ImagineFSL: Self-Supervised Pretraining Matters on Imagined Base Set for VLM-based Few-shot Learning

Haoyuan Yang [1,*], Xiaoou Li [2,*], Jiaming Lv [1], Xianjun Cheng [1], Qilong Wang [3], Peihua Li [†]

[1] Dalian University of Technology , [2] Beijing University of Posts and Telecommunications,

[3] Tianjin University

Our Lab

Code

[*] Equal contribution.

[†] Corresponding author.

Content

- Introduction
  - Sub-optimal strategies of using synthetic images for CLIP adaption
- Proposed Method
  - Pretraining on imagined base set
  - Fine-tuning for downstream tasks
  - Synthesizing captions and images
- Experiments
  - Few-shot recognition
  - Domain generalization and zero-shot recognition
  - Ablation
- Conclusion

Introduction

- **Sub-optimal strategies of using synthetic images for CLIP Adaption**

  - Morden **text-to-image models** offer a promising solution for adapting CLIP in few-shot tasks

  - Existing approaches [17–19] simply treat **synthetic images as complements**, directly fine-tuning CLIP models using few-shot real images augmented with the synthetic ones during adaptation.

  - However, we argue that **such a strategy is sub-optimal** because it treats synthetic images merely as complementsto real images, instead of as **standalone knowledge repositories** derived from distinct foundation models.

[17] Victor G Turrisi da Costa, Nicola Dall'Asen, Yiming Wang, Nicu Sebe, and Elisa Ricci. Diversified in-domain synthesis with efficient fine-tuning for few-shot classification. *arXiv preprint arXiv:2312.03046*, 2023.

[18] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *International Conference on Learning Representations*, 2023.

[19] Jae Myung Kim, Jessica Bader, Stephan Alaniz, Cordelia Schmid, and Zeynep Akata. DataDream: Few-shot guided dataset generation. In *European Conference on Computer Vision*, pages 252–268, 2025

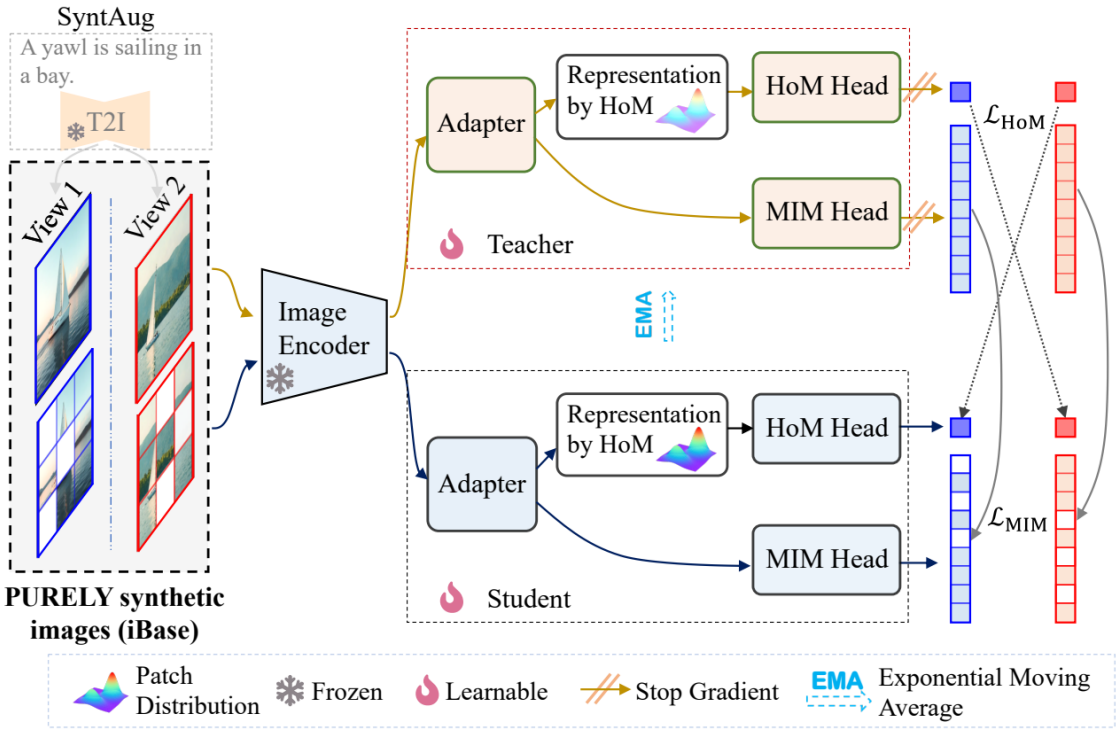
Introduction

Main Contribution

- We frame synthetic images as ***standalone knowledge repositories*** and present a CLIP adaptation methodology that ***pretrains on PURELY synthetic images*** before ***fine-tuning*** for few-shot tasks.
  - This marks a clear departure from existing one-stage fine-tuning methods that simply *treat synthetic images as complements* to real images.
- We propose ***an improved Self-SL method for pretraining*** based on DINO, specifically tailored for FSL.
  - It introduces *higher-order moments* for image representation and employs *synthetic augmentation* for effective view construction.
- We develop **a systematic and scalable pipeline** for synthesizing both captions and images, enabling generation of large-scale base sets for pretraining and task-specific datasets.
  - Distinct from existing arts, *we leverage CoT and ICL techniques* for diverse, realistic image generation.

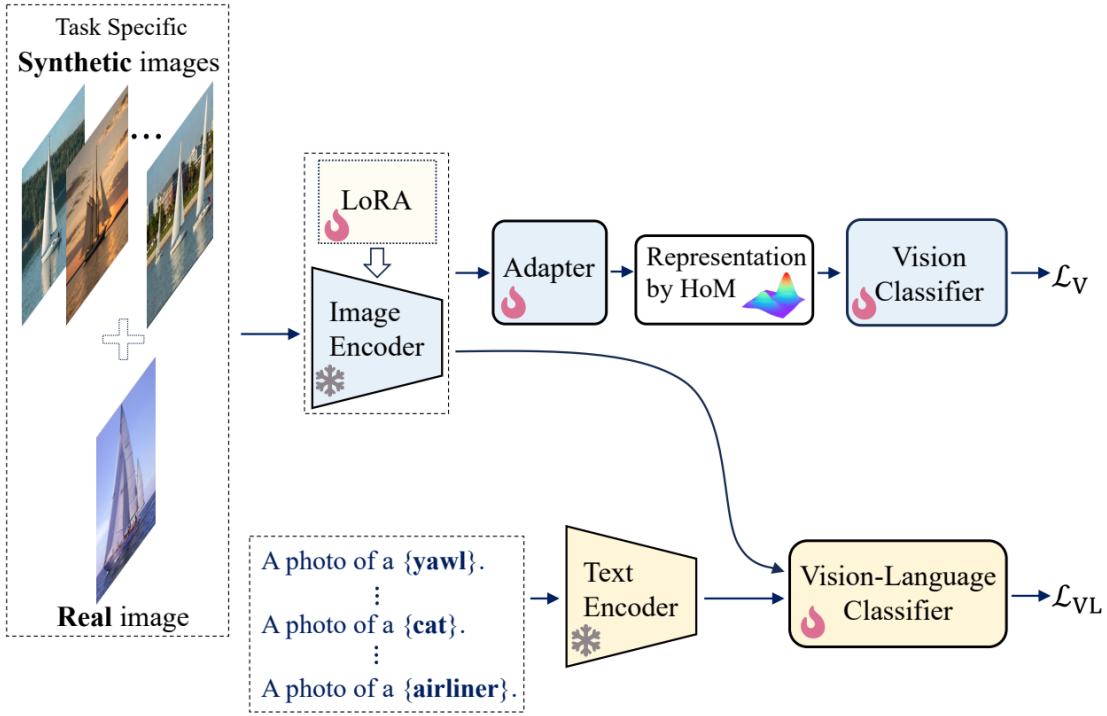
Proposed Method

Overview



(a) Self-supervised pretraining on PURELY *synthetic dataset of iBase*.

Pretraining



(b) Fine-tuning with *real and task-specific synthetic images*.

Fine-tuning

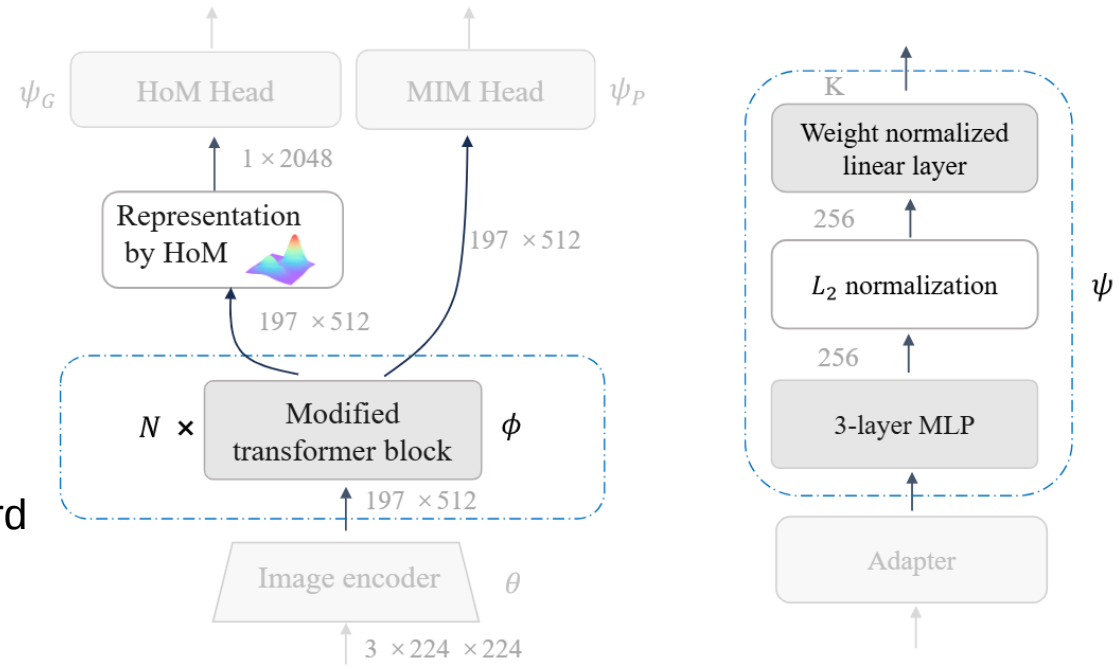
Proposed Method

Overview

➢ Architecture of adapter & projection head

☐ **Adapter** (for ViT-B/16):

- Multi-Head Self-Attention:
  - 6 attention heads, each with a 64-dim Q/K/V.
  - Concatenated attention output to 384-dim. Projected back to 512-dim via a linear layer.
- Feed-Forward Network:
  - Hidden layer size: 1024.
  - This is 2× the input dimension instead of 4× in standard Transformer architectures.



(a) Adapter. (b) Projection head.

☐ **Projection Head**:

- Design of architecture is adopted from [21].

[21] Maxime Oquab, Timothee Darcet, Theo Moutakanni, et al. DINOv2: Learning robust visual features without supervision. In *TMLR* (2024): 1-31.

Proposed Method

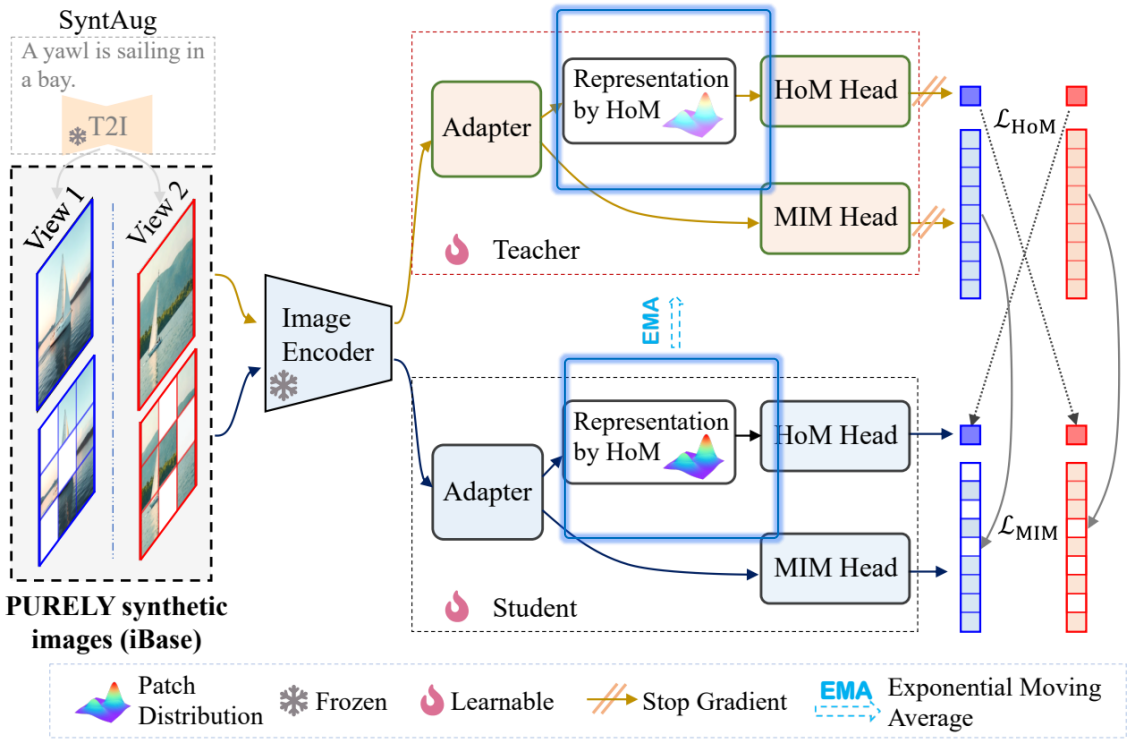
Pretraining on imagined base set (HoM-DINO)

- Low- to High-order Moments (HoM) as image representations
- **Local image regions** are crucial in few-shot tasks [22-24]

Let $\{\mathbf{f}_i\}_i^n = 1$ be p-dimensional patch tokens from adapter ϕ

- [CLS] Token: $[\mathrm{CLS}]$
- 1-st moment: $\mathbf{m}_1 = \frac{1}{n}\sum_i \mathrm{f}_i$
- 2-nd moment: $\mathbf{m}_2 = \left(\frac{1}{n}\sum_i (\mathbf{f}_i - \mathbf{m}_1)^2\right)^{\frac{1}{2}}$
- 3-rd moment: $\mathbf{m}_3 = \left(\frac{1}{n}\sum_i (\mathbf{f}_i - \mathbf{m}_1)^3\right)^{\frac{1}{3}}$

HoM representation: $\left[[\mathrm{CLS}]; \mathbf{m}_1; \mathbf{m}_2; \mathbf{m}_3\right] \in \mathbb{R}^{4p}$



(a) Self-supervised pretraining on PURELY *synthetic dataset of iBase*.

[22] Carl Doersch, Ankush Gupta, and Andrew Zisserman. CrossTransformers: spatially-aware few-shot transfer. In *NeurIPS*, 2020.

[23] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. DeepEMD: Differentiable earth mover's distance for few-shot learning. In *IEEE TPAMI* 45.5 (2022): 5632-5648.

[24] Marc Lafon, Elias Ramzi, Clement Rambour, Nicolas Audebert, and Nicolas Thome. GalLoP: Learning global and local prompts for vision-language models. In *ECCV*, 2024

Proposed Method

Pretraining on imagined base set (HoM-DINO)

➢ Low- to High-order Moments (HoM) as image representations

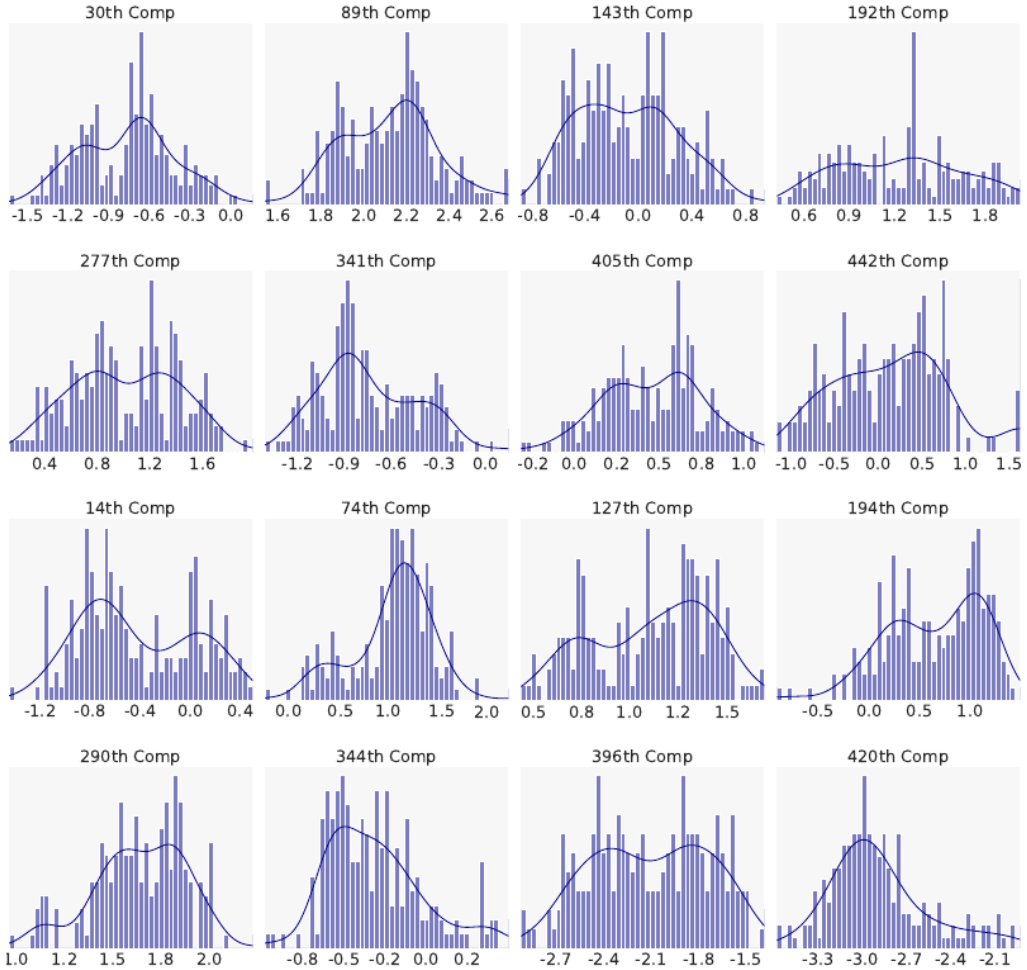
Why use HoM?

- We prefer using statistical moments over assuming a prior distribution (e.g., Gaussian) for modeling feature distributions.

- We extract 512-dim features of input image from the last block of adapter and visualize histogram for each feature component (Comp).

- These histograms are then fitted with Kernel Density Estimation (KDE).

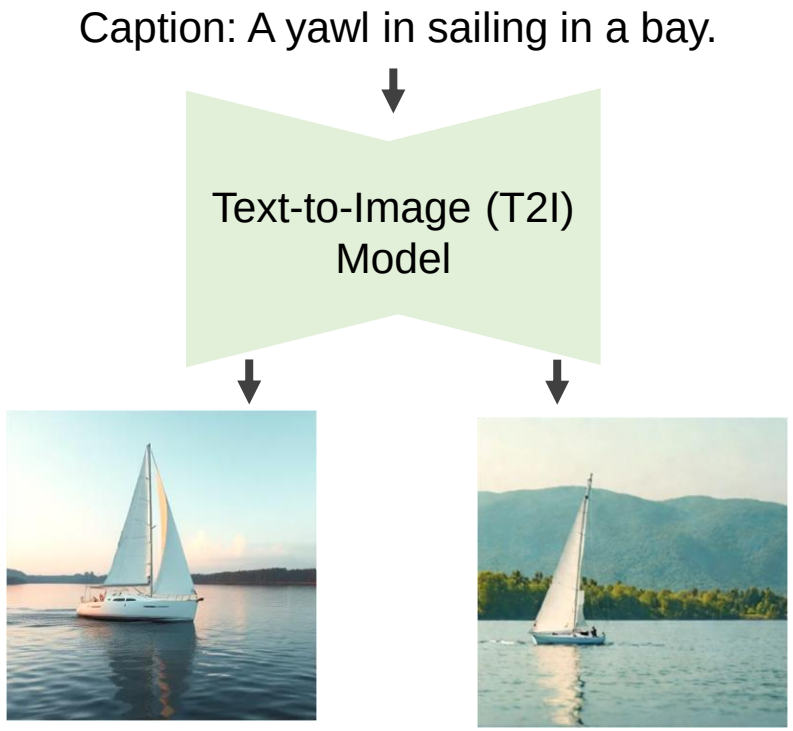
- The distributions are complex and varied, suggesting that assuming specifically a prior distribution may be sub-optimal.



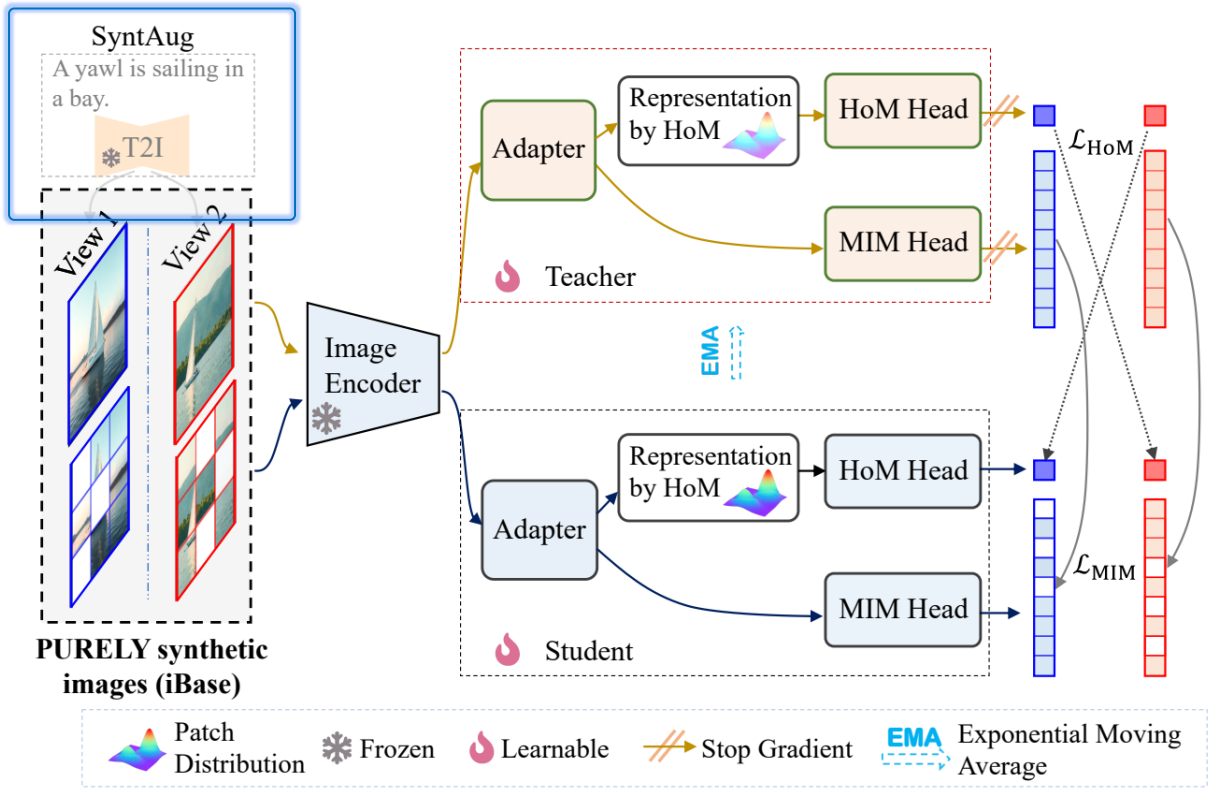
Proposed Method

Pretraining on imagined base set (HoM-DINO)

- Synthetic Augmentation (SyntAug)



They convey the ***same scenery*** but ***with natural variations*** due to the randomness of noise in the generative models.



(a) Self-supervised pretraining on PURELY *synthetic dataset of iBase*.

Proposed Method

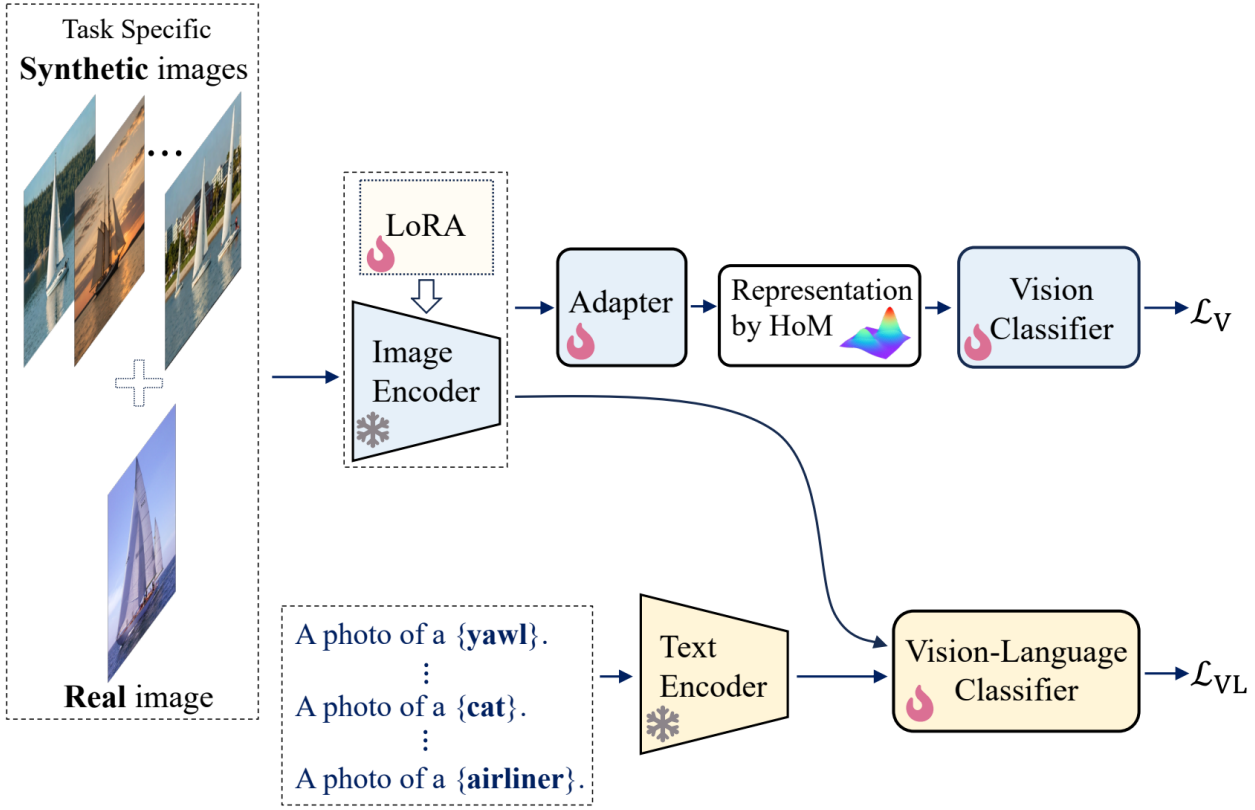
Fine-tuning for downstream tasks

ImagineFSL:

- We attach to the adapter a vision (V) classifier with a cross-entropy loss (CE) $\mathcal{L}_V$ for vision-based recognition.
- We tune a learnable vision-language (VL) classifier with a CE loss $\mathcal{L}_{VL}$ [18, 53] for integration with language knowledge.

ImagineFSL$_{\text{LoRA}}$:

- We further tune the image encoder, e.g., by introducing LoRA.



[18] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *ICLR*, 2023

[53] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *CVPR*, 2022.

Proposed Method

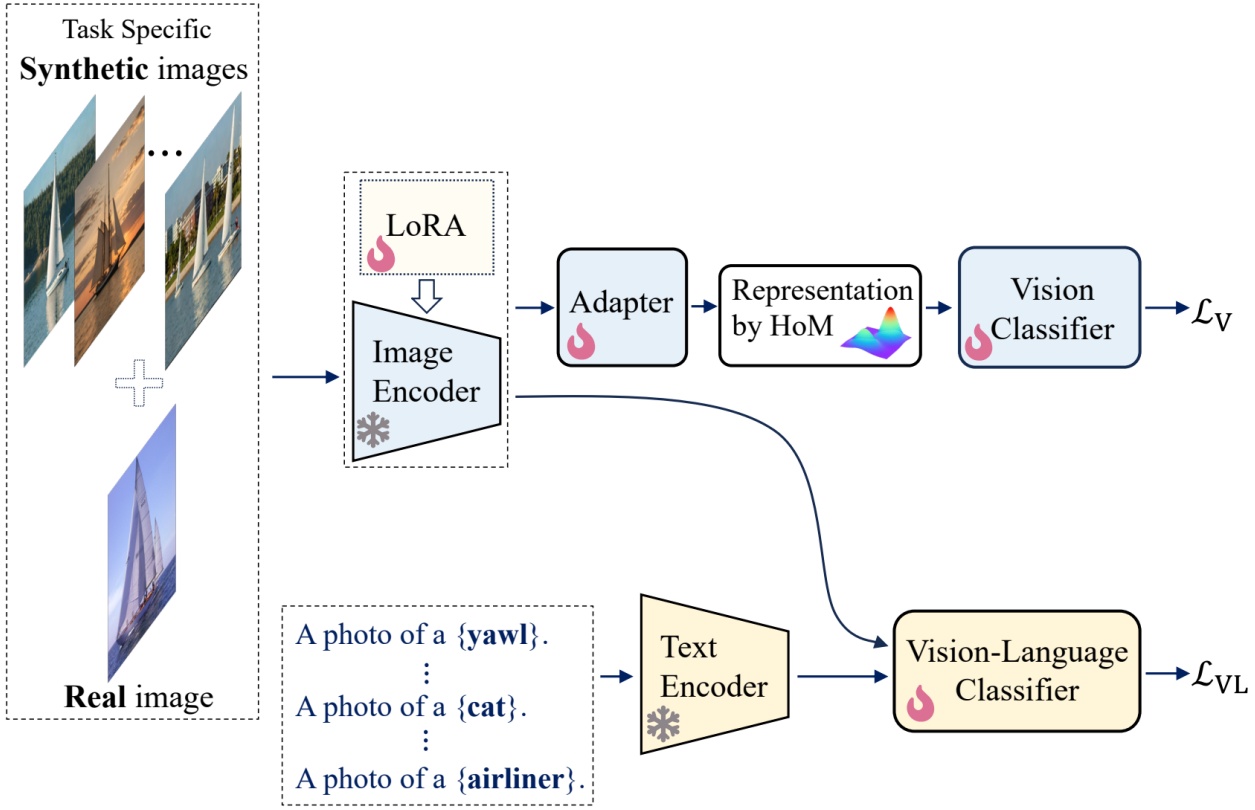
Fine-tuning for downstream tasks

Training:

- We use mix training [17, 18, 19] that simultaneously uses real images and synthetic images in a batch
- We adopt an ensemble of textual templates hand-engineered by CLIP and CuPL [72] for initialization of vision-language classifier.

Inference:

- We average the logits from two classifiers with a fusion weight.



[17] Victor G Turrisi da Costa, Nicola Dall'Asen, Yiming Wang, Nicu Sebe, and Elisa Ricci. Diversified in-domain synthesis with efficient fine-tuning for few-shot classification. *arXiv preprint arXiv:2312.03046*, 2023.

[18] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *International Conference on Learning Representations*, 2023.

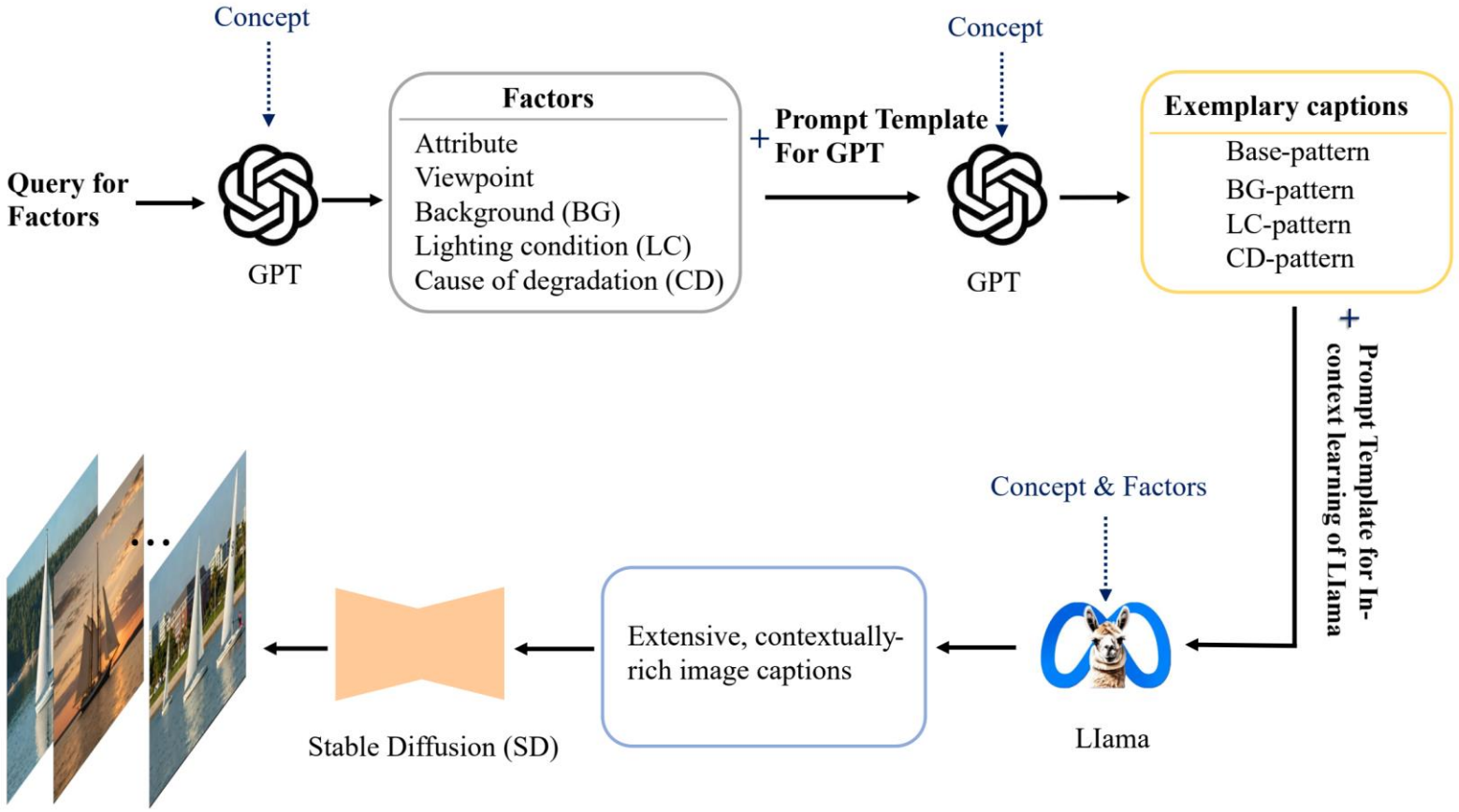
[19] Jae Myung Kim, Jessica Bader, Stephan Alaniz, Cordelia Schmid, and Zeynep Akata. DataDream: Few-shot guided dataset generation. In *European Conference on Computer Vision*, pages 252–268, 2025

[72] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. Whatdoes a platypus look like? Generating customized promptsfor zero-shot image classification. In *CVPR,* 2023.

Proposed Method

Synthesizing captions and images

- Overview



Proposed Method

Synthesizing captions and images

➢ Query GPT-4 for factors

- We identify five factors that play an important role in constituting image caption:
- Attribute
- Viewpoint
- Background (BG)
- Lighting condition (LC)
- Cause of degradation (CD)

- Example for **Background (BG)**:

PTe for GPT-4 to analyze *factors*

Generate a list of 20 distinct visual backgrounds that distinguish this particular {concept} from the others while accommodating variations within the {concept} in everyday photographs. Present this list as a Python array, with each element being a concise phrase that describes a unique background.

- Example concept: **Leonberger**
- Output:

```
“lonberger": [
        "distinctive lion-like mane",
        "large and muscular build",
        "long, thick fur",
        "black mask on the face",
        "gentle, expressive eyes",
        "bushy tail carried low",
        "broad head shape",
        "triangular, pendant ears",
        "dense double coat",
        "fawn to reddish-brown coat color",
        "prominent chest",
        "strong, straight legs",
        "long, feathered legs",
        "slightly arched neck",
        "white markings on toes or chest",
        "deep and wide muzzle",
        "large, rounded feet",
        "relaxed and calm demeanor",
        "large, blocky nose",
        "distinctive facial markings"
],
```

Proposed Method

Synthesizing captions and images

➢ Generate examples via GPT-4

- Based on these factors, we design four patterns for systemic caption generation.

	Attribute	Viewpoint	Background	Lighting condition	Causes of degradation
Base-pattern	✓	✓			
BG-pattern	✓	✓	✓		
LC-pattern	✓	✓		✓	
CD-pattern	✓	✓			✓

Table S-4. Four patterns to synthesize captions.

- Example for **BG-Pattern**:

PTe for GPT-4 to generate exemplary captions

Your task is to create diverse and contextually rich captions for {concept}, which will serve as prompts for text-to-image models such as stable diffusion to generate images. To achieve this, consider the essential factors that influence visual image generation as follows: attribute that visually distinguishes the {concept}, viewpoint of the camera to capture the scenario, and background where the {concept} is photographed. These factors are crucial not only for generating images representative of various categories but also for ensuring the synthesized images reflect common photographic practices in everyday life.

For the {concept}, you will be provided with specific forms for the factors, i.e., attribute: {attribute}, viewpoint: {viewpoint} and background: {background}. By integrating them, you will generate a caption of less than 36 words. Use concise, clear and straightforward language, avoiding extravagant embellishments and vague expressions.

- Example concept: **Leonberger**
- Output:

A *large, muscular* ***leonberger*** in *profile view*, sitting majestically on a *window sill*, sunlight softly illuminating its fur through the glass.

Proposed Method

Synthesizing captions and images

➢ Generate large scale captions via Llama 3 8B

- We deploy the lightweight Llama 3 8B for generating extensive captions.
- We use exemplary captions generated by GPT-4 to facilitate in-context learning for Llama 3 8B.

- Prompt Template for **BG-Pattern**:

PTe for Llama to generate captions for *BG-pattern*

Your task is to generate an image caption for a {concept}, by considering the following factors: attribute, viewpoint, and background. The caption should be suitable for use as a textual prompt for Stable Diffusion, ensuring that the generated image resembles a real-life photo. Use the three examples provided below to guide the generation of the caption:
{}, {} and {} → {}
{}, {} and {} → {}
{}, {} and {} → {}
{}, {} and {} →

- Example concept: **Leonberger**

- Examples for Llama 3 8B:

{*large and muscular build*}, {*profile view*} and {*window sill*}
→ {A *large, muscular* leonberger in *profile view*, sitting majestically on a *window sill*, sunlight softly illuminating its fur through the glass.}

{*white markings on toes or chest*}, {*low-angle shot*} and {*Café*}
→ {A leonberger with *white markings on its toes and chest*, captured from a *low-angle shot*, sitting comfortably in a bustling *café*.}

{*relaxed and calm demeanor*}, {*front view*} and {*staircase*}
→ {A leonberger with a *relaxed and calm demeanor*, photographed from a *front view*, sitting on a *staircase* in a cozy home.}

- Factors:

{*slightly arched neck*}, {*profile view*} and {*couch*} →

- Output:

{Leonberger with a *slightly arched neck*, captured from a slightly elevated *front view*, resting on a *couch*.}

Proposed Method

Synthesizing captions and images

➢ Generate large scale captions via Llama 3 8B

- We deploy the lightweight Llama 3 8B for generating extensive captions.
- We use exemplary captions generated by GPT-4 to facilitate in-context learning for Llama 3 8B.

- Prompt Template for **BG-Pattern**:

PTe for Llama to generate captions for *BG-pattern*

Your task is to generate an image caption for a {concept}, by considering the following factors: attribute, viewpoint, and background. The caption should be suitable for use as a textual prompt for Stable Diffusion, ensuring that the generated image resembles a real-life photo. Use the three examples provided below to guide the generation of the caption:
{}, {} and {} → {}
{}, {} and {} → {}
{}, {} and {} → {}
{}, {} and {} →

- Example concept: **Leonberger**

- Factors:

{*slightly arched neck*}, {*profile view*} and {*couch*} →

- Output:

{Leonberger with a *slightly arched neck*, captured from a slightly elevated *front view*, resting on a *couch*.}

- Synthesizing image by SD3:

Experiments

➢ Few-shot recognition

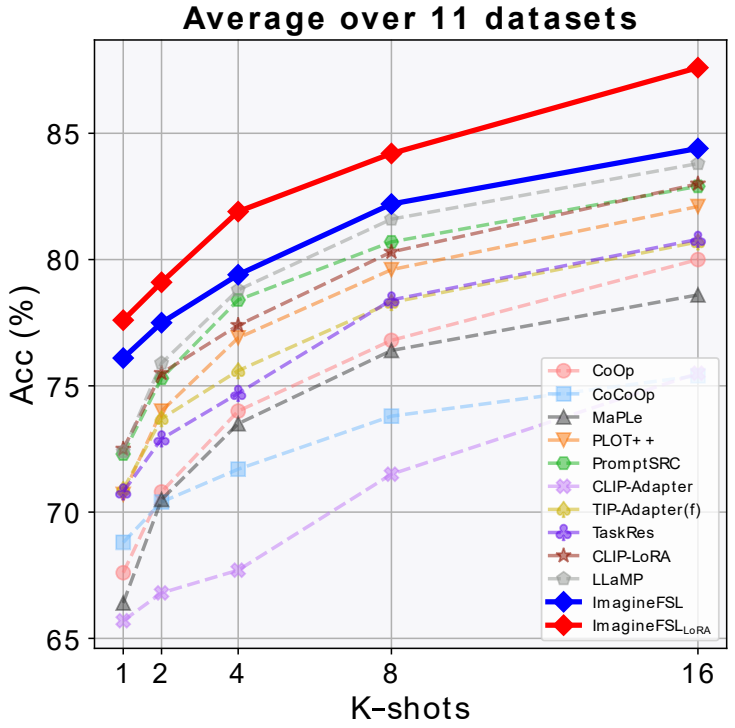
- Comparison to previous methods that *using synthetic images*.

Method	T2I	EN	Syn Data	ImageNet	Caltech	Aircraft	Cars	Food	Pets	Flowers	DTD	EuroSAT	SUN	Avg Acc
IsSynth [18]	❄	❄	[18]	–/73.4	–/96.0	–/46.7	–/82.6	–/87.3	–/92.8	–/86.5	–/73.6	–/87.1	–/76.2	–/80.2
CaFo [13]	❄	❄	[13]	70.2/73.1	93.9/96.9	30.1/46.2	71.5/85.4	85.6/87.6	91.5/**94.9**	74.9/97.6	59.4/72.0	66.5/86.3	71.2/76.9	71.5/81.7
IsSynth†	❄	❄	Ours	70.7/73.9	94.4/96.8	31.1/46.5	71.7/85.4	85.7/87.4	91.6/93.8	85.4/95.7	60.1/73.9	75.9/87.2	70.9/76.5	73.8/81.7
CaFo†	❄	❄	Ours	70.2/73.9	94.2/96.9	30.6/47.4	72.6/85.7	86.1/87.6	91.8/**94.9**	76.1/97.8	60.1/72.5	68.7/86.7	71.3/76.9	72.2/82.0
ImagineFSL	❄	❄	Ours	**71.6/74.7**	**95.3/97.1**	**31.5/54.8**	**82.6/90.5**	**86.7/87.8**	**92.3**/94.3	**87.0/98.7**	**62.7/75.8**	**76.2/89.6**	**72.0/78.1**	**75.8/84.1**
DISEF [17]	❄	🔥	[17]	–/73.9	–/96.9	–/ 63.9	–/88.6	–/87.1	–/94.3	–/98.9	–/75.4	–/94.3	–/77.4	–/85.1
DataDream [19]	🔥	🔥	[19]	–/74.1	–/96.9	–/72.3	–/92.4	–/87.6	–/94.8	–/99.4	–/**81.6**	– /93.4	–/77.5	–/87.0
DISEF†	❄	🔥	Ours	71.0/74.0	93.2/97.0	32.0/67.9	72.6/91.2	86.3/87.2	92.5/94.5	88.0/98.9	60.5/76.0	82.3/94.3	70.6/77.1	74.9/85.8
ImagineFSL$_{\text{LoRA}}$	❄	🔥	Ours	**71.8/75.2**	**95.4/97.9**	**34.0/74.1**	**82.8/92.9**	**86.8/88.3**	**92.8/95.4**	**90.2/99.7**	**63.7**/78.0	**82.7/95.0**	**72.8/78.8**	**77.3/87.5**

Experiments

➢ Few-shot recognition

- Comparison to previous methods that *only using real images*.



	Method	1-shot	2-shot	4-shot	8-shot	16-shot	Avg Acc
PT	CoOp [11]	67.6	70.8	74.0	76.8	80.0	73.8
	CoCoOp [9]	68.8	70.4	71.7	73.8	75.4	72.0
	MaPLe [73]	66.4	70.5	73.5	76.4	78.6	73.1
	ProGrad [74]	69.5	72.4	74.7	77.4	79.9	74.8
	PLOT++ [75]	70.7	74.0	76.9	79.6	82.1	76.7
	PromptSRC [76]	72.3	75.3	78.4	80.7	82.9	77.9
AT	CLIP-Adapter [41]	65.7	66.8	67.7	71.5	75.5	69.4
	TIP-Adapter(f) [12]	70.9	73.7	75.6	78.3	80.7	75.8
	TaskRes [77]	70.8	72.9	74.7	78.4	80.8	75.5
	ImagineFSL	76.1	77.5	79.4	82.2	84.4	79.9
ET	CLIP-LoRA [39]	72.5	75.5	77.4	80.3	83.0	77.7
ET+PT	LLaMP [42]	72.4	75.9	78.8	81.6	83.8	78.5
PT+AT	GalLoP [24]	72.8	76.4	79.1	82.2	84.5	79.0
ET+AT	ImagineFSL$_{LoRA}$	**77.6**	**79.1**	**81.9**	**84.2**	**87.6**	**82.1**

Experiments

➢ Domain generalization and zero-shot recognition

Domain generalization:

- We specifically synthesize images for IN-S and IN-R datasets

	Method	Source	Target				
		ImageNet	-V2	-S	-A	-R	Avg Acc
PT	CoOp [11]	71.5	64.2	48.0	49.7	75.2	59.3
	CoCoOp [9]	71.0	64.1	48.8	50.6	76.2	59.9
	MaPLe [73]	70.7	64.1	49.2	50.9	77.0	60.3
	PromptSRC [76]	71.3	64.4	49.6	50.9	77.8	60.7
	CoCoLe [14]	73.9	65.9	50.9	**51.8**	78.9	61.9
AT	TaskRes [77]	73.1	65.3	49.1	50.4	77.7	60.6
	GraphAdapter [78]	73.4	65.6	49.2	50.6	77.7	60.8
	ImagineFSL	74.7	67.0	52.7	51.5	79.7	62.7
ET+AT	ImagineFSL$_{\text{LoRA}}$	**75.2**	**67.5**	**53.8**	51.5	**80.0**	**63.2**
PT+AT	GalLoP [24]	75.1	**67.5**	49.5	50.3	77.8	61.3

Zero-shot recognition

- We perform fine-tuning using solely the synthetic images, without touch of any real images.

	Method	*ImageNet*	*Caltech*	*Aircraft*	*Cars*	*Food*	*Pets*	*Flowers*	*DTD*	*EuroSAT*	*SUN*	*UCF101*	*Avg Acc*
PT	VCD [79]	68.0	–	–	–	88.5	86.9	–	45.6	48.8	–	–	–
	Sus-X [80]	69.9	94.0	28.7	66.1	86.1	90.6	73.8	54.6	57.5	67.7	66.6	68.7
	TPT [81]	69.0	94.2	24.8	66.9	84.7	87.8	69.0	47.8	42.4	65.5	68.0	65.5
	DiffTPT [82]	70.3	92.5	25.6	67.0	87.2	88.2	70.1	47.0	43.1	65.7	68.2	65.9
	DMN [83]	**72.3**	**95.3**	30.0	68.0	85.1	92.0	**74.5**	55.9	59.4	70.2	72.5	70.5
	CODER [84]	71.5	–	–	–	**89.8**	92.0	–	55.7	60.5	–	–	–
AT	ImagineFSL	70.9	94.7	30.2	78.6	86.5	**92.3**	73.9	59.2	65.6	**70.3**	73.2	72.3
ET+AT	ImagineFSL$_{\text{LoRA}}$	71.0	94.7	**30.4**	**78.7**	86.5	**92.3**	73.6	**60.8**	**71.7**	**70.3**	**73.3**	**73.0**

Experiments

➢ Ablation

Is pretraining necessary?

Pretrain	Syn Img	ImageNet	Aircraft	Flowers	EuroSAT	Avg Acc
–	✗	68.8/72.8	24.1/48.5	78.9/96.5	60.4/84.4	58.1/75.6
–	✓	70.9/73.9	27.5/49.3	84.1/96.5	66.2/86.2	62.2/76.5
SL	✓	71.0/73.9	28.7/49.5	85.3/97.0	67.3/86.8	63.1/76.8
HoM-DINO	✓	**71.5**/**74.3**	**29.4**/**51.2**	**86.3**/**98.3**	**74.5**/**87.8**	**65.4**/**77.9**

How HoM-DINO against other Self-SL methods?

Self-SL	ImageNet	Aircraft	Flowers	EuroSAT	Avg Acc
DINOv2	71.1/73.9	28.9/50.6	85.2/97.5	67.8/86.8	63.3/77.2
SynCLR	**71.5**/74.0	27.4/51.0	84.4/97.7	74.4/86.9	64.4/77.4
MAE	71.2/**74.3**	26.4/47.6	83.6/97.4	66.5/86.7	61.9/76.3
HoM-DINO	**71.5**/**74.3**	**29.4**/**51.2**	**86.3**/**98.3**	**74.5**/**87.8**	**65.4**/**77.9**

Experiments

➢ Ablation

How image representation (image-repr) and view construction impact?

	Image Repr	Dim	View	MIM	ImageNet	Aircraft	Flowers	EuroSAT	Avg Acc
1	[CLS]	512	HA	✗	71.2/73.5	27.2/50.0	83.9/97.2	66.7/86.9	62.3/76.9
2	[CLS]	512	HA	✓	71.1/73.9	28.9/50.6	85.2/97.5	67.8/86.8	63.3/77.2
3	[CLS]	2048	HA	✓	71.4/73.6	29.0/49.9	84.3/97.6	69.8/86.8	63.6/77.0
4	AvgP	512	HA	✓	71.4/73.7	29.2/50.3	84.2/97.4	69.6/87.1	63.6/77.1
5	AttnP	512	HA	✓	71.3/73.2	28.8/56.9	85.7/97.6	69.4/87.4	63.8/77.2
6	G^2De-Net	8385	HA	✓	71.1/73.6	29.3/50.3	85.3/97.5	70.2/87.7	64.0/77.3
7	MP	2087	HA	✓	71.3/73.6	29.3/48.9	82.7/97.4	68.6/86.2	63.0/76.5
8	HoM	2048	HA	✓	71.4/74.0	29.1/50.9	85.6/98.1	72.0/87.3	64.5/77.6
9	HoM	2048	SA	✓	**71.5/74.3**	**29.4/51.2**	**86.3/98.3**	**74.5/87.8**	**65.4/77.9**

Experiments

➢ Ablation

Influence of textual prompts.

CLIP	CuPL	ImageNet	Aircraft	Flower	EuroSAT	Avg Acc
✓	✗	71.5/**74.3**	29.4/51.2	**86.3**/98.3	**74.5**/**87.8**	65.4/77.9
✗	✓	**71.6**/**74.3**	30.7/**51.6**	85.9/**98.4**	72.0/**87.8**	65.0/78.0
✓	✓	**71.6**/**74.3**	**30.8**/51.5	**86.3**/98.3	73.7/**87.8**	**65.6**/**78.0**

Effects of VL-classifier tuning and encoder tuning.

V-clf	VL-clf	I-EN	ImageNet	Aircraft	Flowers	EuroSAT	Avg Acc
🔥	❄	❄	71.6/74.3	30.8/51.5	86.3/98.3	73.7/87.8	65.6/78.0
🔥	🔥	❄	71.6/74.7	31.5/54.8	87.0/98.7	76.2/89.6	66.6/79.5
🔥	🔥	🔥	**71.8**/**75.2**	**34.0**/**74.1**	**90.2**/**99.7**	**82.7**/**95.0**	**69.7**/**86.0**

Experiments

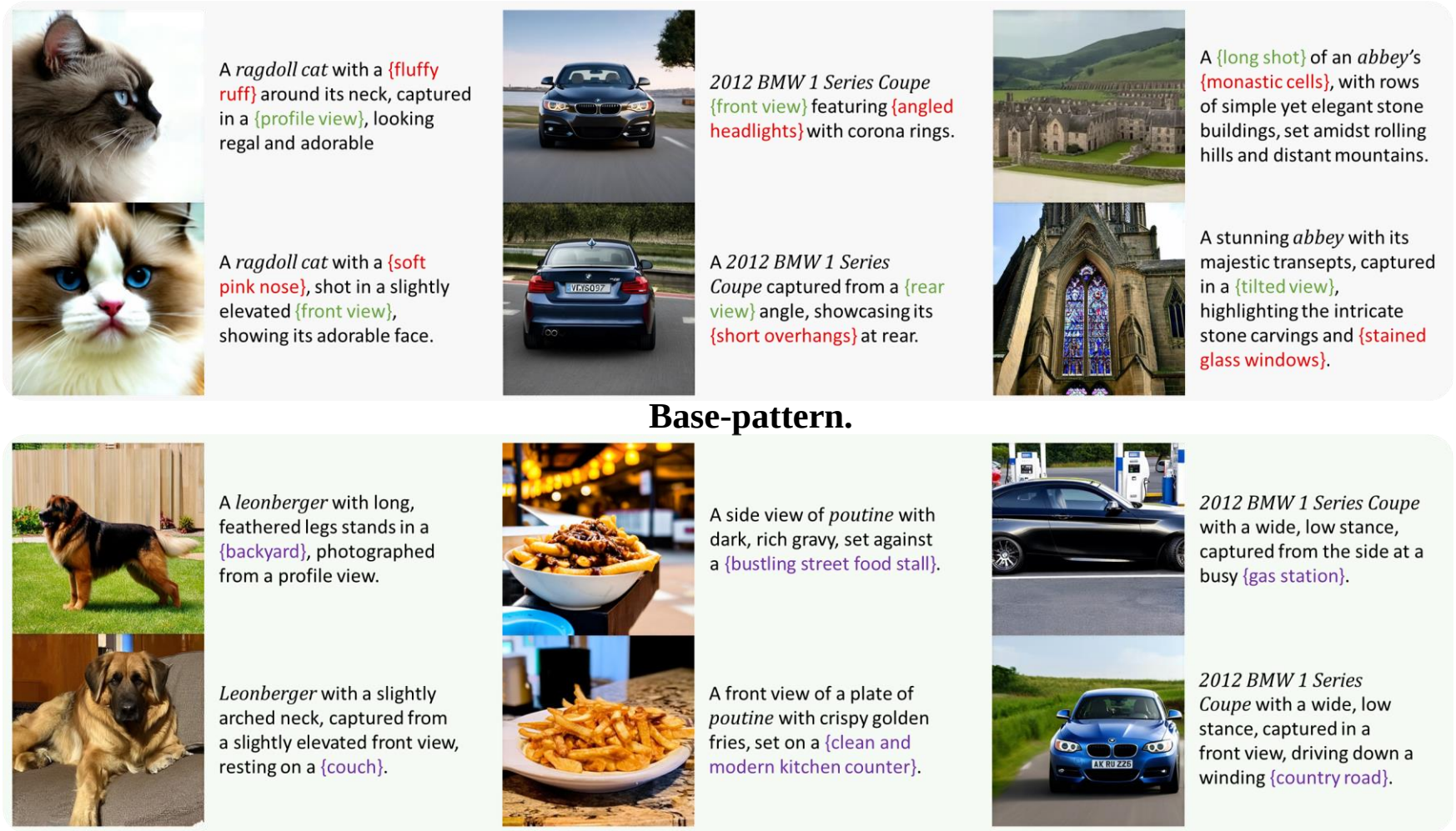
➢ Ablation

How CoT affect generated images?

Method	Chain-of-Thought	Fidelity ↓	Diversity ↑
Handcrafted prompts	✗	45.92	0.95
Ours	✗	40.07	1.06
Ours	✓	**32.59**	**1.21**

Experiments

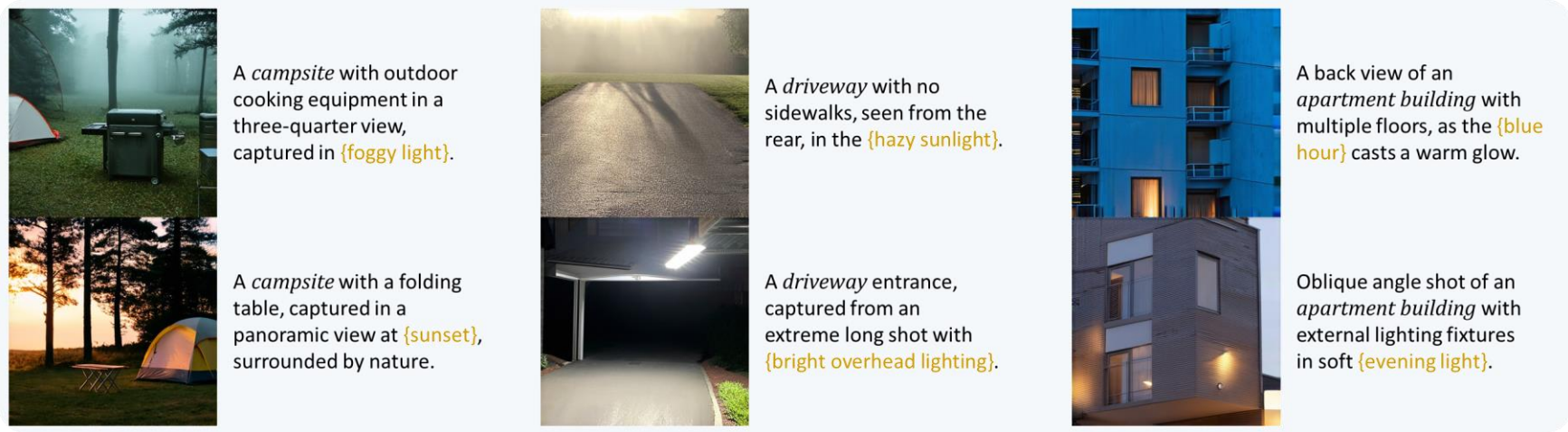
- **Visualization**



Visualization of synthetic images for the four patterns. We highlight the *concept*, and the factors including attributes, viewpoints, backgrounds, lighting conditions and causes of degradation.

Experiments

- **Visualization**

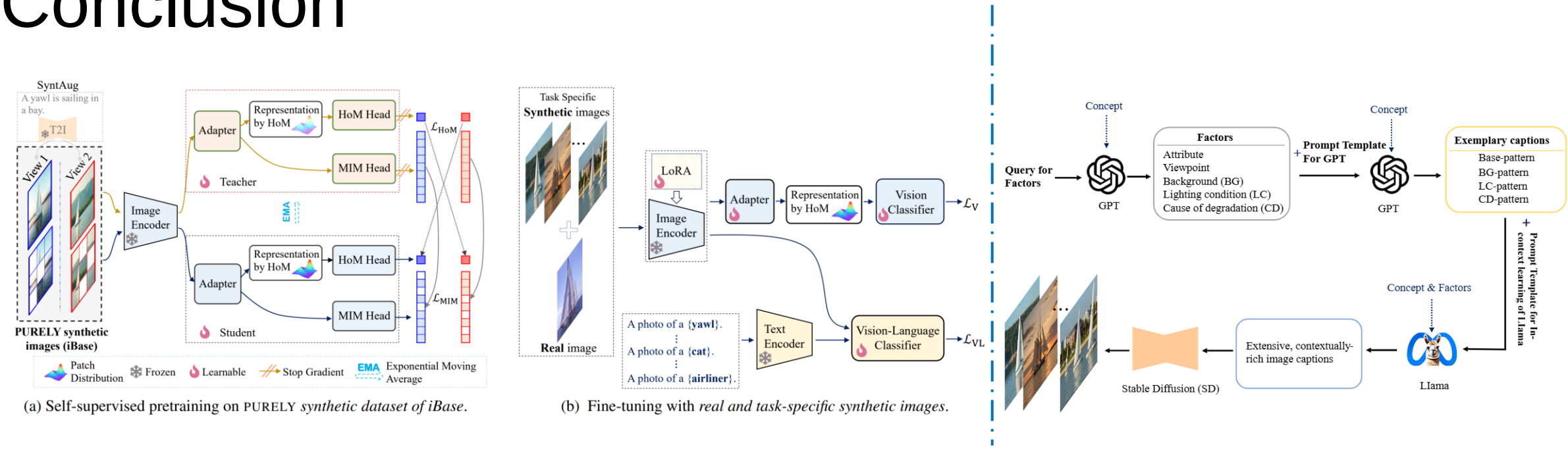


LC-pattern.

CD-pattern.

Visualization of synthetic images for the four patterns. We highlight the *concept*, and the factors including attributes, viewpoints, backgrounds, lighting conditions and causes of degradation.

Conclusion



(a) Self-supervised pretraining on PURELY *synthetic dataset of iBase*.

(b) Fine-tuning with *real and task-specific synthetic images*.

- We introduce **a novel methodology for CLIP adaptation**, involving self-supervised pretraining on an iBase, followed by fine-tuning with real and task-specific synthetic images.
  - Unlike previous methods, **we treat synthetic images as standalone knowledge repositories** of diverse concepts.
- We propose **an improved Self-SL method** based on DINO, specifically tailored for FSL.
- We exploit **CoT and in-context learning techniques** and develop a systematic and scalable pipeline for synthesizing both captions and images.

Thanks for your attention

Contact Us

- Haoyuan Yang: *yanghaoyuan@mail.dlut.edu.cn*

Our Lab

Code