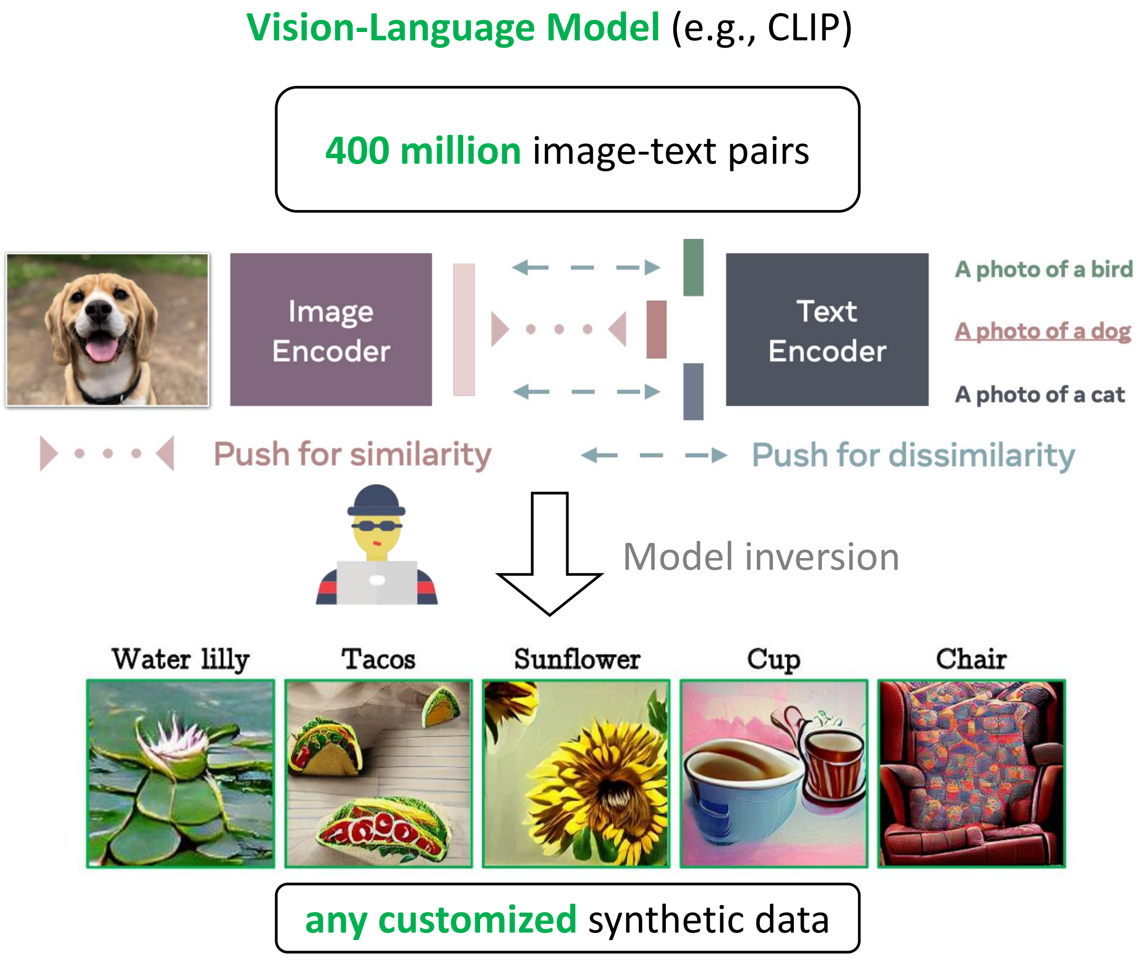
Open-Vocabulary Customization from CLIP via Data-Free Knowledge Distillation

Yongxian Wei[1], Zixuan Hu[2], Li Shen[3], Zhenyi Wang[4], Chun Yuan[1], Dacheng Tao[2]

1Tsinghua University; 2 Nanyang Technological University;
3 Shenzhen Campus of Sun Yat-sen University; 4 University of Maryland, College Park

Overview

Learning from CLIP Model (Data-Free Knowledge Distillation):



Vision-Language Model (e.g., CLIP)

400 million image-text pairs

any customized synthetic data

Goal:

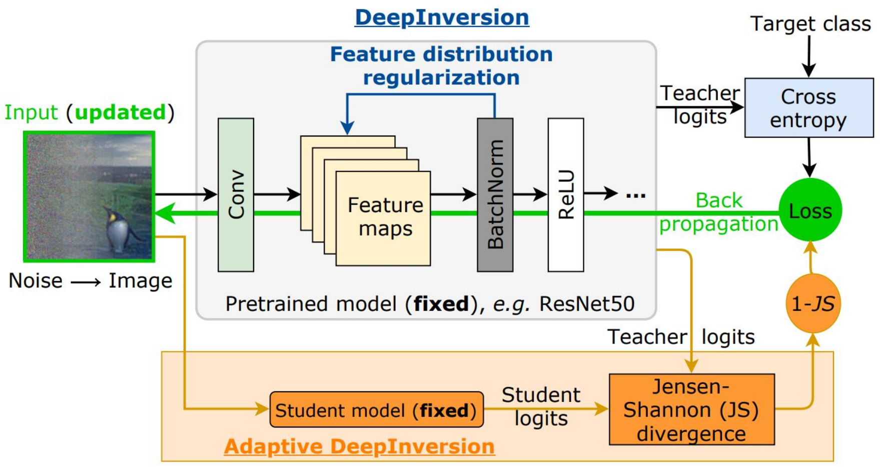
Vision-language models (e.g., CLIP) have demonstrated strong **zero-shot performance**, but their considerable size and inefficient inference limit **customizable deployment for users**. While KD is a solution, it still requires the original data, which is not always available due to copyrights and privacy concerns. For many users seeking open-vocabulary customization, Data-Free Knowledge Distillation (DFKD) emerges as a promising direction.

Challenges:

- **Data-free:** no access to the real training data
- **Lightweight:** the student model not only has a more lightweight visual backbone but also omits the text encoder entirely
- **Privacy-preserving:** no privacy leakage of original training data
- **Open-vocabulary:** different users have varying needs for downstream tasks (e.g., arbitrary combinations of class texts or few example images)

Preliminary

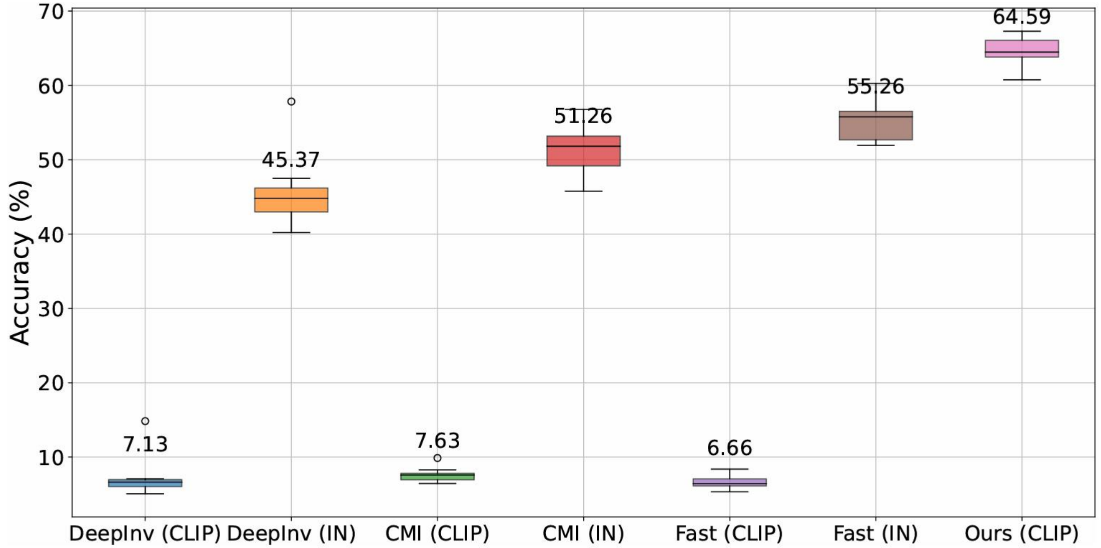
How do existing DFKD methods perform model inversion?



- Classification loss： $\min\limits_{\boldsymbol{Z},\boldsymbol{\theta}_G} \mathcal{L}_{cls}(\hat{\boldsymbol{X}}) = \frac{1}{|\hat{\boldsymbol{X}}|} \sum\limits_{(\hat{\boldsymbol{x}},y)\in(\hat{\boldsymbol{X}},\boldsymbol{Y})} l_{cls}(\hat{\boldsymbol{x}}, y), \text{ s.t. } \hat{\boldsymbol{X}} = \mathrm{G}(\boldsymbol{Z}; \boldsymbol{\theta}_{\mathrm{G}}).$
- Regularization loss： $\min\limits_{\hat{\boldsymbol{x}}} \mathcal{L}_{\mathrm{BN}} = \sum\limits_{l} \|\mu_l(\hat{\boldsymbol{x}}) - \mu_l^{\mathrm{BN}}\| + \|\sigma_l^2(\hat{\boldsymbol{x}}) - \sigma_l^{\mathrm{BN}}\|,$

Gradient backward propogation to optimize inputs instead of parameters

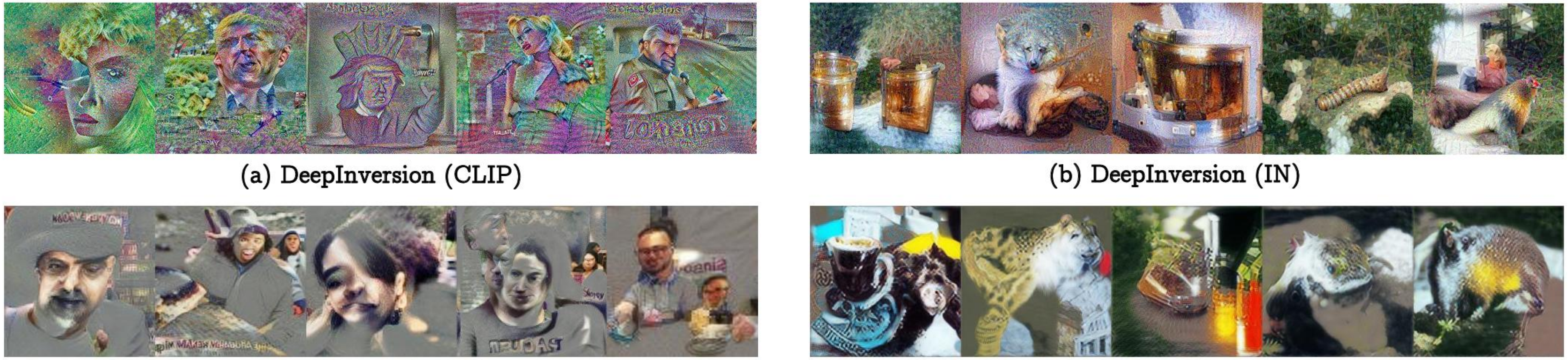
We utilize the ResNet-50 backbone of CLIP as the teacher model. CLIP has a visual encoder E_{img} to extract features. Consequently, we construct a linear classifier W upon the backbone. We then fine-tune this classifier using the testing set to form a classification model f(·) $= E_{img}(\cdot)^T W$.



[1] Hongxu Yin, et al. Dreaming to Distill: Data-free Knowledge Transfer via DeepInversion.

Finding

DFKD methods are only effective when the teacher’s BN stored distribution closely matches the testing distribution.



(a) DeepInversion (CLIP)

(b) DeepInversion (IN)

(c) CMI (CLIP)

(d) CMI (IN)

We observe that CLIP tends to encode facial features into statistics of its BatchNorm layers.

When using models pre-trained on ImageNet, these methods can synthesize informative images for training.

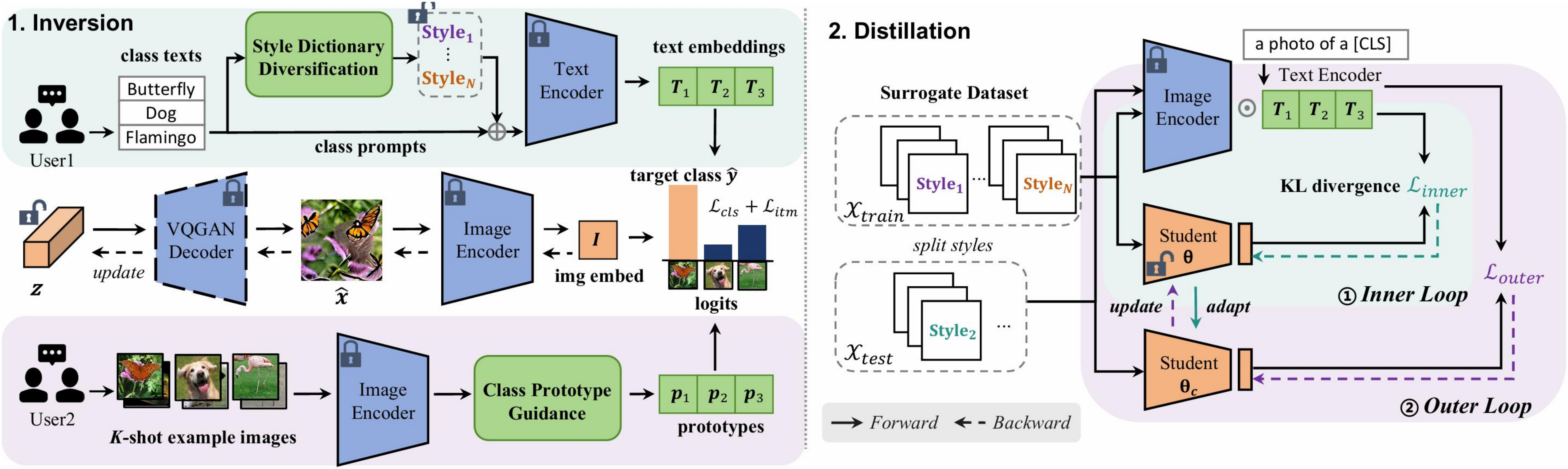
large-scale, web-crawled datasets invariably contain humans, even though the text descriptions may not mention people, leading to model bias.

[2] Gongfan Fang, et al. Contrastive Model Inversion for Data-free Knowledge Distillation.

[3] Gongfan Fang, et al. Up to 100× Faster Data-free Knowledge Distillation.

Methodology

Overall framework:



Methodology

Image-Text Matching:

1. Objective function

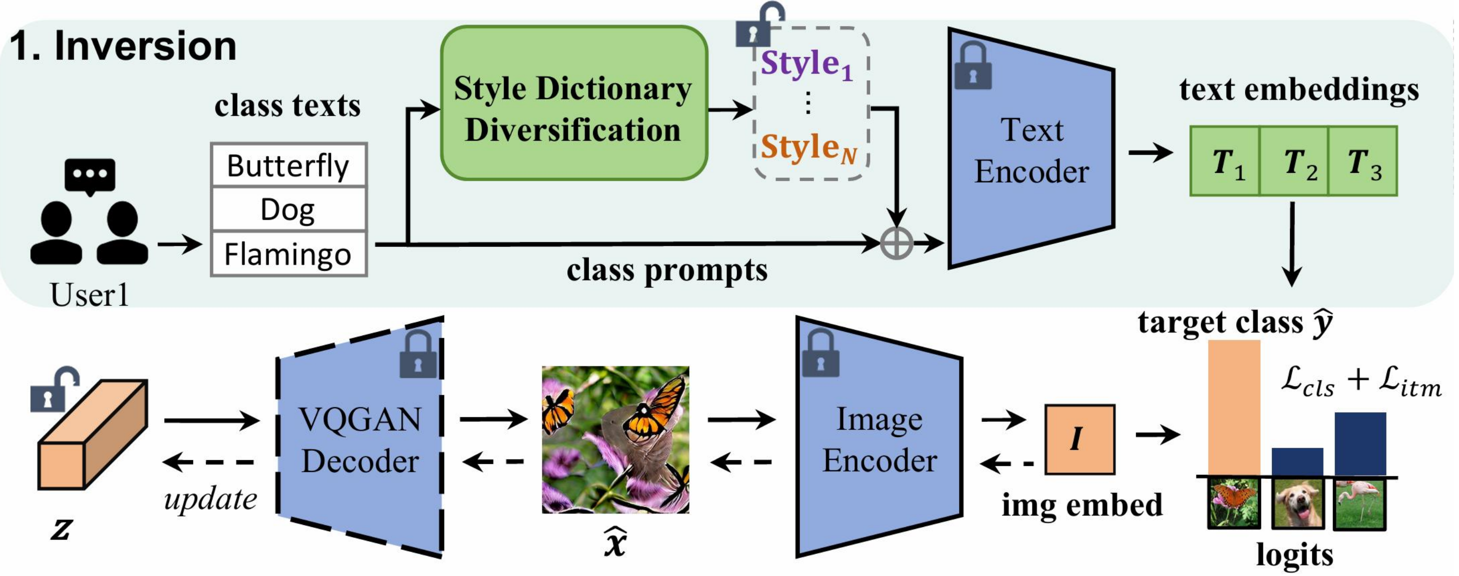
$$\min_{\boldsymbol{z}} \mathcal{L}_{itm} = \arcsin^2 \left(\| E_{\text{img}}(\mathcal{G}(\boldsymbol{z})) - E_{\text{txt}}(\boldsymbol{t}) \| \right)$$

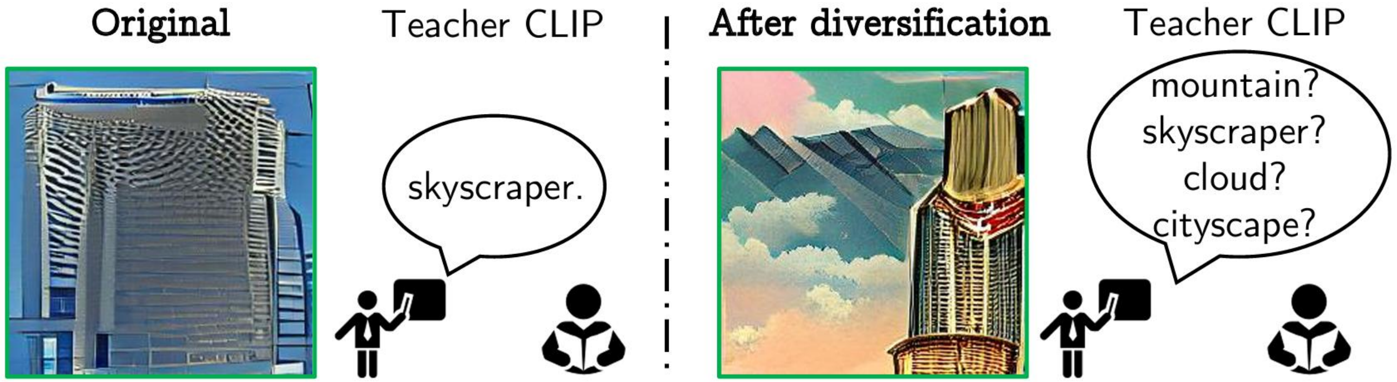
2. Style dictionary diversification

Prompt templates using a style dictionary {d1,··· ,dN}, such as “[t] in the style of [d]”

3. Class consistency maintaining

$$\min_{\boldsymbol{z}} \mathcal{L}_{cls} = CE(E_{\text{img}}(\mathcal{G}(\boldsymbol{z})) \cdot E_{\text{txt}}(\boldsymbol{s}), \hat{y})$$





Methodology

Meta knowledge distillation:

1. Objective function

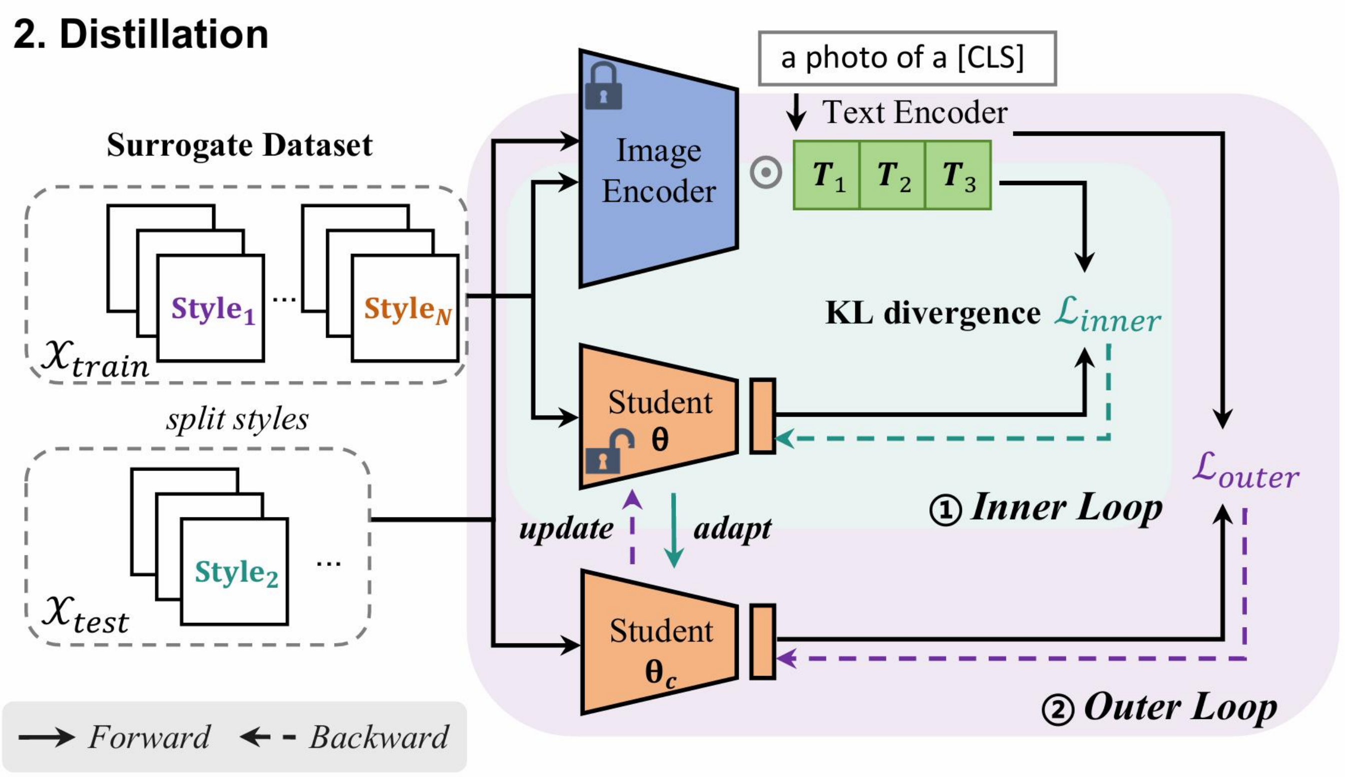
$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathcal{X}_{te}\sim\mathcal{X}} \mathcal{L}_{outer} = \mathcal{L}_{\mathrm{KD}}(\mathcal{X}_{tr};\boldsymbol{\theta}) + \mathcal{L}_{\mathrm{KD}}(\mathcal{X}_{te};\boldsymbol{\theta}_c),$$

$$\text{s.t. } \boldsymbol{\theta}_c = \min_{\boldsymbol{\theta}} \mathcal{L}_{inner} = \mathcal{L}_{\mathrm{KD}}(\mathcal{X}_{tr};\boldsymbol{\theta})$$

2. Gradient alignment

$$\nabla\mathcal{L}_{outer} = \nabla\mathcal{L}_{\mathrm{KD}}(\mathcal{X}_{tr};\boldsymbol{\theta}) + \nabla\mathcal{L}_{\mathrm{KD}}(\mathcal{X}_{te};\boldsymbol{\theta}) - \alpha\nabla\underbrace{\left(\nabla\mathcal{L}_{\mathrm{KD}}(\mathcal{X}_{tr};\boldsymbol{\theta})\cdot\nabla\mathcal{L}_{\mathrm{KD}}(\mathcal{X}_{te};\boldsymbol{\theta})\right)}_{Style\ Alignment} + \mathcal{O}(\alpha^2)$$

To learn invariant representations

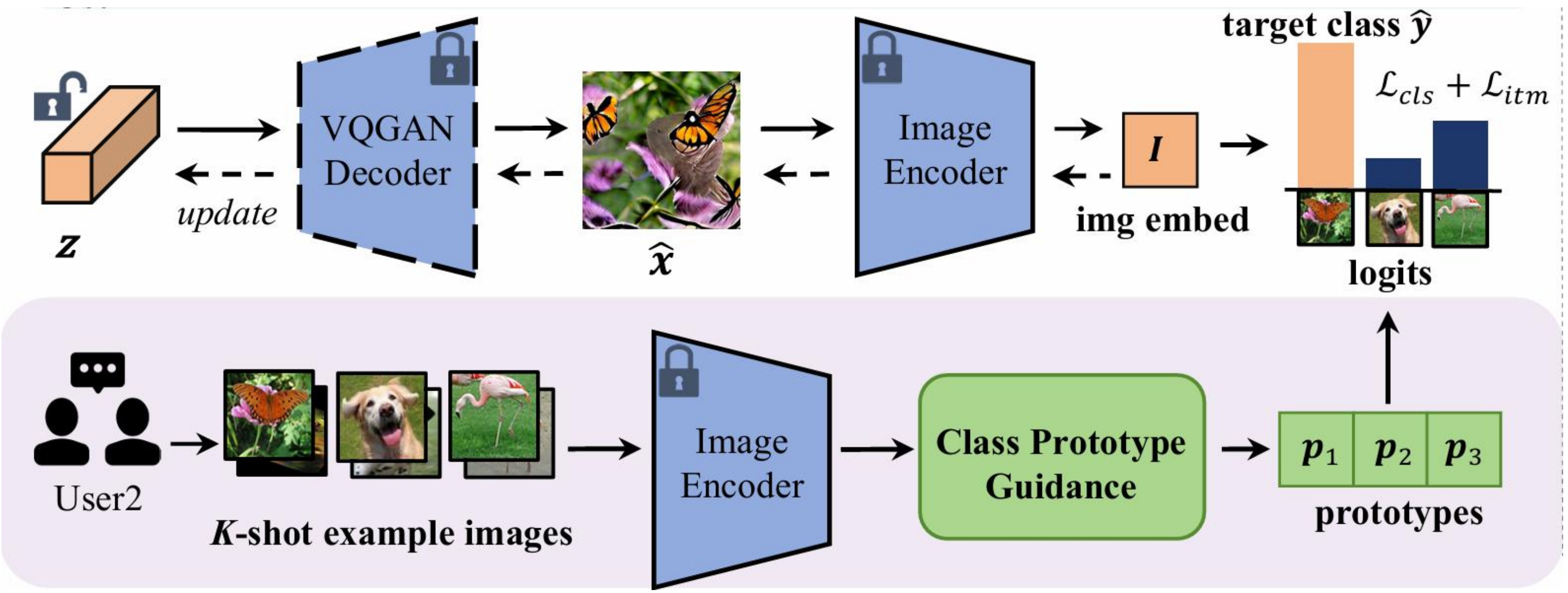


Methodology

Image-based customization:

1. Objective function

$$\min_{\boldsymbol{z}} \mathcal{L}_{iim} + \mathcal{L}_{cls} = \arcsin^2\left(\|E_{\text{img}}(\mathcal{G}(\boldsymbol{z})) - \boldsymbol{p}_{\hat{y}}\|\right) + CE(E_{\text{img}}(\mathcal{G}(\boldsymbol{z})) \cdot \boldsymbol{p}, \hat{y})$$



Experiment

Our method is an open-vocabulary, customized approach suitable for any category recognized by CLIP. Therefore, we randomly divide ImageNet-1K into 10 splits.

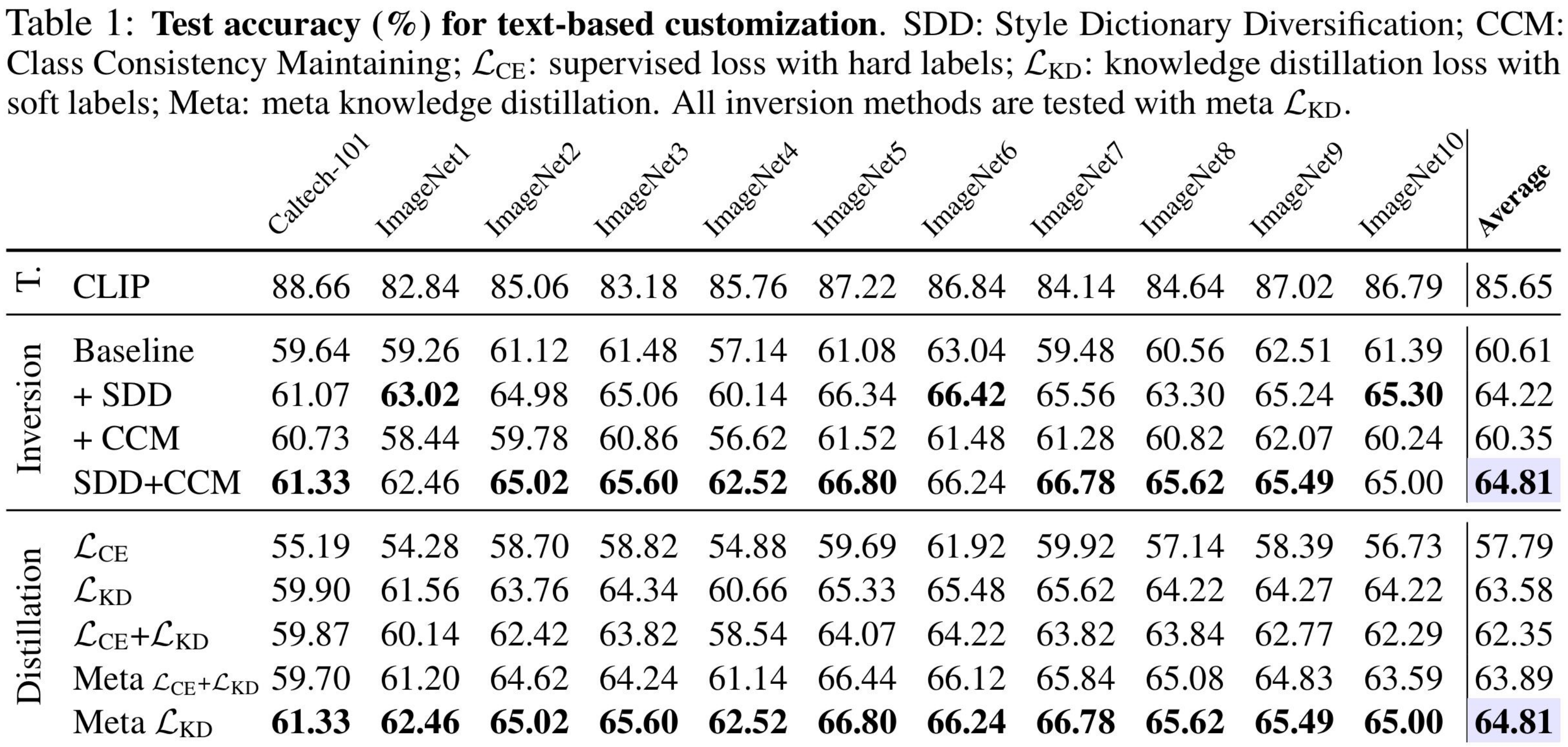
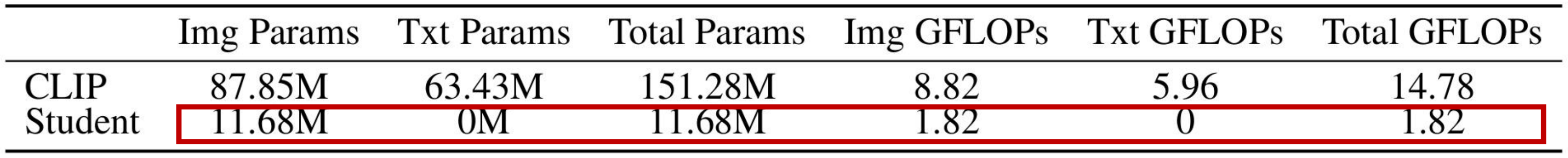


Table 1: **Test accuracy (%) for text-based customization**. SDD: Style Dictionary Diversification; CCM: Class Consistency Maintaining; $\mathcal{L}_{\text{CE}}$: supervised loss with hard labels; $\mathcal{L}_{\text{KD}}$: knowledge distillation loss with soft labels; Meta: meta knowledge distillation. All inversion methods are tested with meta $\mathcal{L}_{\text{KD}}$.

		Caltech-101	ImageNet1	ImageNet2	ImageNet3	ImageNet4	ImageNet5	ImageNet6	ImageNet7	ImageNet8	ImageNet9	ImageNet10	**Average**
T.	CLIP	88.66	82.84	85.06	83.18	85.76	87.22	86.84	84.14	84.64	87.02	86.79	85.65
Inversion	Baseline	59.64	59.26	61.12	61.48	57.14	61.08	63.04	59.48	60.56	62.51	61.39	60.61
	+ SDD	61.07	**63.02**	64.98	65.06	60.14	66.34	**66.42**	65.56	63.30	65.24	**65.30**	64.22
	+ CCM	60.73	58.44	59.78	60.86	56.62	61.52	61.48	61.28	60.82	62.07	60.24	60.35
	SDD+CCM	**61.33**	62.46	**65.02**	**65.60**	**62.52**	**66.80**	66.24	**66.78**	**65.62**	**65.49**	65.00	**64.81**
Distillation	$\mathcal{L}_{\text{CE}}$	55.19	54.28	58.70	58.82	54.88	59.69	61.92	59.92	57.14	58.39	56.73	57.79
	$\mathcal{L}_{\text{KD}}$	59.90	61.56	63.76	64.34	60.66	65.33	65.48	65.62	64.22	64.27	64.22	63.58
	$\mathcal{L}_{\text{CE}}$+$\mathcal{L}_{\text{KD}}$	59.87	60.14	62.42	63.82	58.54	64.07	64.22	63.82	63.84	62.77	62.29	62.35
	Meta $\mathcal{L}_{\text{CE}}$+$\mathcal{L}_{\text{KD}}$	59.70	61.20	64.62	64.24	61.14	66.44	66.12	65.84	65.08	64.83	63.59	63.89
	Meta $\mathcal{L}_{\text{KD}}$	**61.33**	**62.46**	**65.02**	**65.60**	**62.52**	**66.80**	**66.24**	**66.78**	**65.62**	**65.49**	**65.00**	**64.81**



	Img Params	Txt Params	Total Params	Img GFLOPs	Txt GFLOPs	Total GFLOPs
CLIP	87.85M	63.43M	151.28M	8.82	5.96	14.78
Student	11.68M	0M	11.68M	1.82	0	1.82

Visualization

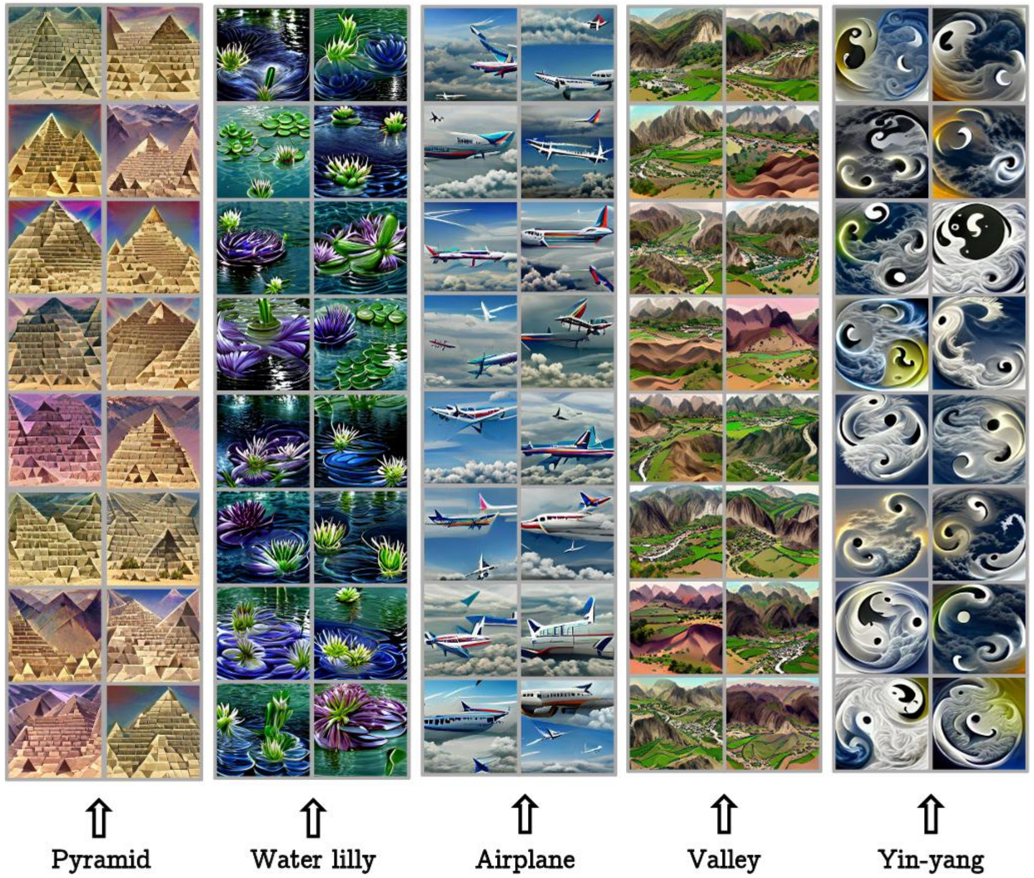


Figure 6: **Visualizations of the text-based customization.** Our approach enhances the diversity of synthesized images, preventing repetitive content while ensuring accurate representation of class semantics.

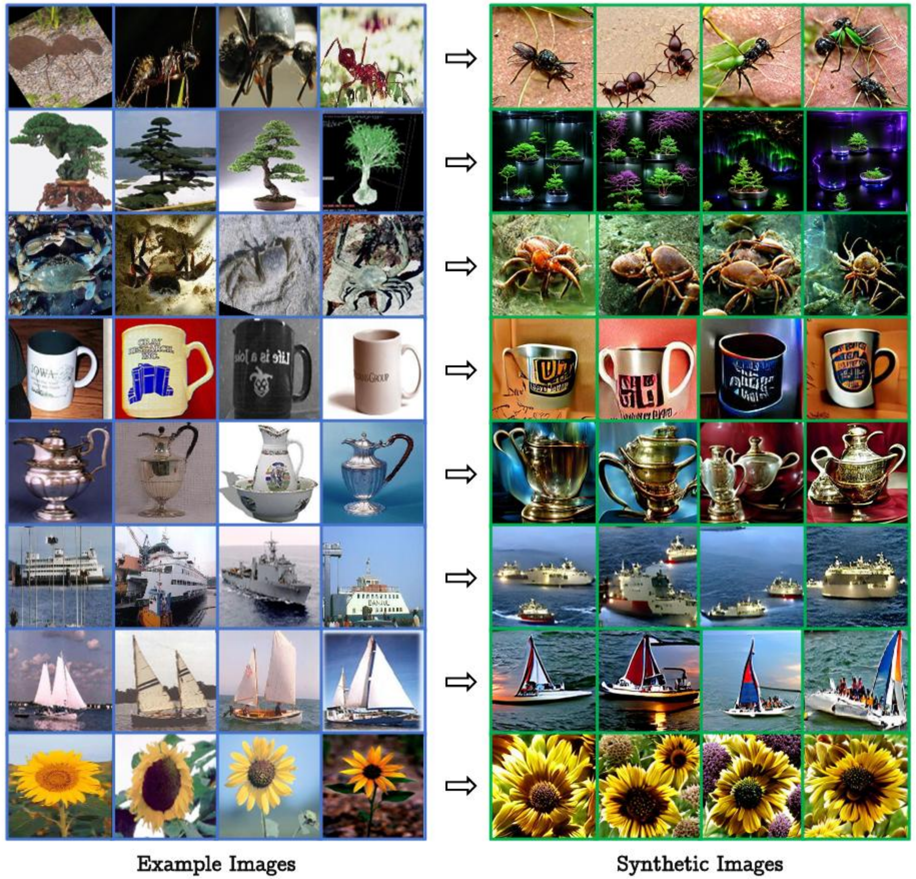


Figure 7: **Visualizations of the image-based customization.** Using only few example images, we synthesize corresponding images that serve as valuable supplements to the original images.

Limitation

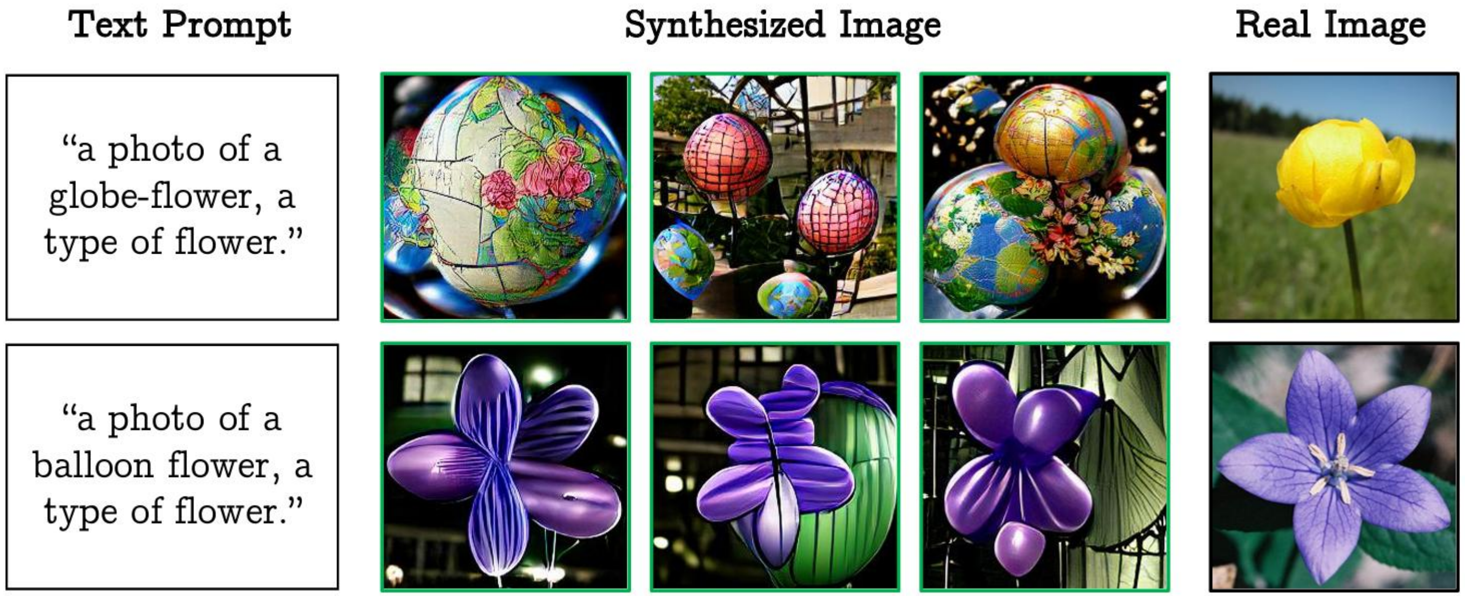

Text Prompt
Synthesized Image
Real Image
“a photo of a globe-flower, a type of flower.”
“a photo of a balloon flower, a type of flower.”

Figure 8: **Examples illustrating misalignment between input texts and synthesized images**, which can lead to suboptimal performance in knowledge distillation. CLIP may struggle with certain specialized terms and fine-grained classes. **Upper:** Request for a globe-flower, but CLIP ambiguously combines globe and flower. **Lower:** Balloon flower refers to a brand of flowers.

Table 6: **Strategies to alleviate text ambiguity (%).**

	Flower-102
Text prompt	15.83
+ Constraint	18.07(+2.24%)
+ Image prompt	74.72(+58.89%)

Thank you for listening.

Nice to meet you in Singapore!