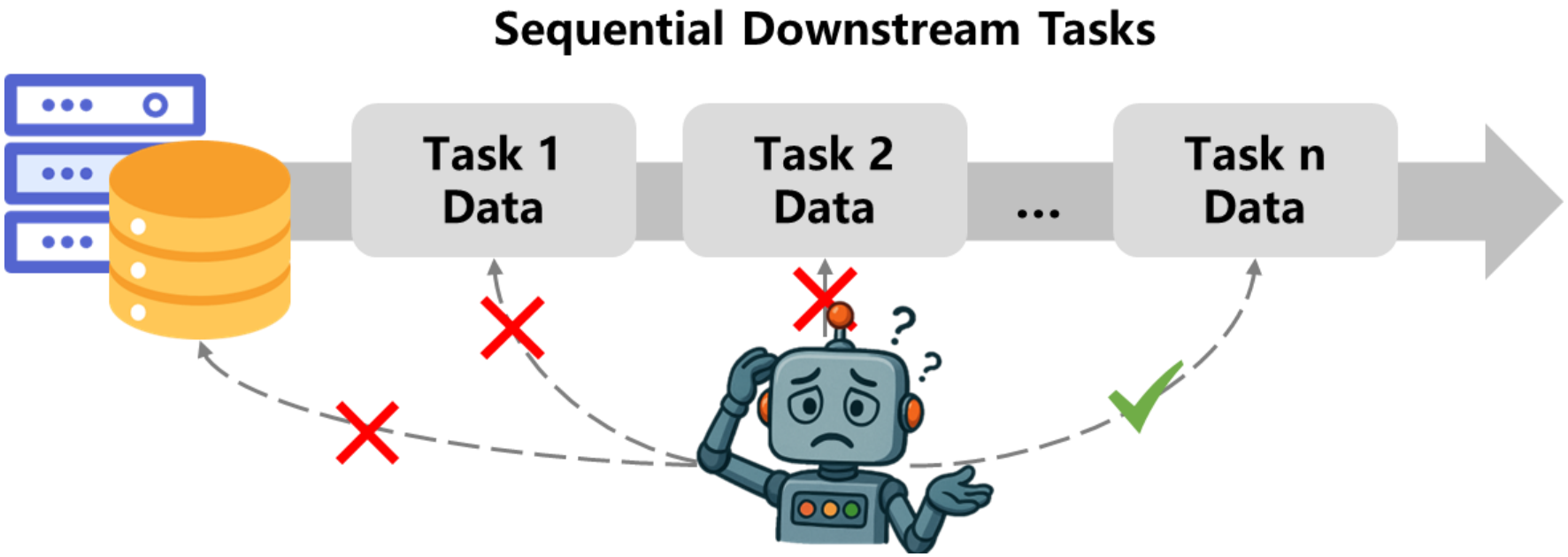
Synthetic Data is an Elegant GIFT for Continual Vision-Language Models

Bin Wu[1]*, Wuxuan Shi[1]*, Jinqiao Wang[2], Mang Ye[1]†
[1]Wuhan University, [2]Wuhan AI Research

https://github.com/Luo-Jiaming/GIFT_CL

Background

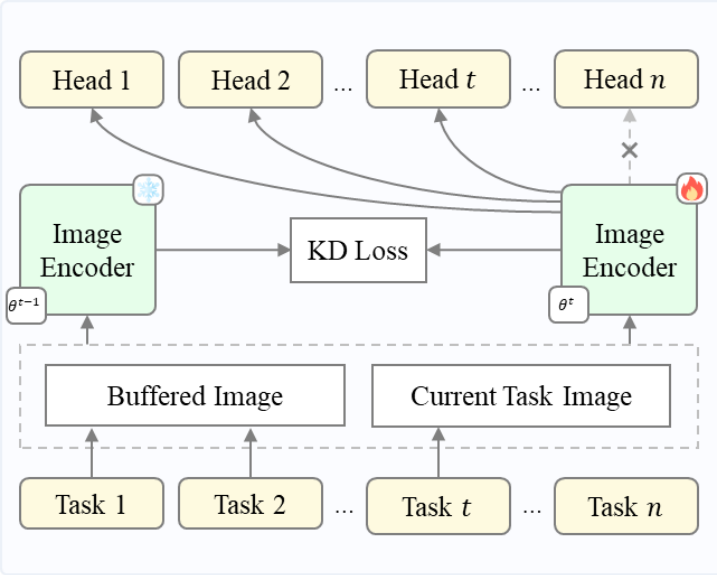
❑ Continual Learning (CL)



➢ When sequentially fine-tuned on multiple downstream tasks, pre-trained vision-language models suffer from catastrophic forgetting of ***general pretraining knowledge*** and ***previously learned downstream tasks***.

Background

❑ Replaying historical data is effective but not necessarily practical



- A straightforward method to alleviate forgetting is to save (part of) historical data for training together with new data.

$$\mathcal{L}_{total} = \mathbb{E}_{(x,y)\sim\mathcal{D}_t}\left[\ell_{ce}\left(y, f(x,\theta_t)\right)\right] + \mathcal{L}_{replay}$$

$\mathcal{L}_{replay}$ can be $\mathbb{E}_{(x,y)\sim\mathcal{D}_{1:t-1}}\left[\ell_{ce}\left(y, f(x,\theta_t)\right)\right]$ or $\mathbb{E}_{(x,y)\sim\mathcal{D}_{1:t-1}}\left[D_{KL}\left(f(x,\theta_{t-1}) \,||\, f(x,\theta_t)\right)\right]$

- Replay-based methods are impractical when ***pre-training data of many models is unavailable*** and ***storing historical data raises privacy concerns***.

Synthetic data is ready for supplement when training data is scarce[1]

*Can **synthetic data from latest diffusion models** help preserve pre-trained VLMs' knowledge during continual learning, and if so, how?*

[1] Is synthetic data from generative models ready for image recognition. ICLR. 2023

Background

❑ Two decoupled sub-questions

Q1: How can diffusion model generate to approximate both pre-training and downstream task data of VLMs?

① Synthetic pre-training data needs to have a wide distribution in the embedding space of the VLM.

② Synthetic downstream data needs to be customized for downstream tasks of different domains.

③ Synthetic images and corresponding text prompts need to be highly aligned in the VLM's embedding space.

Q2: How can the generated data be used to mitigate forgetting?

① Given the generation overhead, we should use as little synthetic data as possible.

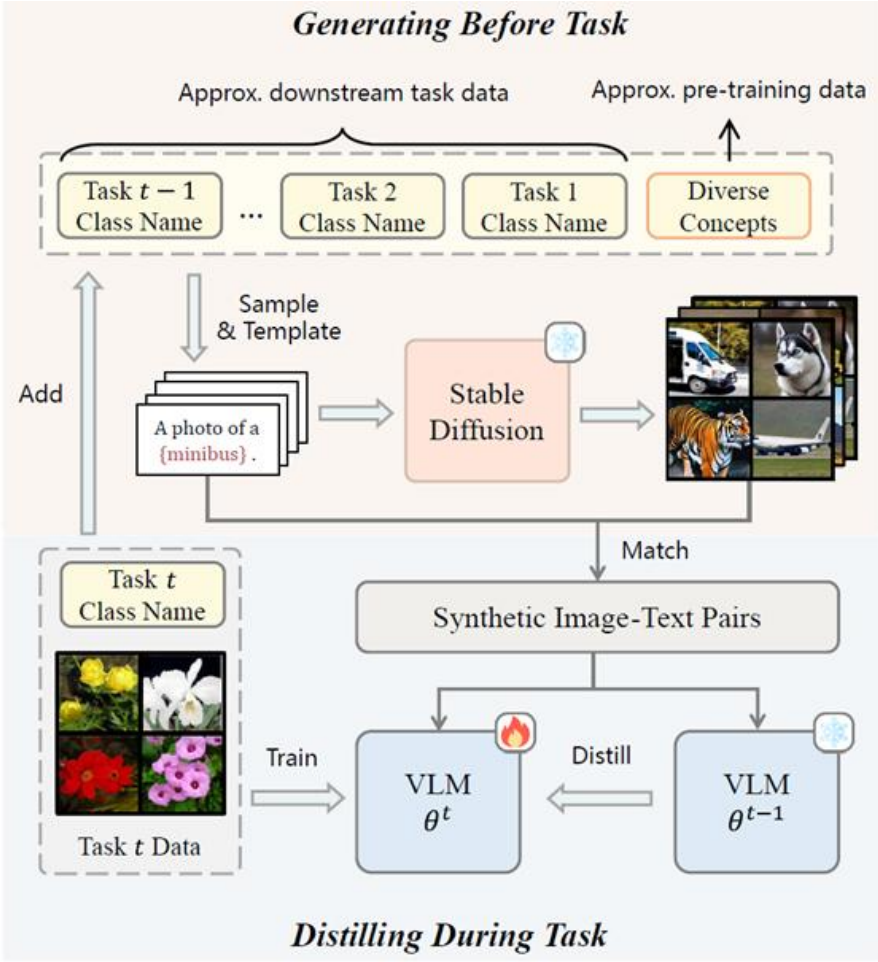
② Aligning model outputs before and after learning new tasks via knowledge distillation is intuitive. However, when the amount of synthetic data is limited, additional regularization is necessary to alleviate overfitting.

Method: Image Generation

Q1: How can diffusion model generate to approximate both pre-training and downstream task data of VLMs?

❑ Generate images from class names

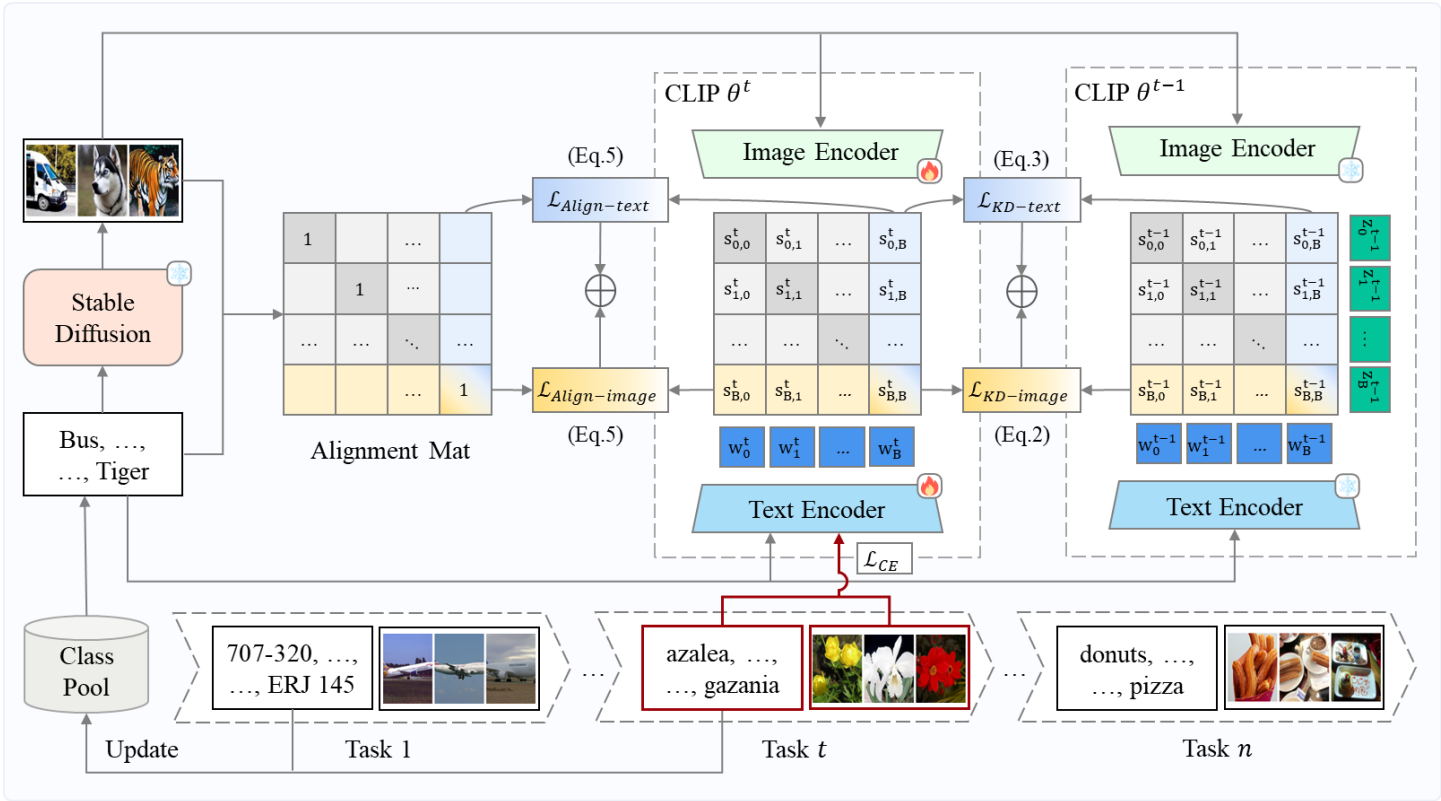
- Maintain a class pool P for class names.
- **Step 1:** Initialize P with C^0 , C^0 is semantically non-overlapping visual concepts sampled ***from different synsets*** to approximate pre-training data.
- **Step 2:** Before learning a new task, sample class names c from P several times and template them with "a photo of a $\{c\}$." to prompt diffusion model to generate images.
- **Step3:** After learning of task t ($t \geq 1$), the class names of task t is added to P, i.e., $P = \cup_{i=0}^{t} C^i$.



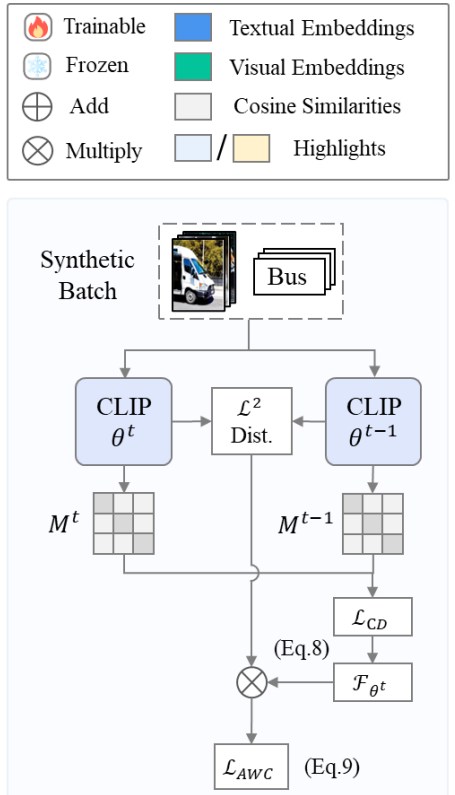
Method: Framework Overview

Q2: How can the generated data be used to mitigate forgetting?

GIFT: Generated data Improves continual Fine-Tuning



(a) Synthetic Data-based Distillation



(b) Adaptive Weight Consolidation

- *Synthetic Data-based Distillation* aligns output of current model with previous model on matching synthetic image-text pairs.
- *Adaptive Weight Consolidation* employs a weighted l_2 penalty to limit parameter changes causing forgetting and overfitting.

Method: Synthetic Data-based Distillation

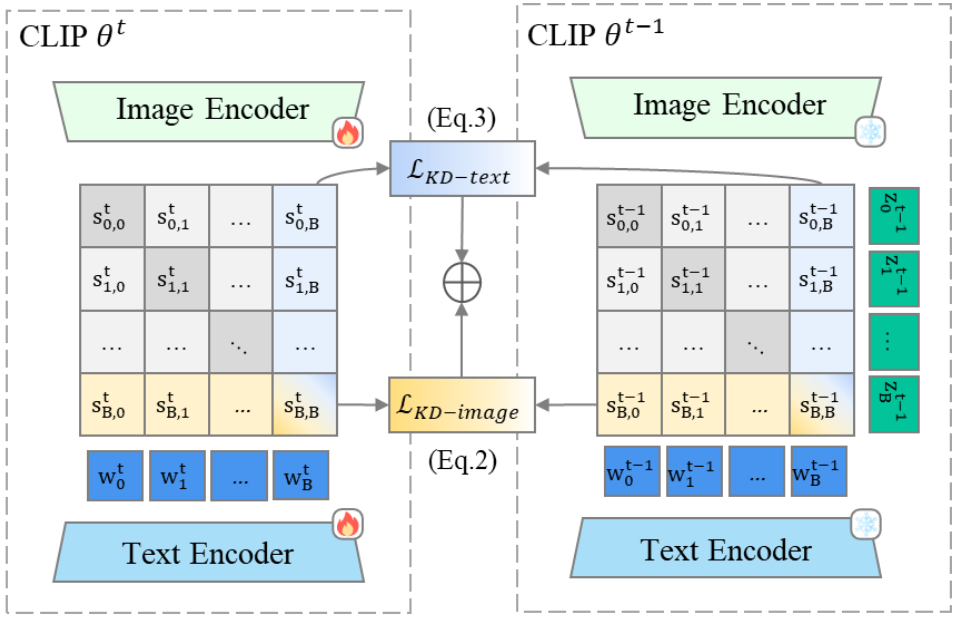
❑ Contrastive Distillation

➢ To better align the modalities, the distillation loss is implement in a contrastive manner similar to CLIP's pre-training objective.

$$\mathcal{L}_{KD\_image} = -\frac{1}{B}\sum_{i=1}^{B} M_{i,:}^{t-1} \cdot \log\left(\frac{M_{i,:}^{t}}{M_{i,:}^{t-1}}\right),$$

$$\mathcal{L}_{KD\_text} = -\frac{1}{B}\sum_{j=1}^{B} M_{:,j}^{t-1} \cdot \log\left(\frac{M_{:,j}^{t}}{M_{:,j}^{t-1}}\right),$$

$$\mathcal{L}_{CD} = \mathcal{L}_{KD\_image} + \mathcal{L}_{KD\_text}.$$



❑ Image-Text Alignment

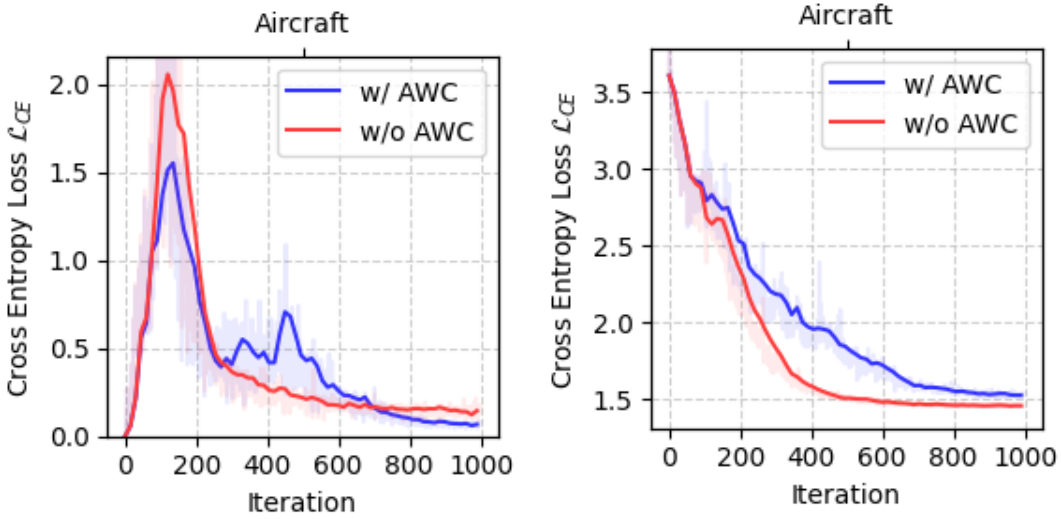
➢ Teacher models also suffer from forgetting. Combining image-text alignment ***hard targets*** with distillation ***soft targets*** helps neutralize error information in the teacher model's outputs.

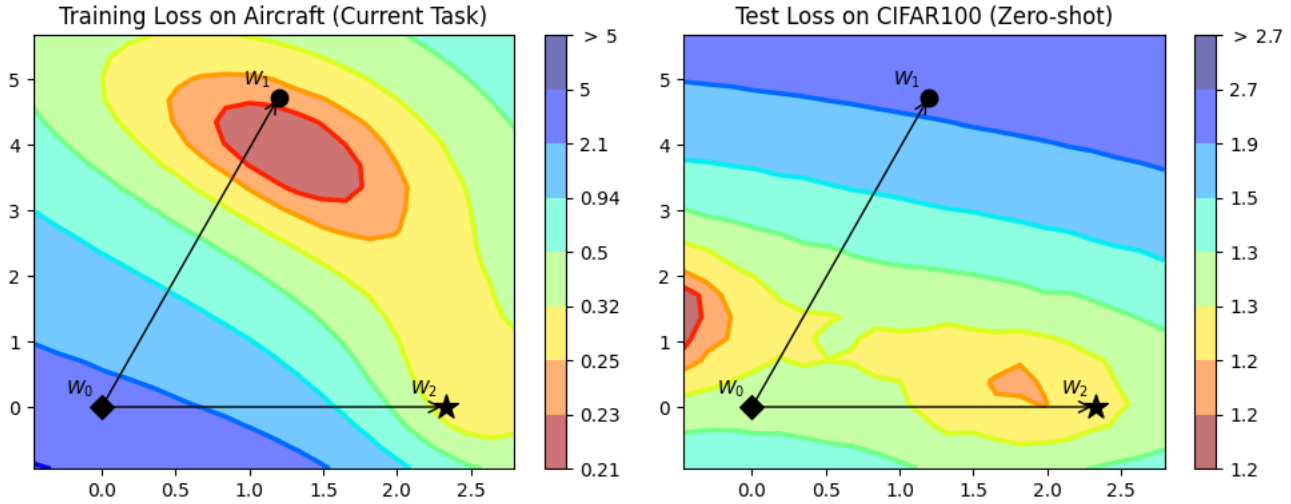
$$\mathcal{L}_{ITA} = \mathcal{L}_{Align\_image} + \mathcal{L}_{Align\_text} = -\frac{1}{B}\sum_{i=1}^{B} I_{i,:} \cdot \log\left(M_{i,:}^{t}\right) + -\frac{1}{B}\sum_{j=1}^{B} I_{:,j} \cdot \log\left(M_{:,j}^{t}\right),$$

$$\mathcal{L}_{Total} = \mathcal{L}_{CE} + (\alpha\mathcal{L}_{CD} + \beta\mathcal{L}_{ITA}).$$

Method: Adaptive Weight Consolidation

❑ Overfitting of Distillation with Limited Synthetic Data





W_0: Initial CLIP W_1: w/o AWC W_2: w/ AWC

❑ Adaptive Weight Consolidation

- A simple l_2 penalty can keep the model in the flat minimum reached during pre-training but greatly sacrifices plasticity.
- We use gradients of the distillation loss to localize parameter updates that cause forgetting and impose a larger penalty.

$$\mathcal{F}_{\theta_i^t}^{(j)} = \left(\frac{\partial(\alpha\mathcal{L}_{KD}^{(j)}+\beta\mathcal{L}_{Align}^{(j)})}{\partial\theta_i^t}\right)^2, \mathcal{L}_{AWC}^{(j)} = \sum_i \mathcal{F}_{\theta_i^t}^{(j)} \cdot \left(\theta_i^t - \theta_i^{t-1}\right)^2.$$

Experiments

❑ Datasets

- ➢ MTIL[2]: Multi-domain Task Incremental Learning
- ➢ 11 datasets from different domains
- ➢ Coarse-grained and fine-grained tasks

Dataset	# classes	# train	# test	Recognition Task
Aircraft	100	3334	3333	aircraft series
Caltech101	101	6941	1736	real-life object
CIFAR100	100	50000	10000	real-life object
DTD	47	1880	1880	texture recognition
EuroSAT	10	21600	5300	satellite location
Flowers	102	1020	6149	flower species
Food	101	75750	25250	food type
MNIST	10	60000	10000	digital number
OxfordPet	37	3680	3669	animal species
StanfordCars	196	8144	8041	car series
SUN397	397	87003	21751	scene category
Total	1201	319352	97109	

❑ Metrics

		Dataset to Test on			
		Task 1	Task 2	.....	Task T
Model to Evaluate	Model 1	$R_{1,1}$	$R_{1,2}$		$R_{1,T}$
	Model 2	$R_{2,1}$	$R_{2,2}$		$R_{2,T}$
	⋮				
	Model T-1	$R_{T-1,1}$	$R_{T-1,2}$		$R_{T-1,T}$
	Task T	$R_{T,1}$	$R_{T,2}$		$R_{T,T}$

Zero shot Learned task Curr. task

Last for final performance and degree of forgetting:

$$Last = \frac{1}{n}\sum_{j=1}^{n} R_{n,j}$$

Transfer for zero-shot transfer ability:

$$Transfer = \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{1}{n-i}\sum_{j=i+1}^{n} R_{i,j}$$

Avg. for stability-plasticity balance:

$$Avg. = \frac{1}{n^2}\sum_{i,j} R_{i,j}$$

Experiments

❑ Main Results

➢ GIFT achieves a balance between learning new knowledge and maintaining general knowledge without raising storage and privacy issues.

Table 1. Comparison of SOTA methods on MTIL Order I.

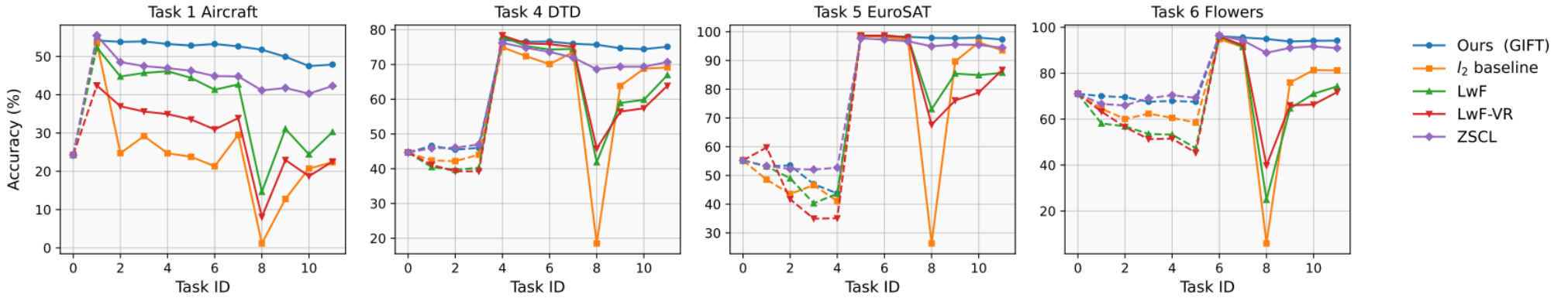
Method	Transfer	Δ	Avg.	Δ	Last	Δ
Zero-shot	69.4	-	65.3	-	65.3	-
Continual Finetune	44.6	-	55.9	-	77.3	-
l_2 baseline	61.0	0.0	62.7	0.0	75.9	0.0
LwF [33]	56.9	-4.1	64.7	+2.0	74.6	-1.3
iCaRL [44]	50.4	-10.6	65.7	+3.0	80.1	+4.2
LwF-VR [11]	57.2	-3.8	65.1	+2.4	76.6	+0.7
WiSE-FT [56]	52.3	-8.7	60.7	-2.0	77.7	+1.8
ZSCL [64]	68.1	+7.1	75.4	+12.7	83.6	+7.7
MoE-Adapter [62]	68.9	+7.9	76.7	+14.0	85.0	+9.1
GIFT (Ours)	**69.3**	**+8.3**	**77.3**	**+14.6**	**86.0**	**+10.1**

Table 2. Comparison of SOTA methods on MTIL Order II.

Method	Transfer	Δ	Avg.	Δ	Last	Δ
Zero-shot	65.4	-	65.3	-	65.3	-
Continual Finetune	46.6	-	56.2	-	67.4	-
l_2 baseline	60.6	0.0	68.8	0.0	77.2	0.0
LwF [33]	53.2	-7.4	62.2	-6.6	71.9	-5.3
iCaRL [44]	50.9	-9.7	56.9	-11.9	71.6	-5.6
LwF-VR [11]	53.1	-7.5	60.6	-8.2	68.3	-3.9
WiSE-FT [56]	51.0	-9.6	61.5	-7.3	72.2	-5.0
ZSCL [64]	64.2	+3.6	74.5	+5.7	83.4	+6.2
MoE-Adapter [62]	64.3	+3.7	74.7	+5.9	84.1	+6.9
GIFT (Ours)	**65.9**	**+5.3**	**75.7**	**+6.9**	**85.3**	**+8.1**

Experiments

❑ Accuracy Curve of Different Tasks (Order I)



❑ Ablation of Distillation Mechanism

➢ The default settings are marked in gray , which employs a contrastive distillation loss, the last CLIP model as the teacher model, and $\beta = 0.25$ for ITA scale.

(a) **Distillation Loss.**

Loss	Transfer	Avg.	Last
Feat. Dist.	64.0	71.6	80.5
Image-only	66.8	75.1	84.1
Text-only	64.7	71.9	81.8
Contrastive	**68.9**	**76.6**	**85.0**

(b) **Teacher Model.**

Teacher	Transfer	Avg.	Last
Initial CLIP	69.1	74.0	80.1
Last CLIP	68.9	**76.6**	**85.0**
WiSE(0.2)	69.1	76.1	83.4
WiSE(0.5)	**69.6**	75.3	81.6

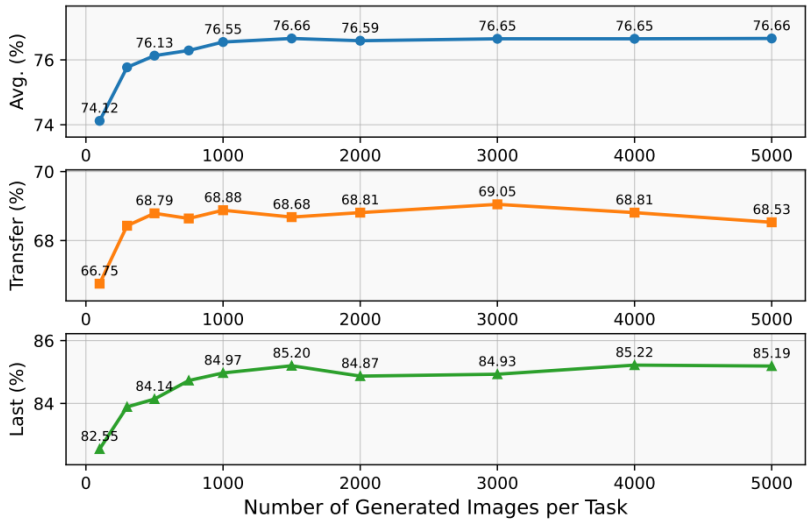
(c) **Scale of Image-Text Alignment.**

ITA Scale	Transfer	Avg.	Last
$\beta = 0.0$	68.3	76.3	84.7
$\beta = 0.25$	**68.9**	**76.6**	**85.0**
$\beta = 0.5$	68.7	76.2	84.2
$\beta = 1.0$	68.5	75.4	82.4

Experiments

❑ Ablation of Image Generation

- Generating 1k per task yields stable performance.



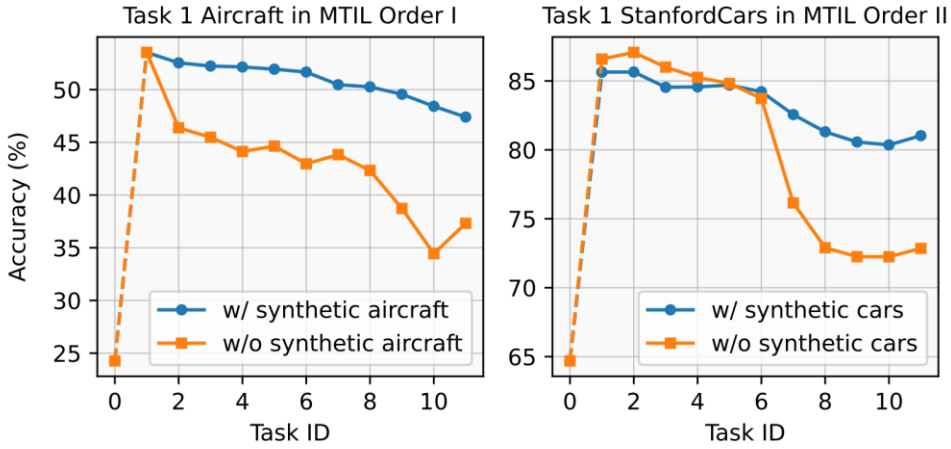
- Compatible with fewer denoising steps and faster generation

Method	Denoising Steps	Transfer	Avg.	Last
GIFT w/ AWC	50 Steps	**69.3**	**77.3**	**86.0**
GIFT w/o AWC	50 Steps	68.9	76.6	85.0
GIFT w/ AWC	25 Steps	69.2	77.2	85.8
GIFT w/o AWC	25 Steps	69.2	76.6	84.8

- Not sensitive to guidance scale value.

Guidance Scale	Image Num	Transfer	Avg.	Last
small	1K	68.2	76.3	85.2
medium		68.9	76.6	85.0
large		68.5	76.3	85.1
small	3K	68.7	76.8	85.0
medium		69.1	76.7	84.9
large		68.8	76.6	85.1

- Eliminating synthetic images for specific downstream tasks exacerbates forgetting of these tasks.



Thank you !

Wu Bin

2025/05/15