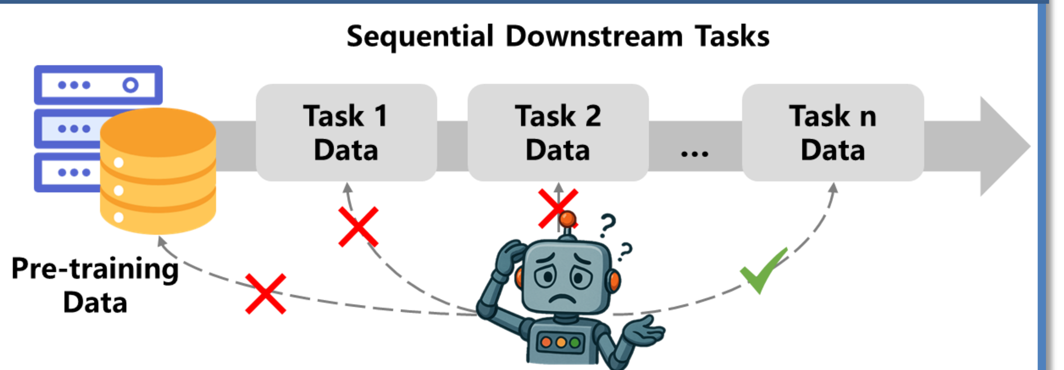
Synthetic Data is an Elegant GIFT for Continual Vision-Language Models

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

Introduction

Background

❑ When sequentially fine-tuned on multiple downstream tasks, pre-trained vision-language models (VLMs) suffer from severe catastrophic forgetting.



Motivation

❑ Replay-based methods are impractical when pre-training data is unavailable and storing historical data raises privacy concerns.

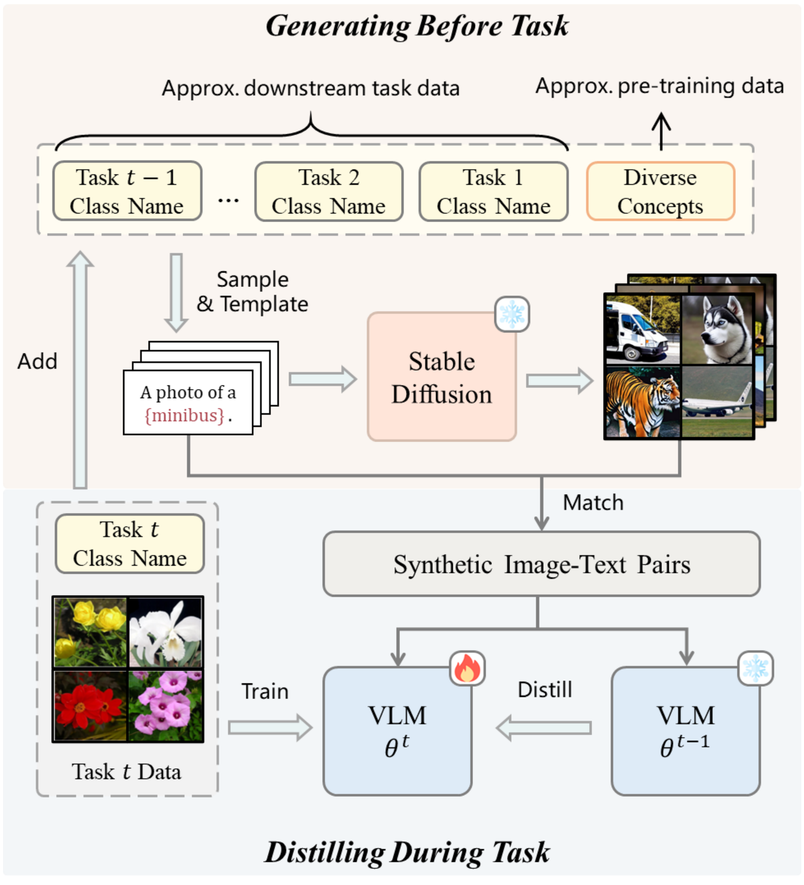
❑ Synthetic data from latest diffusion models is ready for supplement when training data is scarce.

We Want to Explore

When direct access to historical data is not allowed, can synthetic data help preserve VLM's knowledge during continual learning?

Q1: How to generate?

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



Generate Images from Class Names

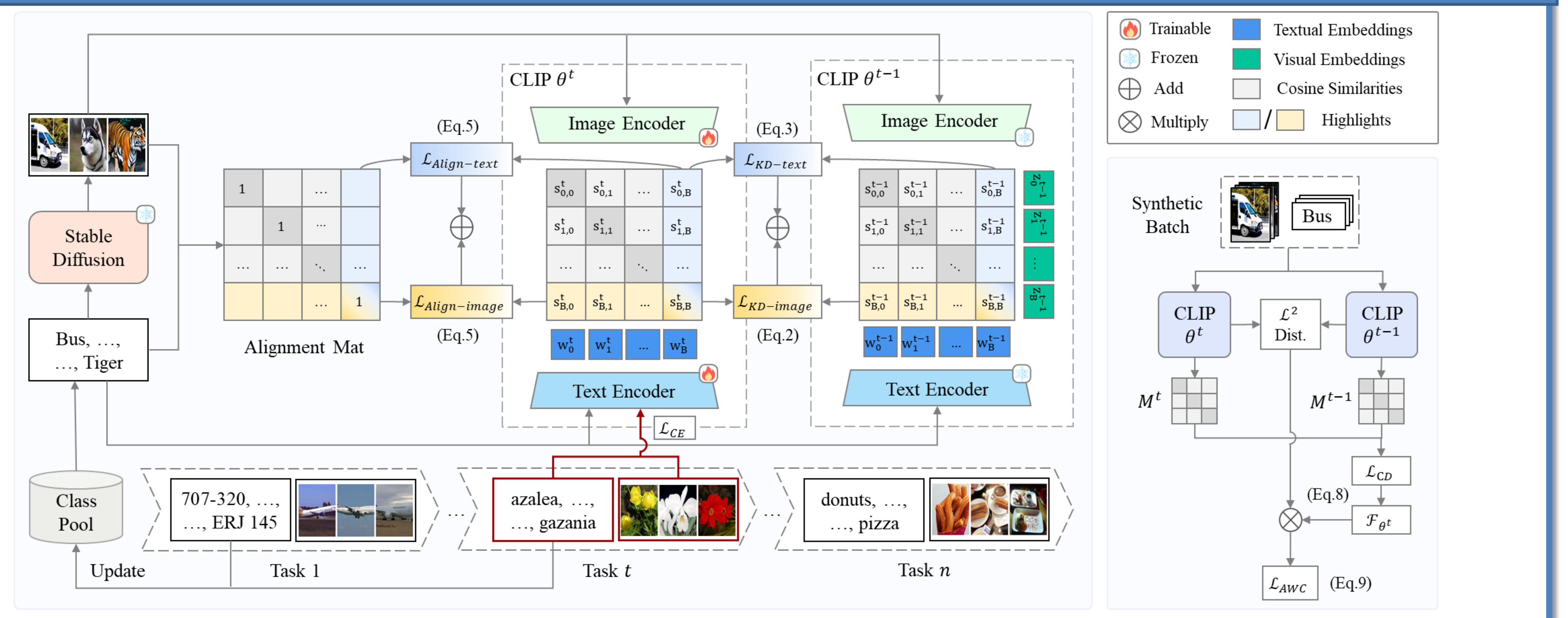
❑ **Step 1**: Start with a pool P of base class names C^0: diverse, non-overlapping visual concepts from different synsets.

❑ **Step 2**: Before task t, sample class names c from P and format prompts for generation: "a photo of a {c}".

❑ **Step 3**: After task t, add its class names C^t to P: $P = \cup_{i=0}^{t} C^i$.

Q2: How to use the generated data to mitigate forgetting?

GIFT: Generated data Improves continual Fine-Tuning



(a) Synthetic Data-based Distillation (b) Adaptive Weight Consolidation

Synthetic Data-based Distillation

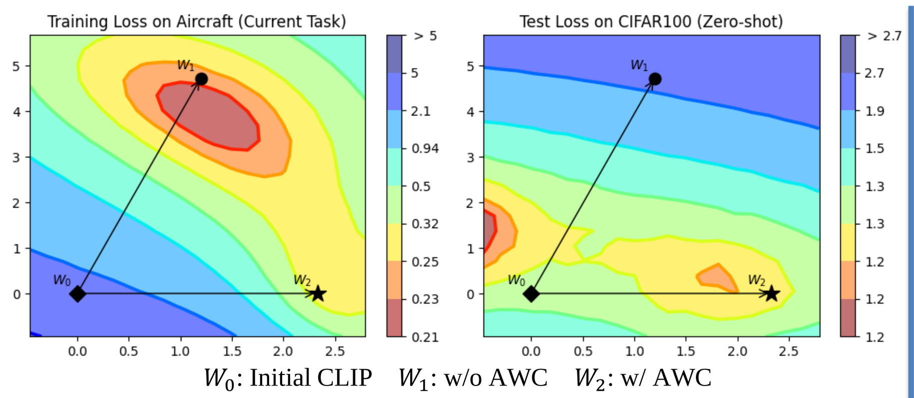
① **Contrastive Distillation:** To align the modalities better, the distillation loss is implemented in a contrastive manner similar to CLIP's pre-training objective:

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

② **Image-Text Alignment:** Combining image-text alignment hard targets with distillation soft targets to neutralize error information in teacher model's outputs:

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

Adaptive Weight Consolidation



An illustration of overfitting from the perspective of the loss surface

❑ Overfitting occurs when the amount of synthetic data is limited.

❑ We use a Fisher information weighted l_2 penalty to mitigate overfitting without sacrificing plasticity.

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

Experiments

Comparison to SOTA

❑ We conduct experiments and achieves SOTA on the MTIL benchmark, which spans 11 datasets across different domains.

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

Ablation of Distillation Mechanism

(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.**

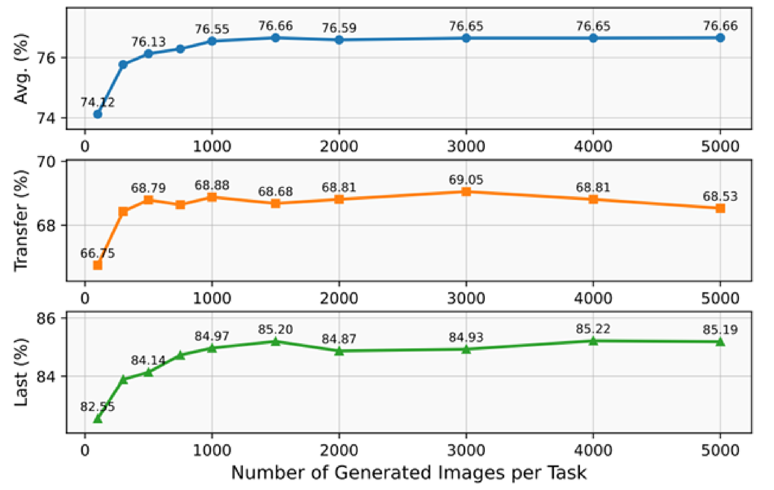
Teacher	Transfer	Avg.	Last
Initial CLIP	69.1	74.0	80.1
Last CLIP	68.9	**76.6**	**85.0**
WiSE(0.2)	69.1	76.1	83.4
WiSE(0.5)	**69.6**	75.3	81.6

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

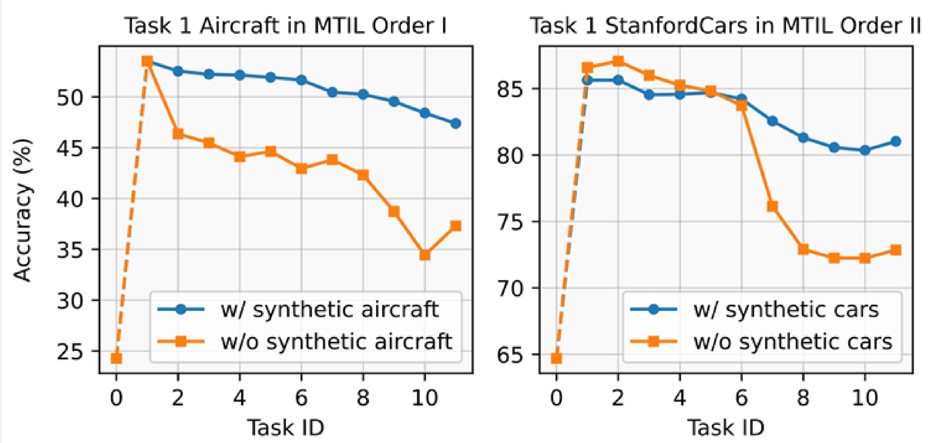
ITA Scale	Transfer	Avg.	Last
$\beta = 0.0$	68.3	76.3	84.7
$\beta = 0.25$	**68.9**	**76.6**	**85.0**
$\beta = 0.5$	68.7	76.2	84.2
$\beta = 1.0$	68.5	75.4	82.4

Ablation of Image Generation

❑ Generating 1k per task yields stable performance.



❑ Removing task-specific synthetic data worsens forgetting.



❑ 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