Deep Learning for Computer Vision

113-1/Fall 2024

https://cool.ntu.edu.tw/courses/41702 (NTU COOL)

http://vllab.ee.ntu.edu.tw/dlcv.html (Public website)

Yu-Chiang Frank Wang 王鈺強, Professor

Dept. Electrical Engineering, National Taiwan University

2024/11/5

Pretrain & Finetune LLM/VLM/MLLM

Stage 1

Pre-training by
self-supervised learning
or supervised learning

Stage 2

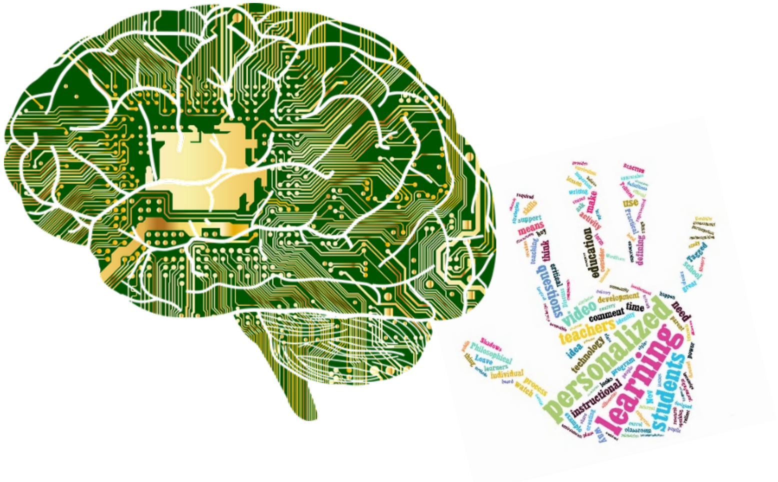
Finetuning
by downstream tasks
in target domains

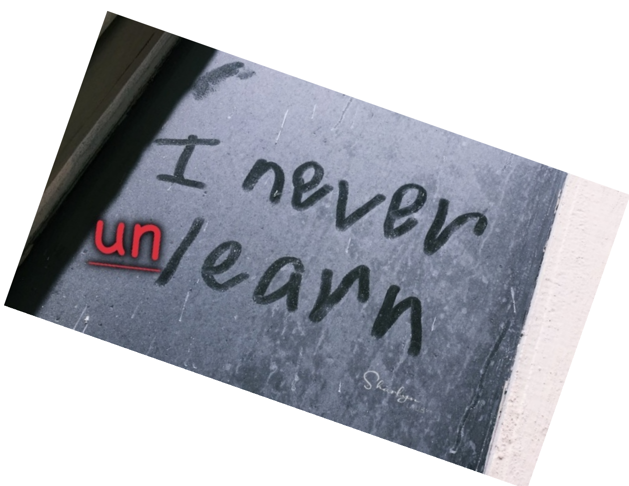
Stage 3

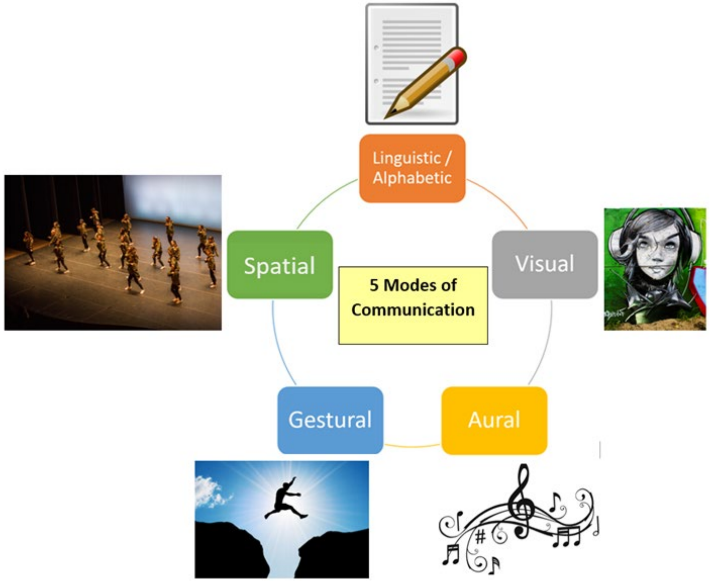
RLHF - Reinforcement Learning
with Human Feedback
(not covered)

What to Be Covered Today…

- **Multimodal LLM**
- **Advanced Topics in LLM/VLM**
  - Concept Editing
  - Concept Unlearning
  - Personalization
  - Continual Learning



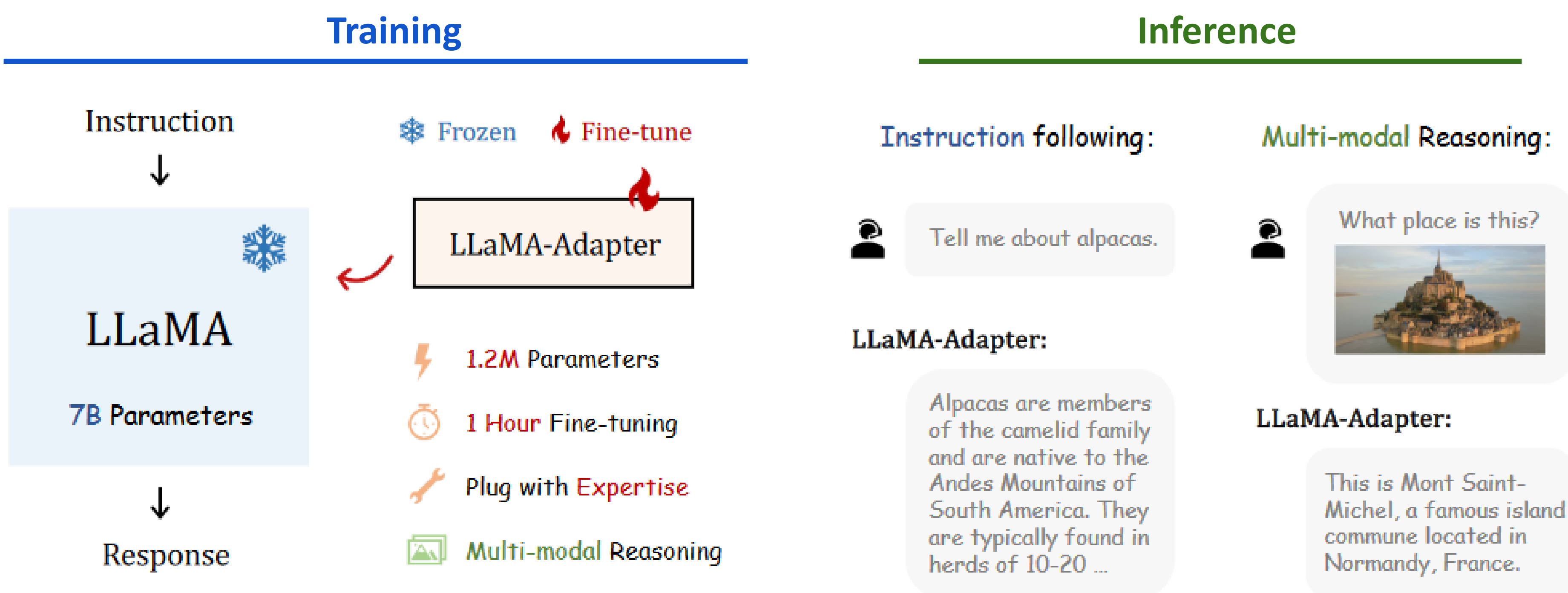




LLaMA-Adapter (SAIL & CUHK, ICLR'24)

- A **lightweight** adaptation method to efficiently finetune LLaMA into an **multi-modal instruction-following** model
- **Result:**

  1. Equipping LLaMA the ability of understanding instruction-following data
  2. Capable of addressing multi-modal reasoning tasks



[LLaMA-Adapter: Efficient Fine-tuning of Large Language Models with Zero-initialized Attention]

LLaMa-Adapter: *Method*

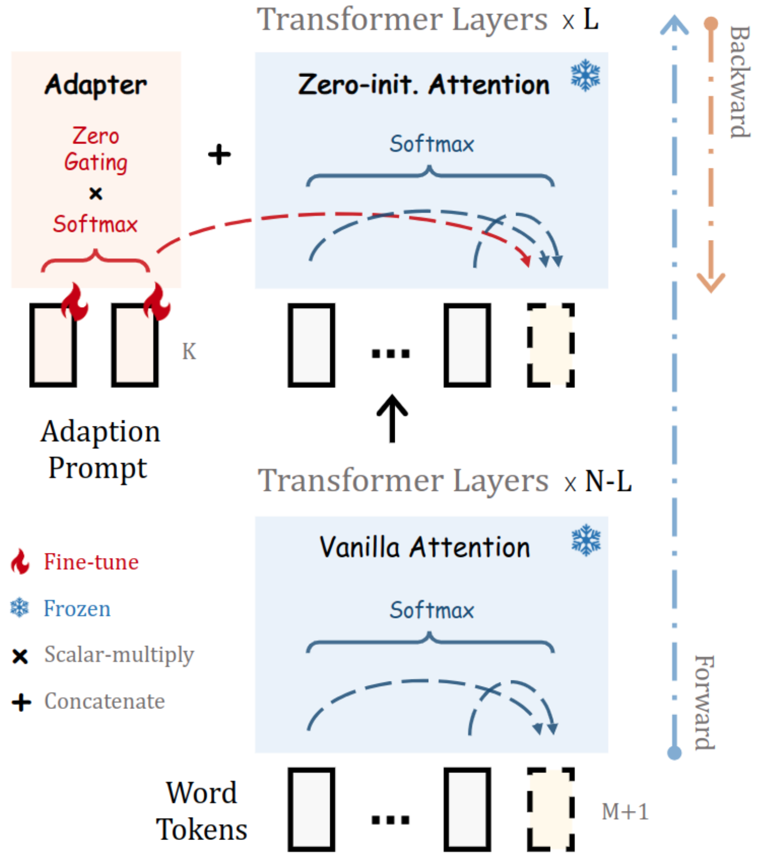
- Append K learnable prompts at each of the last L layers
- Adopt **zero-initialized attention**
  for fine-tuning stability and effectiveness
  - use a zero-initialized learnable gating factor g_l
    to control the prompts' attention scores

$$S_l = Q_l K_l^T / \sqrt{C} \ \in \mathbb{R}^{1\times(K+M+1)}$$

$$S_l^g = [\mathrm{softmax}(S_l^K)\cdot g_l;\ \ \mathrm{softmax}(S_l^{M+1})]^T$$

Stage 1

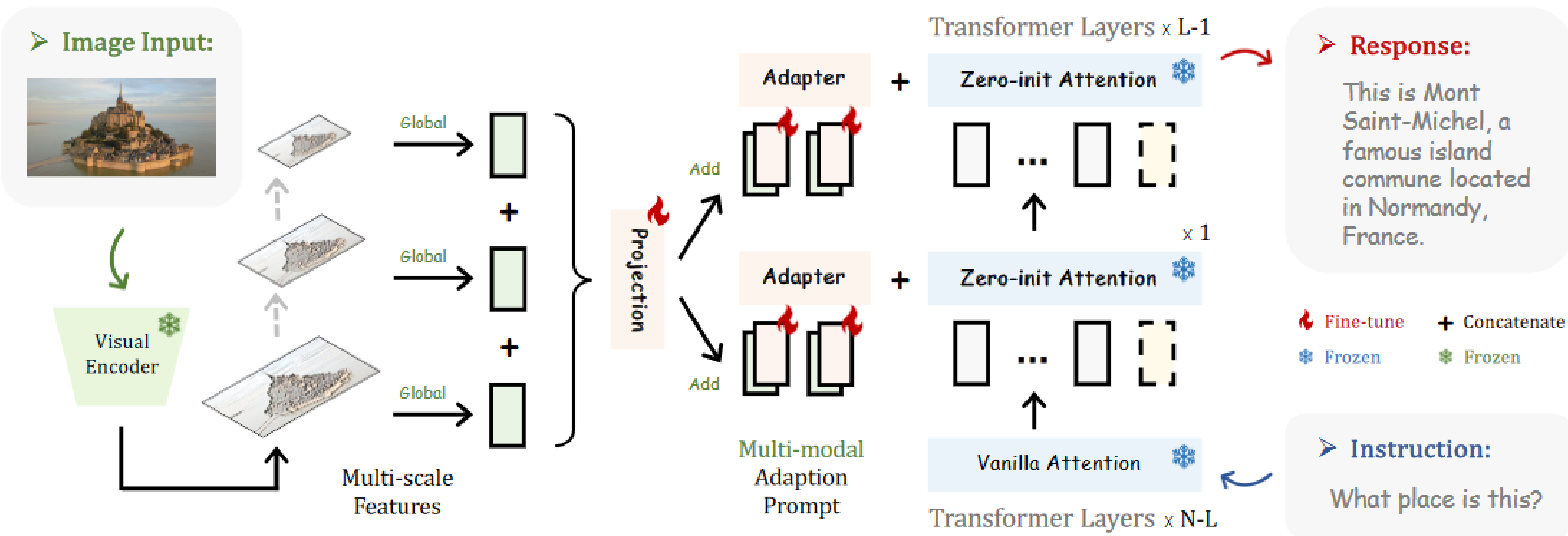
Stage 2



[LLaMA-Adapter: Efficient Fine-tuning of Large Language Models with Zero-initialized Attention]

Multi-Modal LLaMA-Adaptor

- When being finetuned to a multi-modal reasoning task (e.g. VQA), a pre-trained visual encoder + a learnable projection layer are additionally utilized



[LLaMA-Adapter: Efficient Fine-tuning of Large Language Models with Zero-initialized Attention]

Quantitative Result

- Methods comparison on the visual question answering (VQA) task (ScienceQA dataset)

Model	Tuned Params	Avg	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12
Random Choice [41]	-	39.83	40.28	46.13	29.25	47.45	40.08	33.66	39.35	40.67
Human [41]	-	88.40	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42
MCAN [65]	95M	54.54	56.08	46.23	58.09	59.43	51.17	55.40	51.65	59.72
VisualBERT [33, 34]	111M	61.87	59.33	69.18	61.18	62.71	62.17	58.54	62.96	59.92
UnifiedQA [27]	223M	70.12	68.16	69.18	74.91	63.78	61.38	77.84	72.98	65.00
UnifiedQA$_{CoT}$	223M	74.11	71.00	76.04	78.91	66.42	66.53	81.81	77.06	68.82
GPT-3 [4]	0M	74.04	75.04	66.59	78.00	74.24	65.74	79.58	76.36	69.87
GPT-3$_{CoT}$	0M	75.17	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68
ChatGPT$_{CoT}$ [2]	0M	78.31	78.82	70.98	83.18	77.37	67.92	86.13	80.72	74.03
GPT-4$_{CoT}$ [45]	0M	83.99	85.48	72.44	90.27	82.65	71.49	92.89	86.66	79.04
MM-COT$_T$ [74]	223M	70.53	71.09	70.75	69.18	71.16	65.84	71.57	71.00	69.68
MM-COT	223M	84.91	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37
LLaMA-Adapter$_T$	**1.2M**	**78.31**	**79.00**	**73.79**	**80.55**	**78.30**	**70.35**	**83.14**	**79.77**	**75.68**
LLaMA-Adapter	**1.8M**	**85.19**	**84.37**	**88.30**	**84.36**	**83.72**	**80.32**	**86.90**	**85.83**	**84.05**

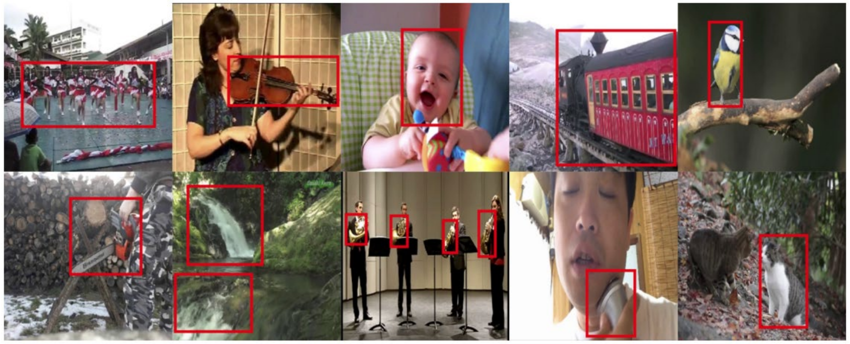
metrics:
accuracy (%)

Easy Visual Sound Localization (EZ-VSL), CMU & UWisc., ECCV 2022

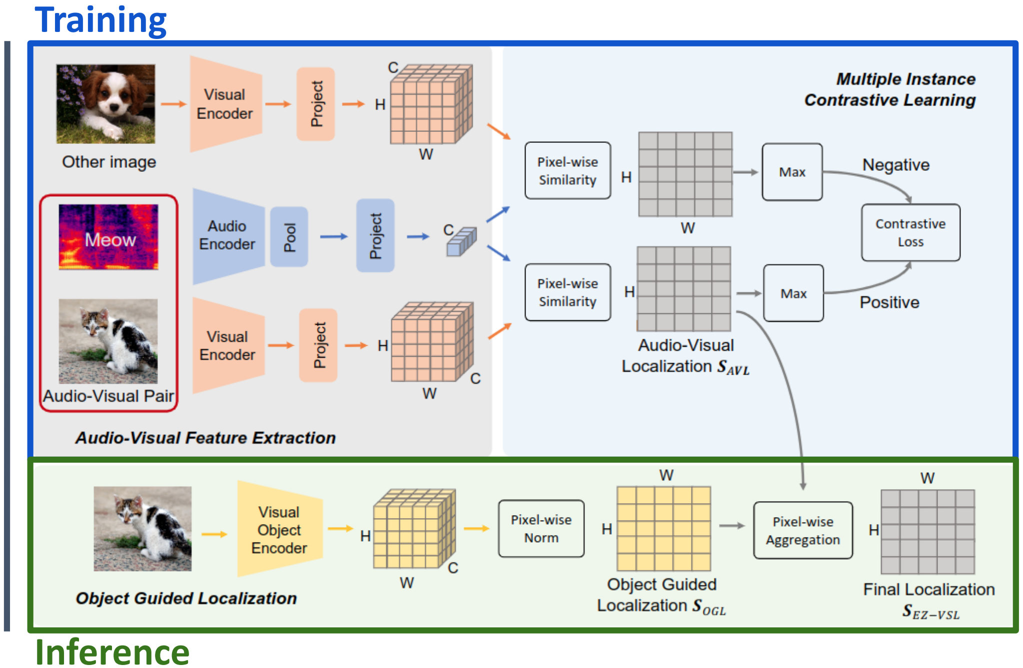
Stage 1

Stage 2

- **Goal:**
  A simple yet effective method to **unsupervised audio-visual sound localization**
  - Task: localize visible sound sources in a video without ground-truth localization
- **Result:** SOTA performance on two popular benchmarks, Flickr SoundNet & VGG-Sound Source



(a) audio-visual sound localization examples

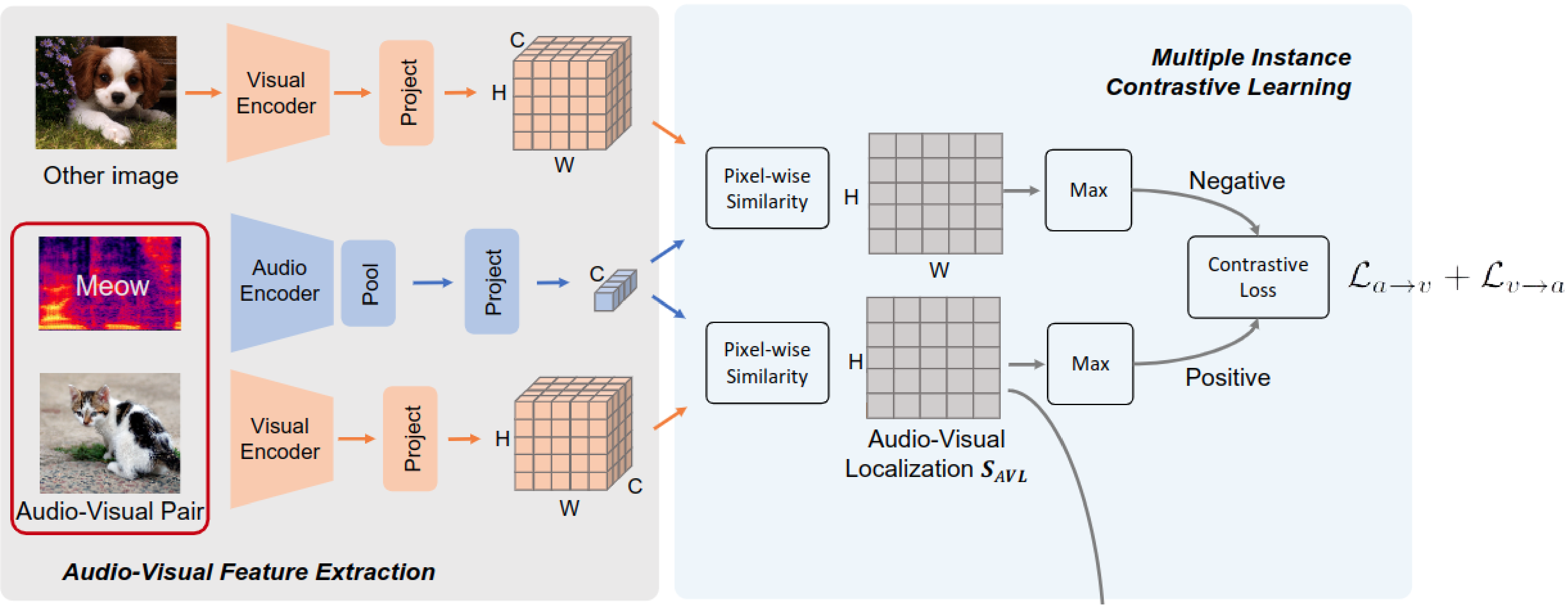


(b) EZ-VSL framework

[Localizing Visual Sounds the Easy Way](#)

Training Stage

- **Audio-Visual Matching by Multiple-Instance Contrastive Learning**
  - Encourage audio representation to be aligned with the associate visual representations **at least at one location** and not being associated with any locations from other images

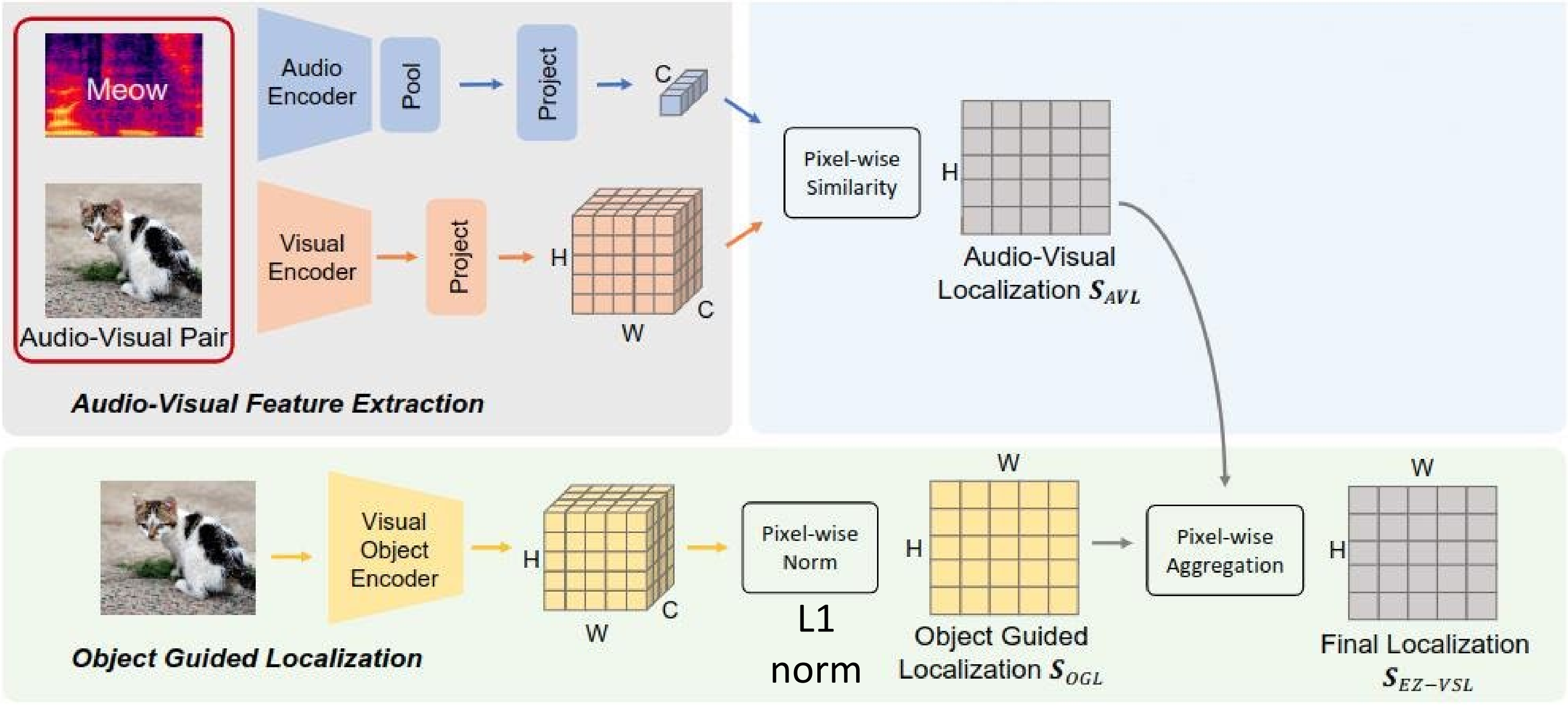


$$\mathcal{L}_{a\to v} = -\log \frac{\exp\left(\frac{1}{\tau}\max_{\hat{\mathbf{v}}\in V_i}\mathtt{sim}(\hat{\mathbf{a}}_i, \hat{\mathbf{v}})\right)}{\sum_k \exp\left(\frac{1}{\tau}\max_{\hat{\mathbf{v}}\in V_k}\mathtt{sim}(\hat{\mathbf{a}}_i, \hat{\mathbf{v}})\right)} \quad \mathcal{L}_{v\to a} = -\log \frac{\exp\left(\frac{1}{\tau}\max_{\hat{\mathbf{v}}\in V_i}\mathtt{sim}(\hat{\mathbf{v}}, \hat{\mathbf{a}}_i)\right)}{\sum_k \exp\left(\frac{1}{\tau}\max_{\hat{\mathbf{v}}\in V_i}\mathtt{sim}(\hat{\mathbf{v}}, \hat{\mathbf{a}}_k)\right)}$$

Localizing Visual Sounds the Easy Way

Inference Stage

- **Object-Guided Localization**
  - Improve localization precision by combining audio-visual similarity map with an **object localization map** from a pre-trained visual model



$$S_{EZ-VSL} = \alpha S_{AVL} + (1-\alpha) S_{OGL}$$

Localizing Visual Sounds the Easy Way

Quantitative Results

- Comparison results on Flickr SoundNet testset

Training set	Method	CIoU(%)	AUC(%)
Flickr 10k	Attention10k [30]	43.60	44.90
	CoarsetoFine [28]	52.20	49.60
	AVObject [2]	54.60	50.40
	LVS [6]	58.20	52.50
	EZ-VSL (ours)	**81.93**	**62.58**
Flickr 144k	Attention10k [30]	66.00	55.80
	DMC [17]	67.10	56.80
	LVS [6]	69.90	57.30
	HardPos [31]	75.20	59.70
	EZ-VSL (ours)	**83.13**	**63.06**

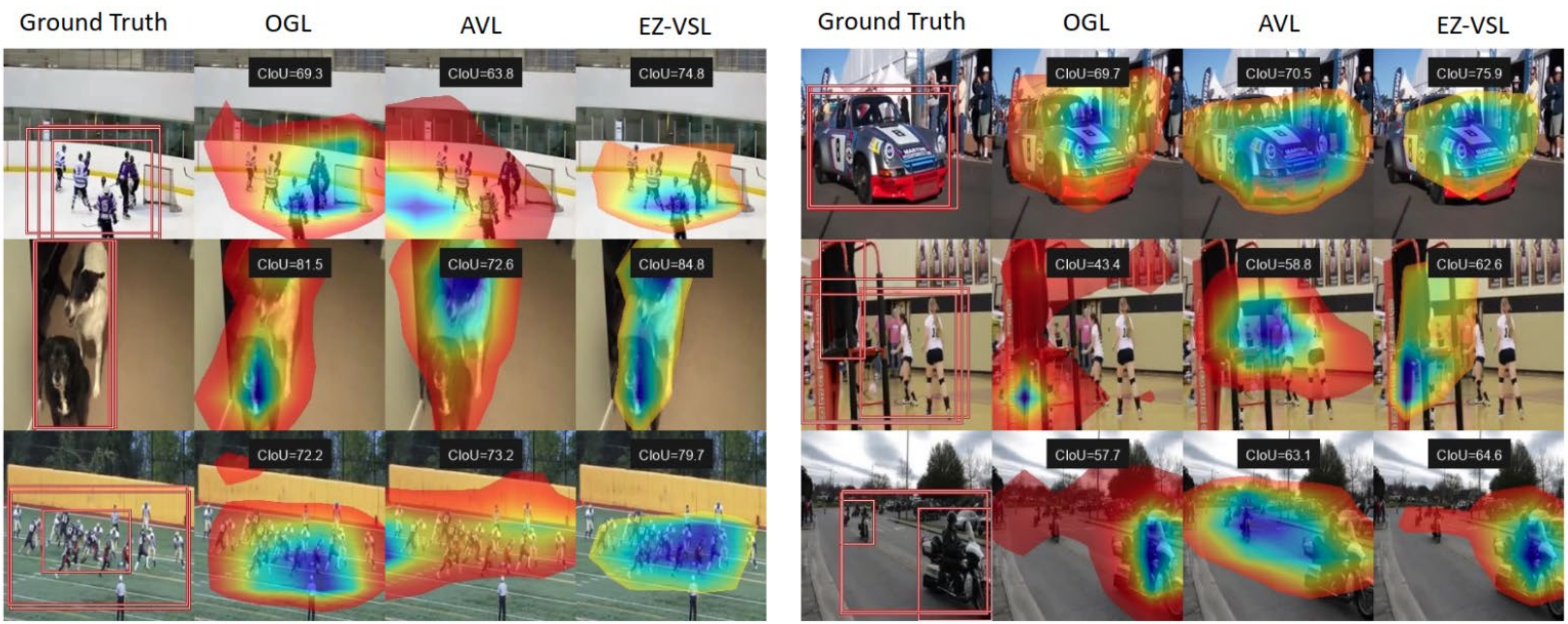
- Methods comparison on Flickr SoundNet and VGG-SS testset

Training set	Method	Flickr-SoundNet		VGG-SS	
		CIoU(%)	AUC(%)	CIoU(%)	AUC(%)
VGG-Sound 144k	Attention10k [30]	66.00	55.80	18.50	30.20
	CoarsetoFine [28]	-	-	29.10	34.80
	AVObject [2]	-	-	29.70	35.70
	LVS [6]	73.50	59.00	34.40	38.20
	HardPos [31]	76.80	59.20	34.60	38.00
	EZ-VSL (ours)	**83.94**	**63.60**	**38.85**	**39.54**

[Localizing Visual Sounds the Easy Way]

Visualization

- Sound source localization predictions with other methods



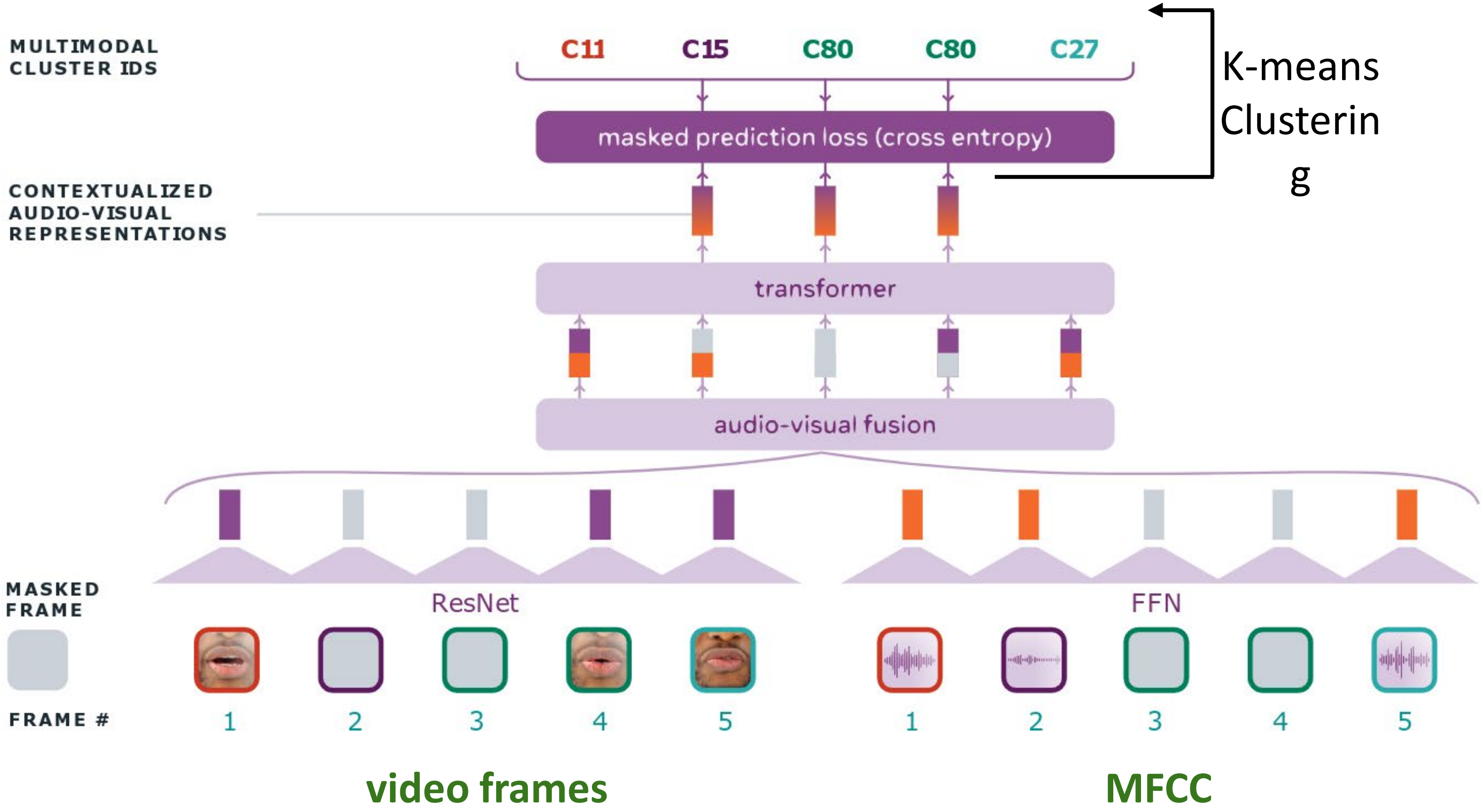
Localizing Visual Sounds the Easy Way

AV-HuBERT, Meta, ICLR'22


Stage 1

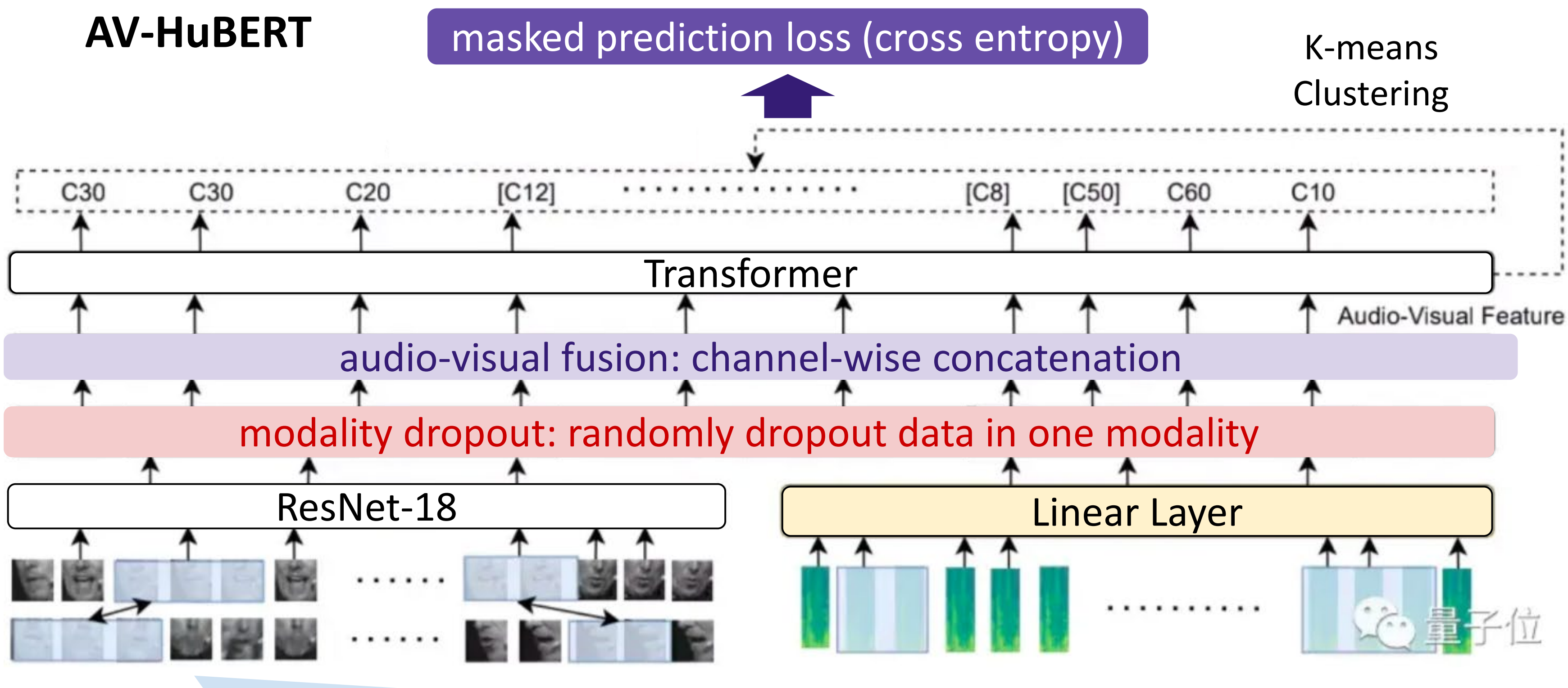

Stage 2

- **Goal:** A **self-supervised** representation learning framework for **audio-visual speech recognition**
- **Result:** SOTA performance on the largest downstream lip-reading benchmark



[Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction]()

Pre-training Framework



Masking by Substitution
Goal: enforce the model to learn the temporal relationship between the frames
Method: replace the masked segments with random segments from the same video

https://zhuanlan.zhihu.com/p/455426545

[Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction](#)

Quantitative Result

- Method comparison on lip-reading benchmark (LRS3 dataset)

Method	Backbone	Criterion	Labeled iso (hrs)	Labeled utt (hrs)	Unlabeled data (hrs)	WER (%)
Supervised						
Afouras et al. (2020)	CNN	CTC	157	433	-	68.8
Zhang et al. (2019b)	CNN	S2S	157	698	-	60.1
Afouras et al. (2018a)	Transformer	S2S	157	1,362	-	58.9
Xu et al. (2020)	RNN	S2S	157	433	-	57.8
Shillingford et al. (2019)	RNN	CTC	-	3,886	-	55.1
Ma et al. (2021b)	Conformer	CTC+S2S	-	433	-	46.9
Ma et al. (2021b)	Conformer	CTC+S2S	157	433	-	43.3
Makino et al. (2019)	RNN	Transducer	-	31,000	-	**33.6**
Semi-Supervised & Self-Supervised						
Afouras et al. (2020)	CNN	CTC	157	433	334	59.8
Ma et al. (2021a)†	Transformer-BASE	S2S	-	30	433	71.9
			-	433	1,759	49.6
Proposed (Self-Supervised & Self-Supervised + Semi-Supervised)						
AV-HuBERT	Transformer-BASE	S2S	-	30	-	94.3
			-	30	433	51.8
			-	30	1,759	**46.1**
			-	433	-	60.3
			-	433	433	44.0
			-	433	1,759	**34.8**
	Transformer-LARGE	S2S	-	30	-	92.3
			-	30	433	44.8
			-	30	1,759	**32.5**
			-	433	-	62.3
			-	433	433	41.6
			-	433	1,759	**28.6**
AV-HuBERT + Self-Training	Transformer-LARGE	S2S	-	30	1,759	**28.6**
			-	433	1,759	**26.9**

CTC: connectionist temporal classification

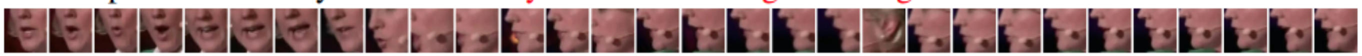
S2S: attention-based sequence-to-sequence cross entropy

Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction

Qualitative Result

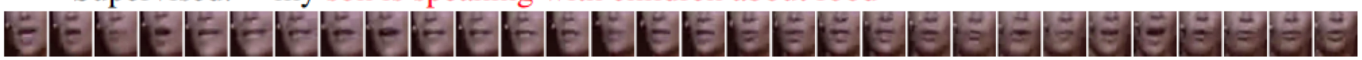
- Transcriptions comparison between AV-HuBERT (Proposed) and a supervised model

(1)	GT:	why not ask all of the states to do that instead
	Proposed:	why not ask all of these things to do that instead
	Supervised:	why can't i actually all of these things do things and

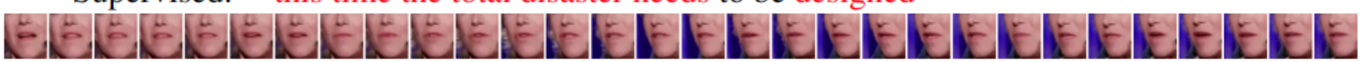


(2)	GT:	indeed we run the risk of making things worse
	Proposed:	indeed we want the risk of making things worse
	Supervised:	in india we roughly receive money in the health world

(3)	GT:	my desire to disappear was still very powerful
	Proposed:	my desire to disappear was still very powerful
	Supervised:	my son is speaking with children about food



(4)	GT:	the silent majority does not need to be silent
	Proposed:	the same majority does not need to be silent
	Supervised:	this time the total disaster needs to be designed



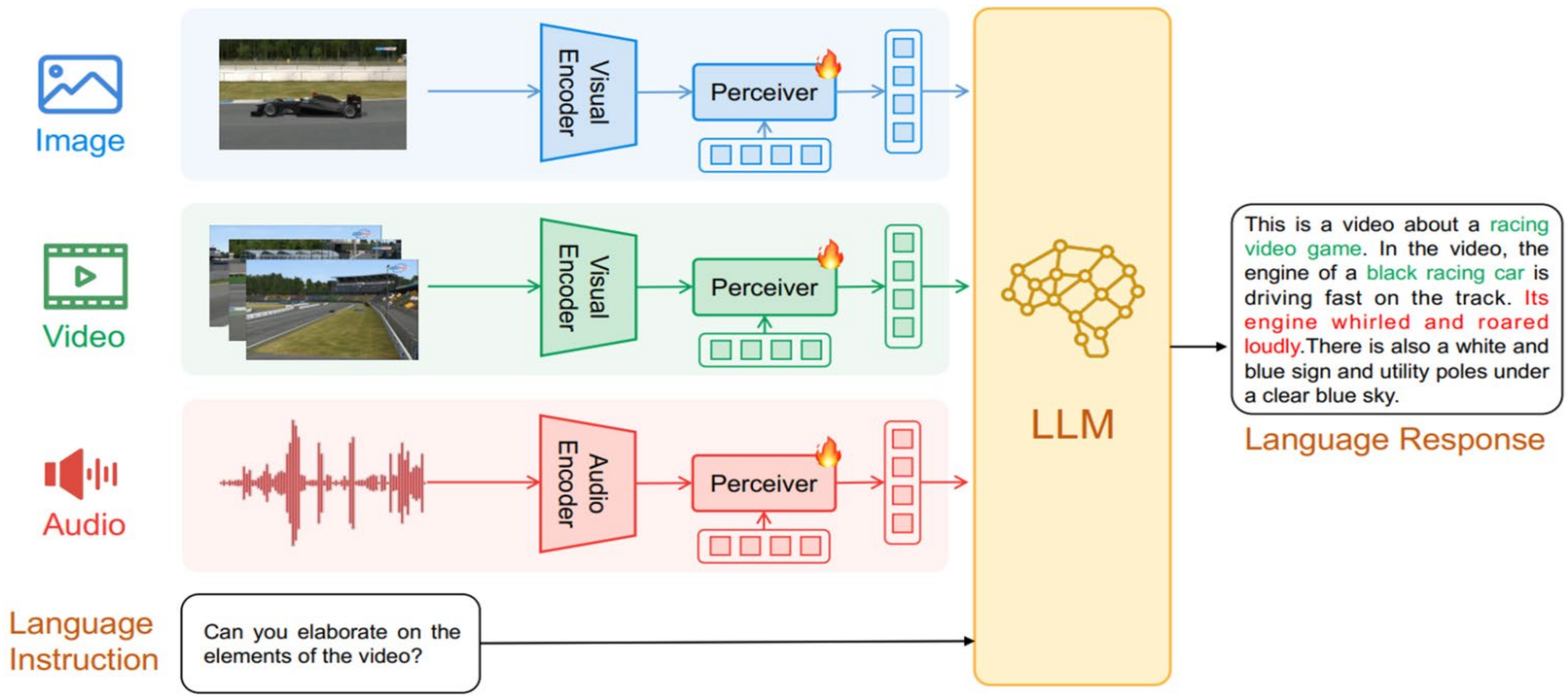
(5)	GT:	mortality is not going down it's going up
	Proposed:	mortality is not going down it's going up
	Supervised:	we're seeing this not only carrying slowly how

[Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction]

ChatBridge (CAS, arxiv'23)

- **Goal:** Transform a LLM into a **multi-modal (text, image, video, audio) language model** using **language** to bridge the gap between various modalities
- **Result:** Outstanding zero-shot performance on various multi-modal tasks

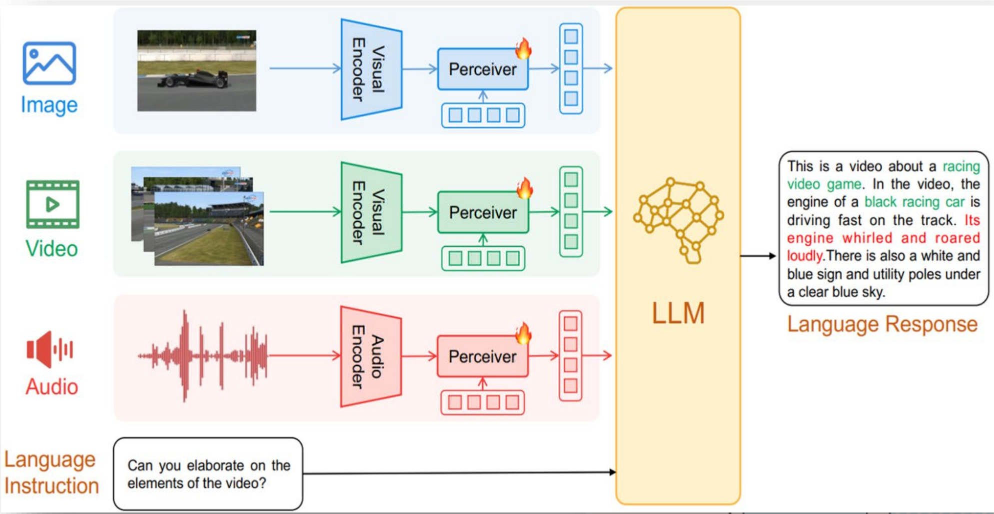
modality-shared perceiver (Q-Former)
but modality-independent learnable queries



ChatBridge: Bridging Modalities with Large Language Model as a Language Catalyst

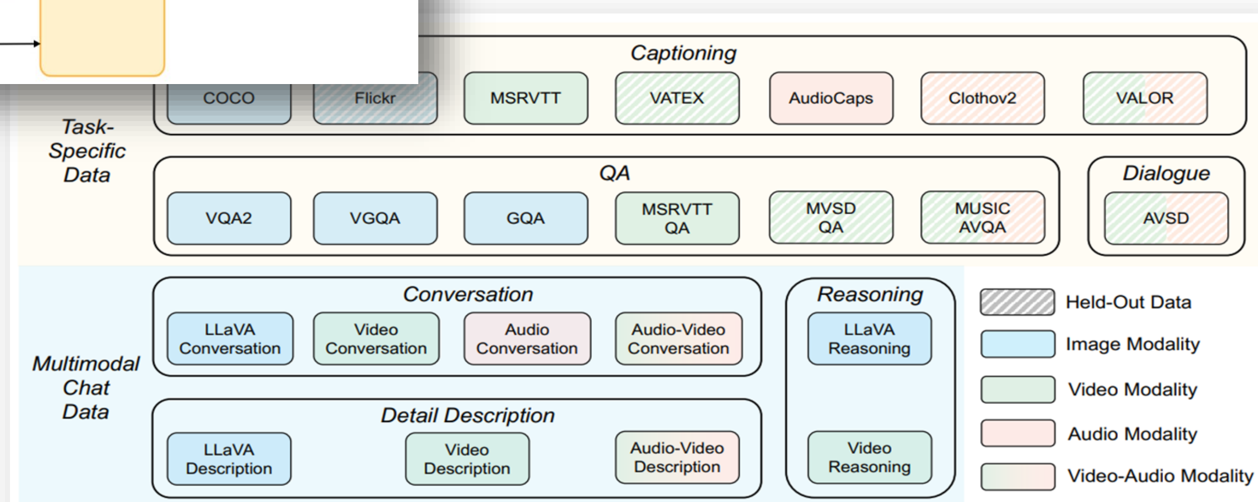
Pre-training Strategy

- **Two-stage** Pre-training
  - Stage 1 multi-modal alignment: pre-train **perceivers** with large-scale language-paired two-modality data (image-text, video-text, audio-text)
  - Stage 2 multi-modal instruction tuning: instruction-finetune **perceivers** with the proposed multi-modal instruction-tuning dataset MULTIS



Stage 1

Stage 2



MULTIS dataset

Quantitative Result

- Zero-shot evaluation of SOTA methods on various downstream tasks

Methods	Image-Text Tasks OKVQA QA	GQA QA	Flickr30k Caption	NoCaps Caption	Video-Text Tasks MSVD QA	VATEX Caption	Audio-Text Tasks Clothov2 Caption
Finetuned SoTA	66.1 [18]	65.1 [70]	67.4 [72]	121.6 [31]	60.0 [12]	95.8 [12]	48.8 [38]
Flamingo-9B [3]	44.7	-	61.5	-	30.2	39.5	-
Flamingo-80B [3]	**50.6**	-	67.2	-	35.6	46.7	-
BLIP-2 (FlanT5-XXL) [17]	-	**42.4**	73.7	98.4	34.4	-	-
BLIP-2 (Vicuna-13B) [17]	-	32.3	71.6	103.9	20.3	-	-
ChatBridge w/o MULTIS	41.4	37.4	77.7	107.5	23.5	47.7	22.4
ChatBridge	45.2	41.8	**82.5**	**115.7**	**45.3**	**48.9**	**26.2**

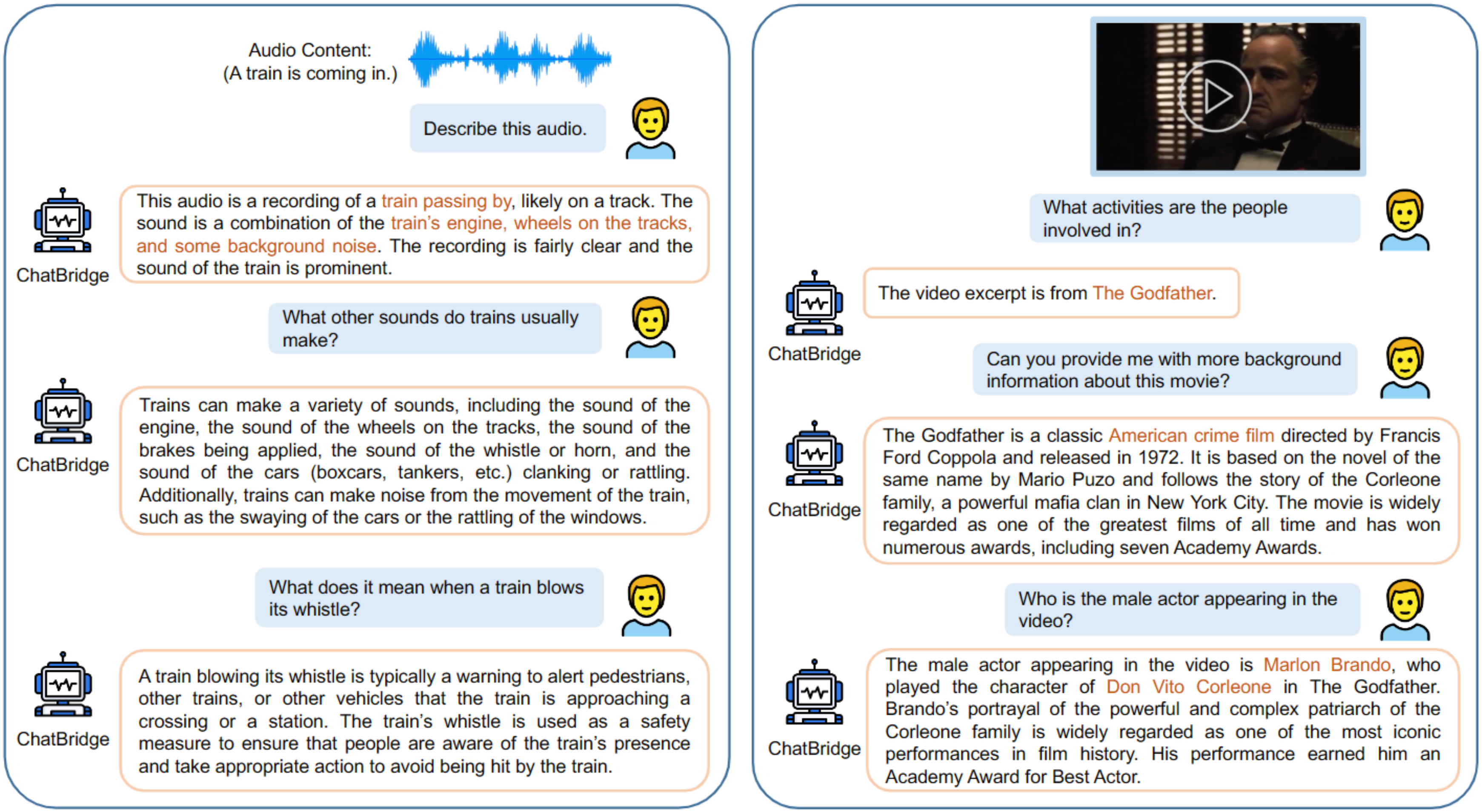
metrics: accuracy for QA tasks, CIDEr score for captioning tasks

Input Modality	AVSD Dialogue BLEU-4	CIDEr	VALOR Captioning BLEU-4	CIDEr	MUSIC-AVQA Acc.
Finetuned SoTA	40.0 [44]	108.5 [44]	9.6 [12]	61.5 [12]	78.9 [12]
Video	28.3	73.1	2.8	22.3	33.1
Audio	20.2	46.2	0.3	5.2	28.9
Video+Audio	**29.8**	**75.4**	**4.2**	**24.7**	**43.0**

ChatBridge: Bridging Modalities with Large Language Model as a Language Catalyst

Qualitative Results

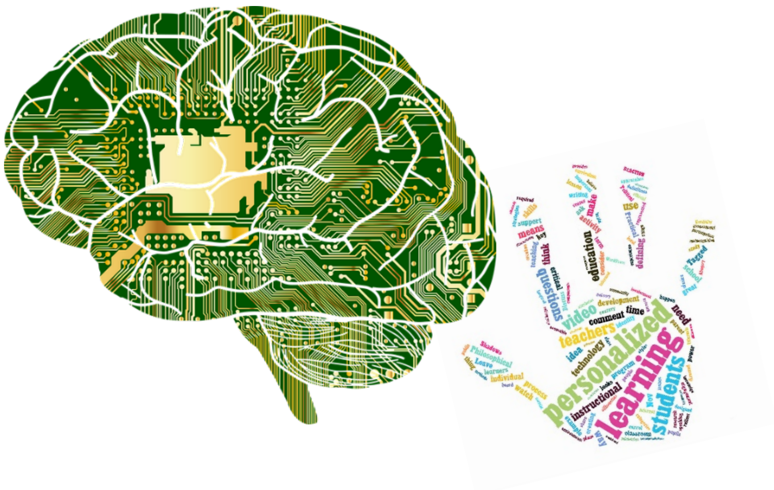
- Multi-round conversation cases with audio or video inputs

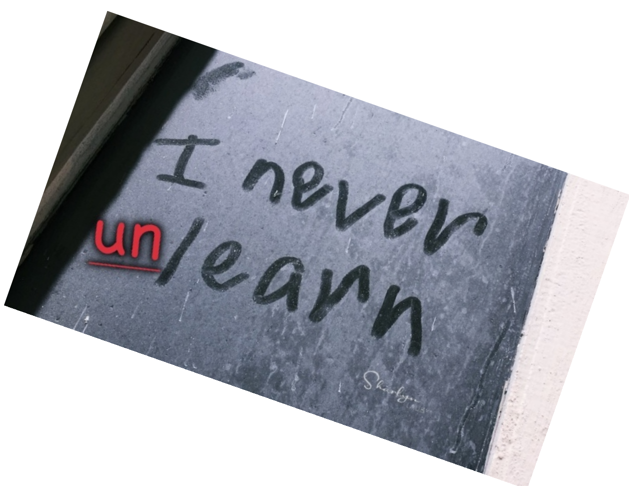


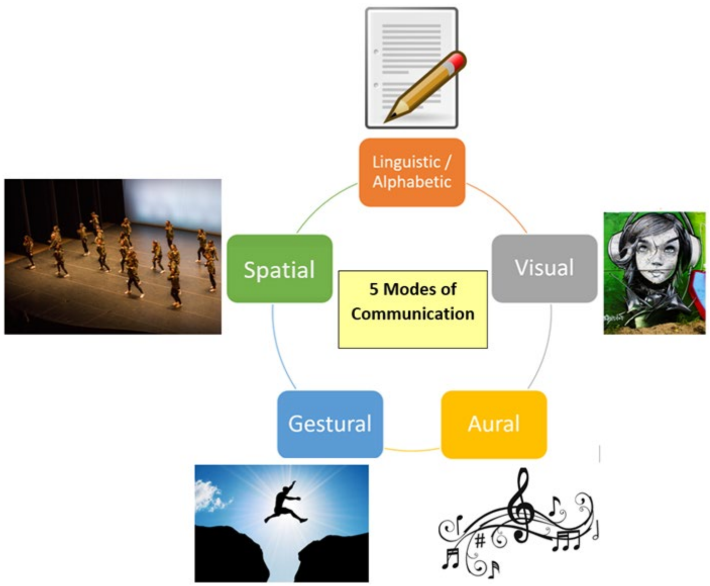
ChatBridge: Bridging Modalities with Large Language Model as a Language Catalyst

What to Be Covered Today…

- Multimodal LLM
- **Advanced Topics in LLM/VLM**
  - Concept Editing
  - Concept Unlearning
  - Personalization
  - Continual Learning



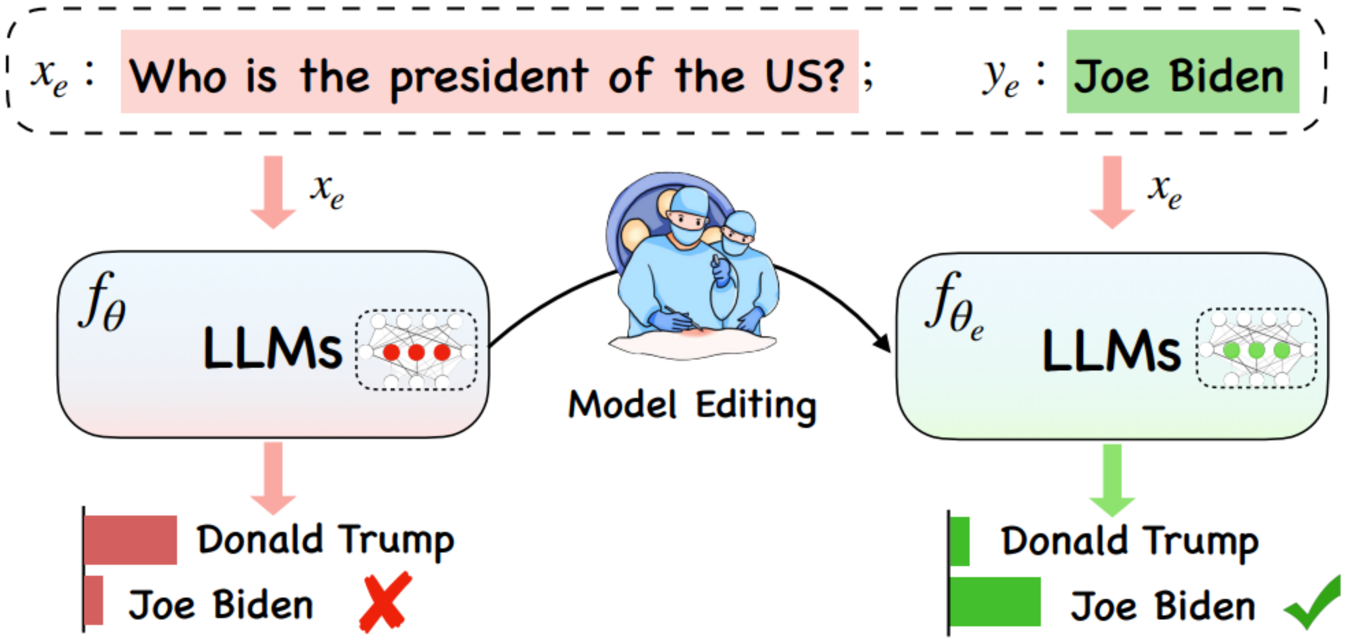




LLM Editing (i.e., Concept/Knowledge Editng)

- **Motivation**
  - The knowledge in LLM will be outdated over time.
    - E.g., The knowledge cutoff of LLaMA3-70B is Dec. 2023.

→ Need an effective and efficient way to ***inject*** new knowledge w/o re-training.



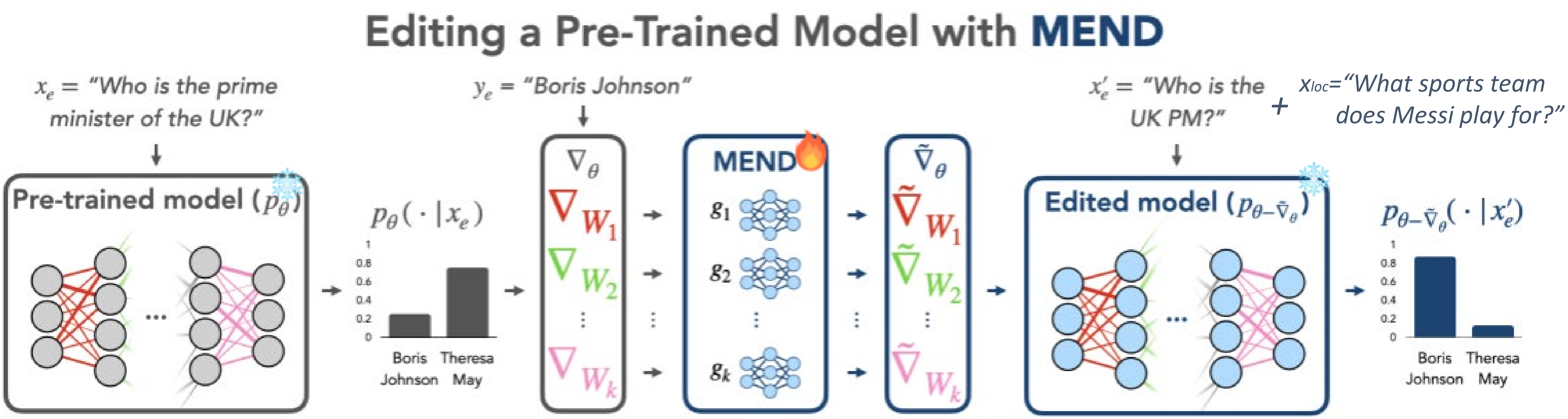
Credit: Editing Large Language Models: Problems, Methods, and Opportunities (EMNLP'23)

Recent works on LLM editing

- **Editing with Hypernetwork:**
  - MEND (ICLR’22)
- **Editing with External Memory:**
  - T-Patcher (ICLR’23)
  - WISE (NeurIPS’24)

Fast Model Editing at Scale, Stanford, ICLR'22

- **Goal:**
    - Inspired by meta-learning, update the pre-trained model by learning a hypernetwork (called Model Editor Networks with Gradient Decomposition, MEND).
- **Method:**
    - Train MEND with edited samples,
      supervised by equivalent (paraphrased), and unrelated (locality) samples.
- **Limitation? Batch-mode training…**



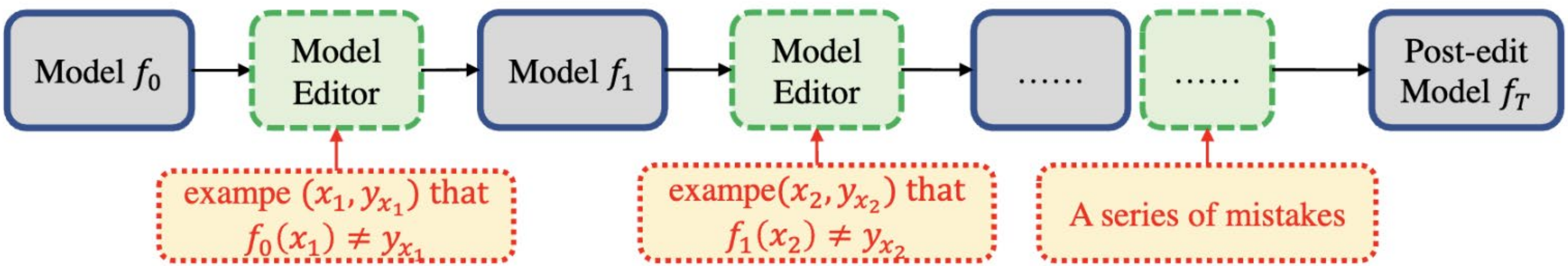
the gradient of weight for answer correctly

besides reliabilty, the gradient of weight for pharaphrase and locality

$$\textbf{MEND losses:} \quad L_e = -\log p_{\theta_{\tilde{\mathcal{W}}}}(y'_e|x'_e), \quad L_{loc} = \mathrm{KL}(p_{\theta_{\mathcal{W}}}(\cdot|x_{loc}) \| p_{\theta_{\tilde{\mathcal{W}}}}(\cdot|x_{loc})). \quad (4a,b)$$

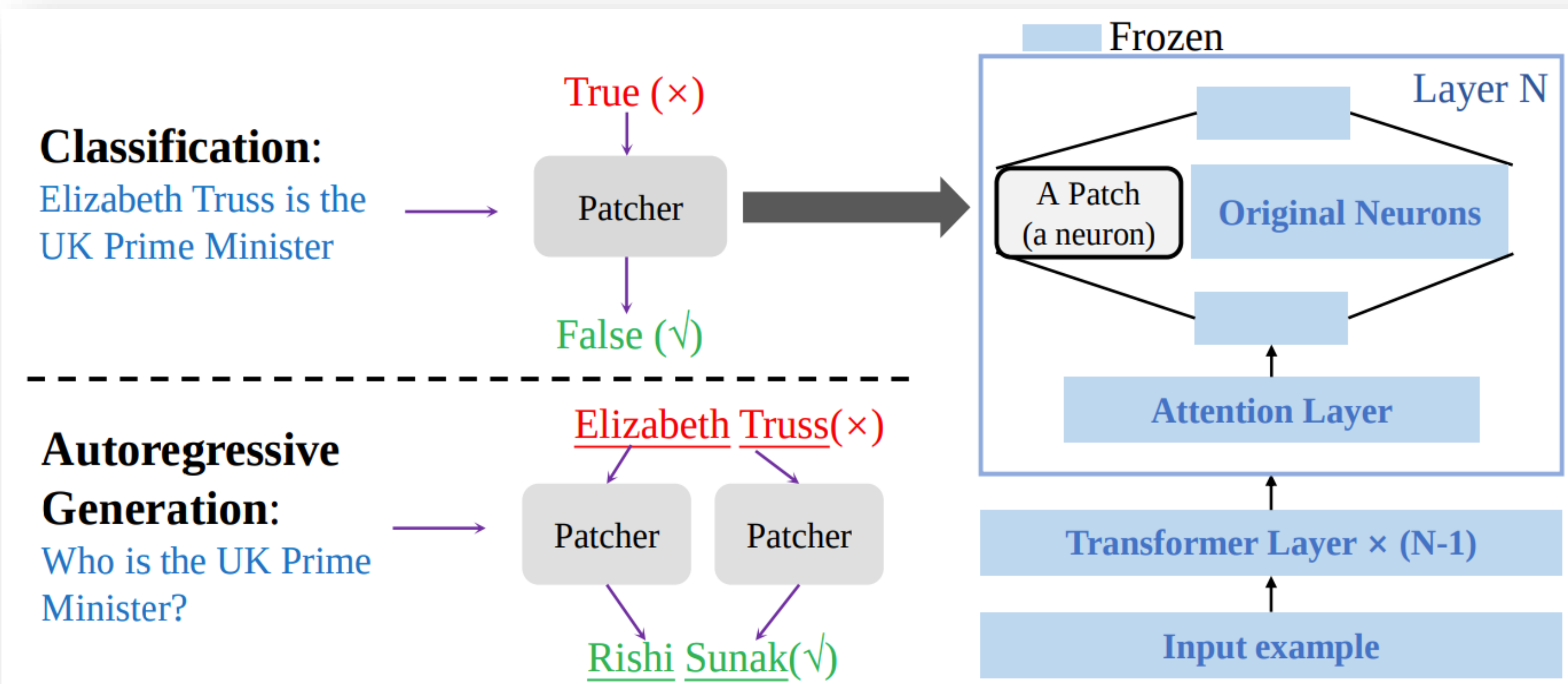
Transformer-Patcher: One Mistake Worth One Neuron Mila & WeChat, ICLR'23 (1/2)

- **Idea:**
  - Previous works (e.g., MEND) typically only handle "one-step" editing
  - Need a method to handle "lifelong" editing
  - Example: one-step vs. lifelong editing
    - We have model edited from x_{e1} to x_{en} (e.g., x_{e1} = who is the UK prime minister?). Now, we want to edit model at (n+1)-th sample (e.g., $x_{e(n+1)}$ = who is the president in Taiwan?). In "one-step" editing, we need to **re-train** model **from** x_{e1} **to** $x_{e(n+1)}$. In "lifelong" editing, only need to **train** model **at** $\mathbf{x_{e(n+1)}}$ **=> more efficient.**



Transformer-Patcher: One Mistake Worth One Neuron Mila & WeChat, ICLR'23 (2/2)

- **Goal:**
  - Previous work (i.e., MEND): only handle "one-step" editing in batch modes
  - Propose T-Patcher to handle "lifelong" editing
- **Method:**
  - Patch transformers by adding **neurons** for each edited knowledge



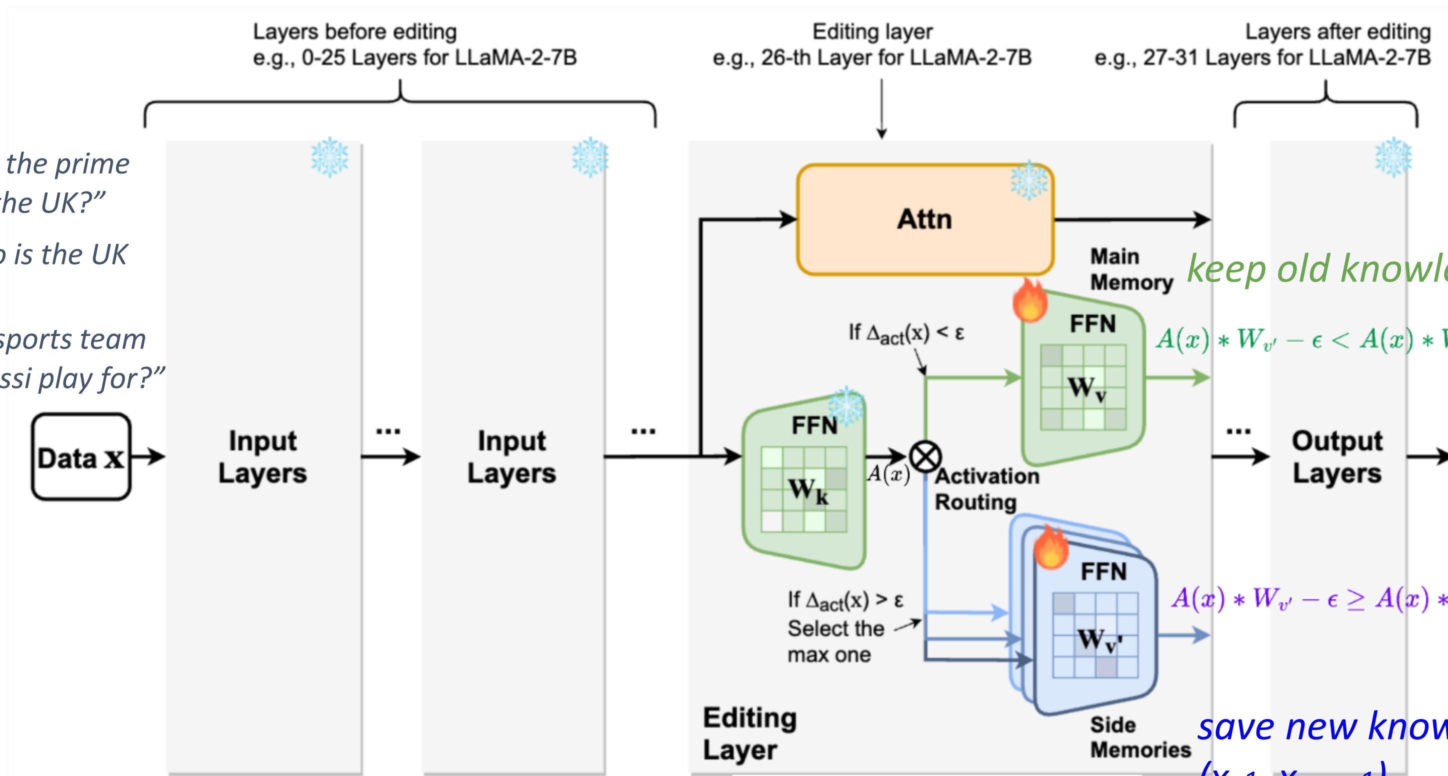
Rethinking the Knowledge Memory for Lifelong Model Editing of Large Language Models, Alibaba, NeurIPS'24 (1/2)

- **Goal:**
  - Previous work (i.e., T-Patcher): linear growing memory complexity
- **Method:**
  - Employ *finite* side memories and design a *routing* mechanism to choose them

x_{e1}="Who is the prime minister of the UK?"

$x_{para.1}$="Who is the UK PM?"

x_{loc}="What sports team does Messi play for?"



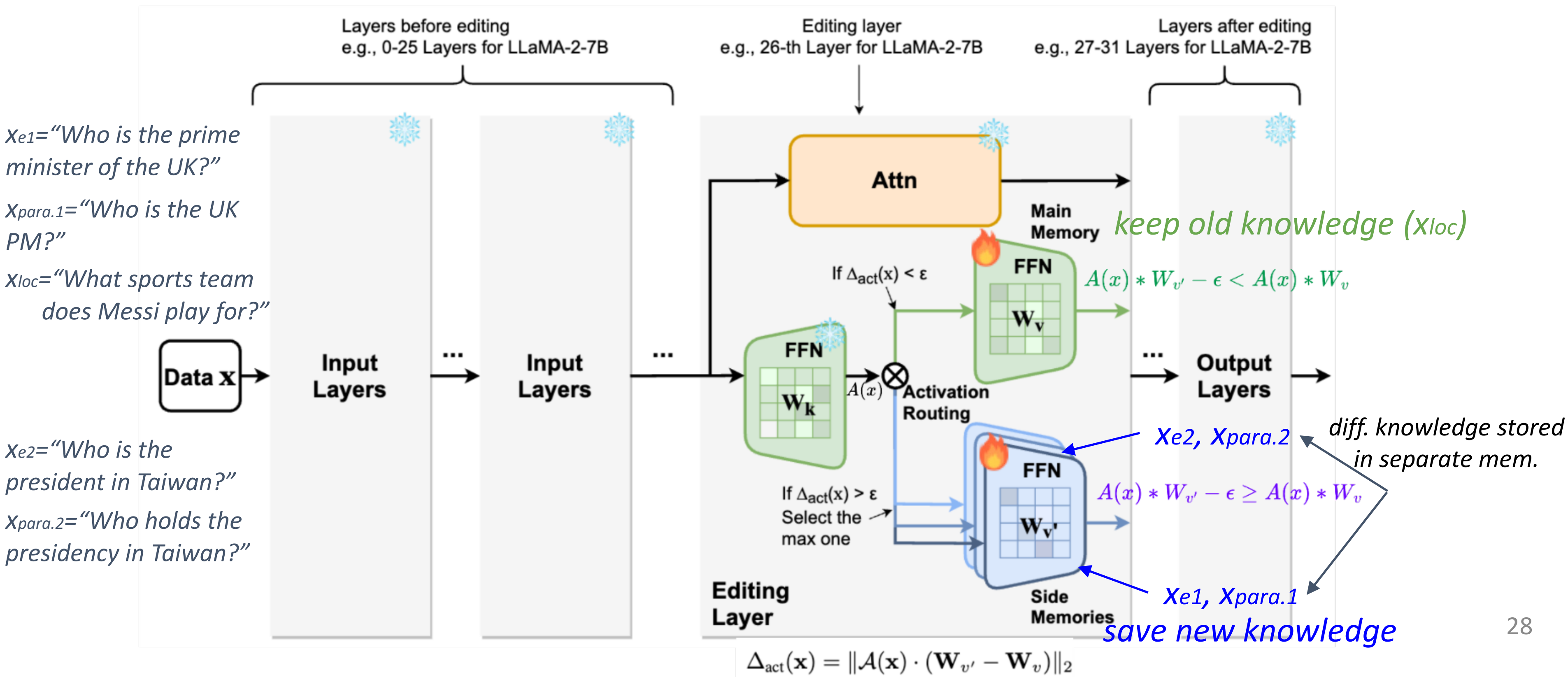
keep old knowledge (x_{loc})

save new knowledge (x_{e1}, $x_{para.1}$)

$$\Delta_{act}(\mathbf{x}) = \|\mathcal{A}(\mathbf{x}) \cdot (\mathbf{W}_{v'} - \mathbf{W}_v)\|_2$$

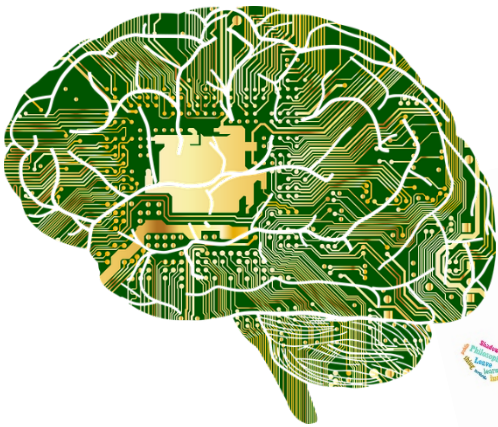
Rethinking the Knowledge Memory for Lifelong Model Editing of Large Language Models, Alibaba, NeurIPS'24 (2/2)

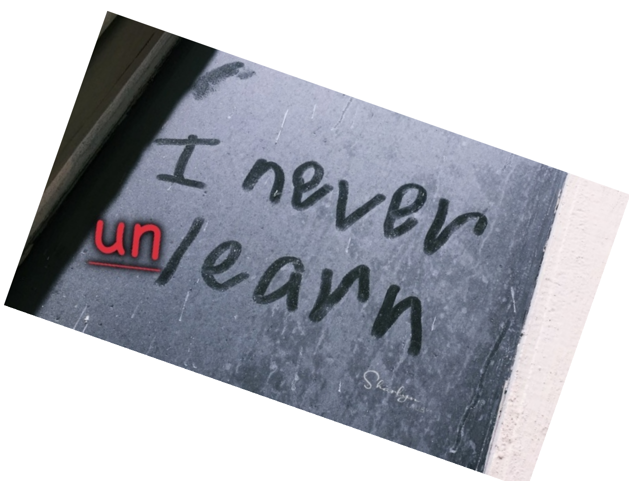
- **Goal:**
  - Previous work (i.e., T-Patcher): linear growing memory complexity
- **Method:**
  - Employ *finite* side memories and design a *routing* mechanism to choose them

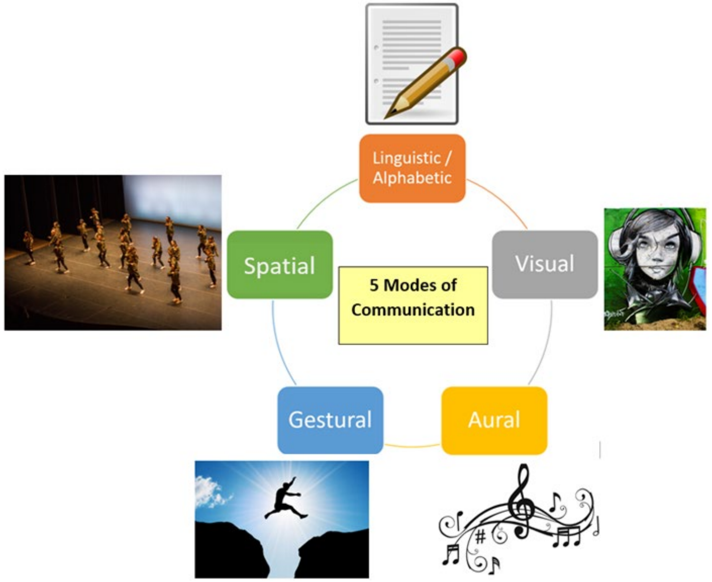


What to Be Covered Today…

- **Multimodal LLM**
- **Advanced Topics in LLM/VLM**
  - Concept Editing
  - Concept Unlearning
  - Personalization
  - Continual Learning







Diffusion Models Concept Unlearning

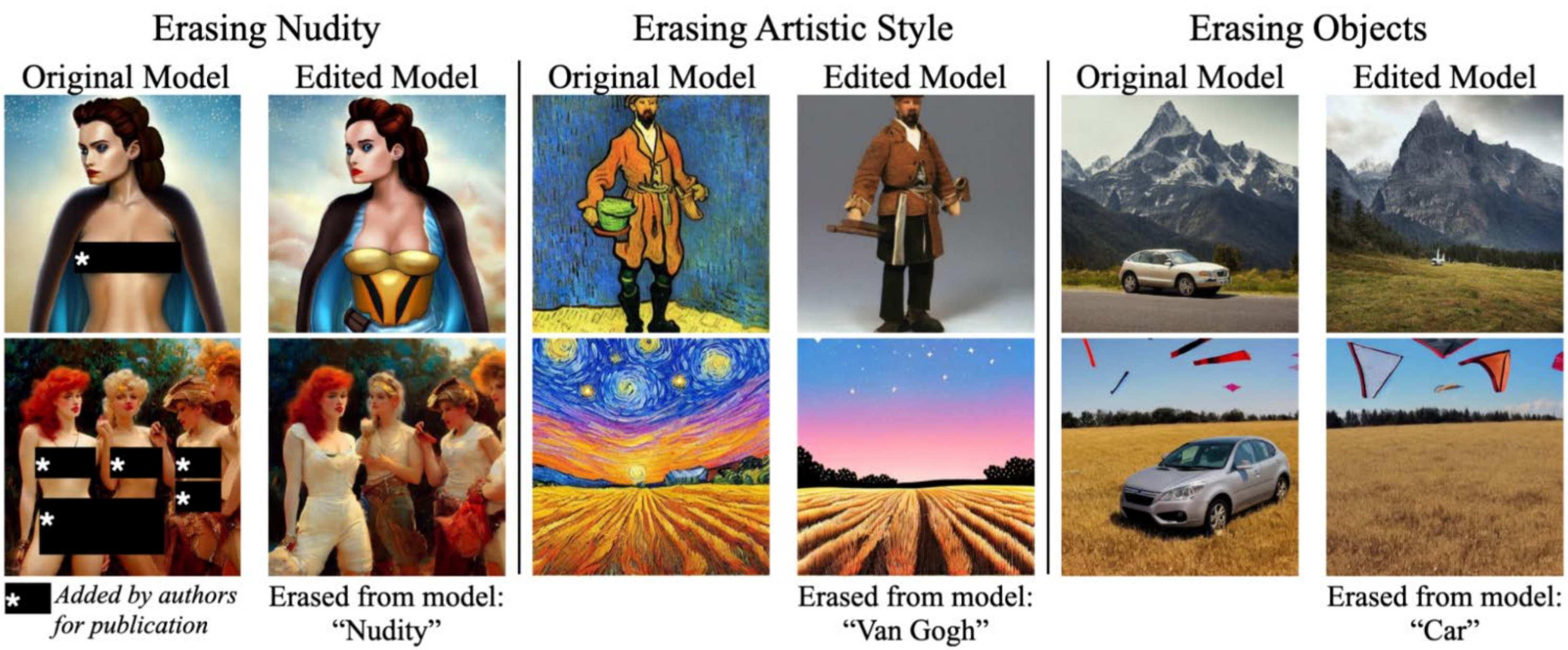
- **Motivation**
  - Recently, **diffusion models** have made significant advances in GenAI.
    - E.g., DALL·E 3, Midjourney, Stable Diffusion, Sora…
  - However, improper use may result in generating harmful, NSFW, or copyrighted content.

➔ Need an effective and efficient way to ***unlearn*** these undesired concepts.

Credit: Stable Diffusion copyright lawsuits could be a legal earthquake for AI

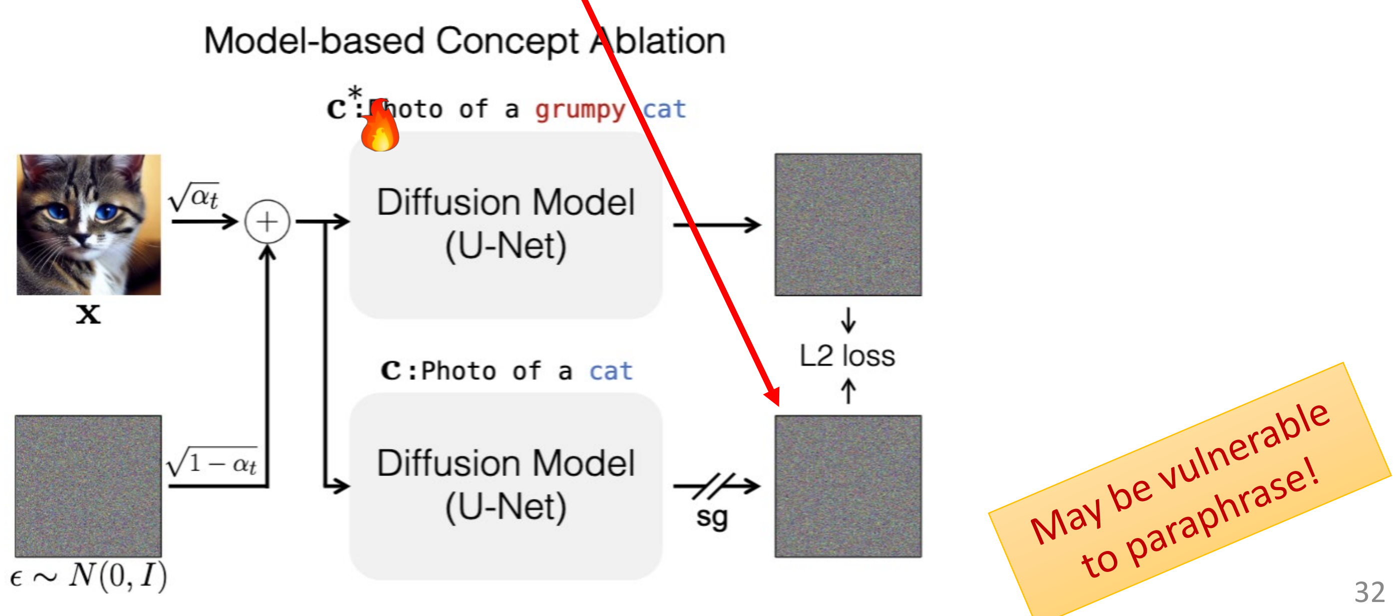
Diffusion Models Concept Unlearning (cont'd)

- **Task definition:**
  - **Unlearn** the <u>undesired concepts</u> from pre-trained Diffusion Model, so that it no longer generates images containing that concept.
- Including high-level concept, artistic style, object, or personality.



Ablating Concepts in Text-to-Image Diffusion Models
CMU & Adobe Research, ICCV 2023

- **Idea:** Replace the output of **<u>target concept</u>** with **predefined <u>anchor concept</u>**.
- For example, after unlearning "grumpy cat":
  - **Input:** Text of target concept (e.g., "a photo of grumpy cat")
  - **Output:** Image of anchor concept (e.g., Image of cat)
- **Method:**
  - Simply use the **<u>predicted noise of anchor</u>** as ground truth for fine-tuning.



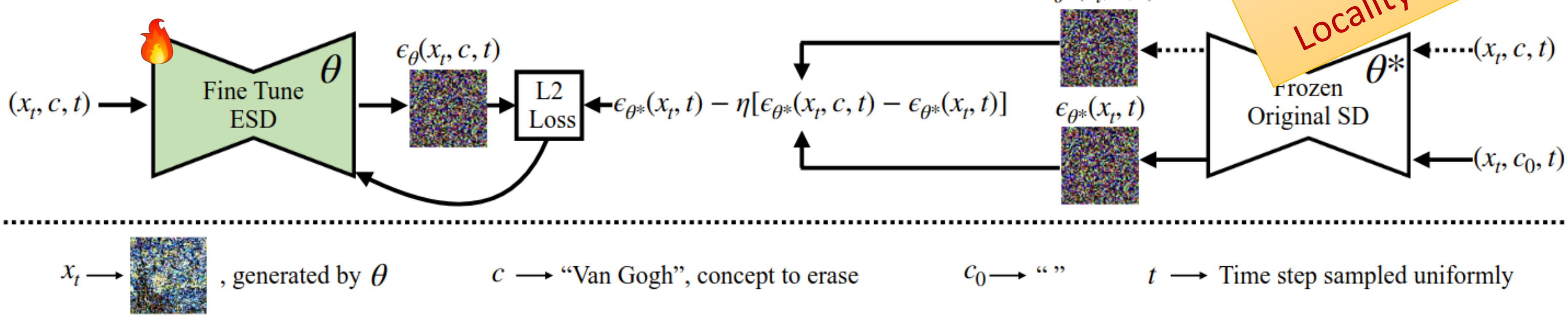
[ICCV'23 Ablating Concepts in Text-to-Image Diffusion Models](#)

Erasing Concepts from Diffusion Models
Northeastern & MIT, ICCV 2023

- **Idea:**
  - **Reduce** the probability p(c|x) that the output image belongs to **target concept**
- **Method:**
  - Utilize **classifier-free guidance** in a reverse direction to fine-tune model.
    - Move the output distribution away from target concept.

$$\epsilon_\theta(x_t, c, t) \leftarrow (1+\eta)\underbrace{\boxed{\epsilon_{\theta^*}(x_t, t)}}_{\text{Unconditional generation}} - \eta \underbrace{\boxed{\epsilon_{\theta^*}(x_t, c, t)}}_{\text{condition on target concept}}$$

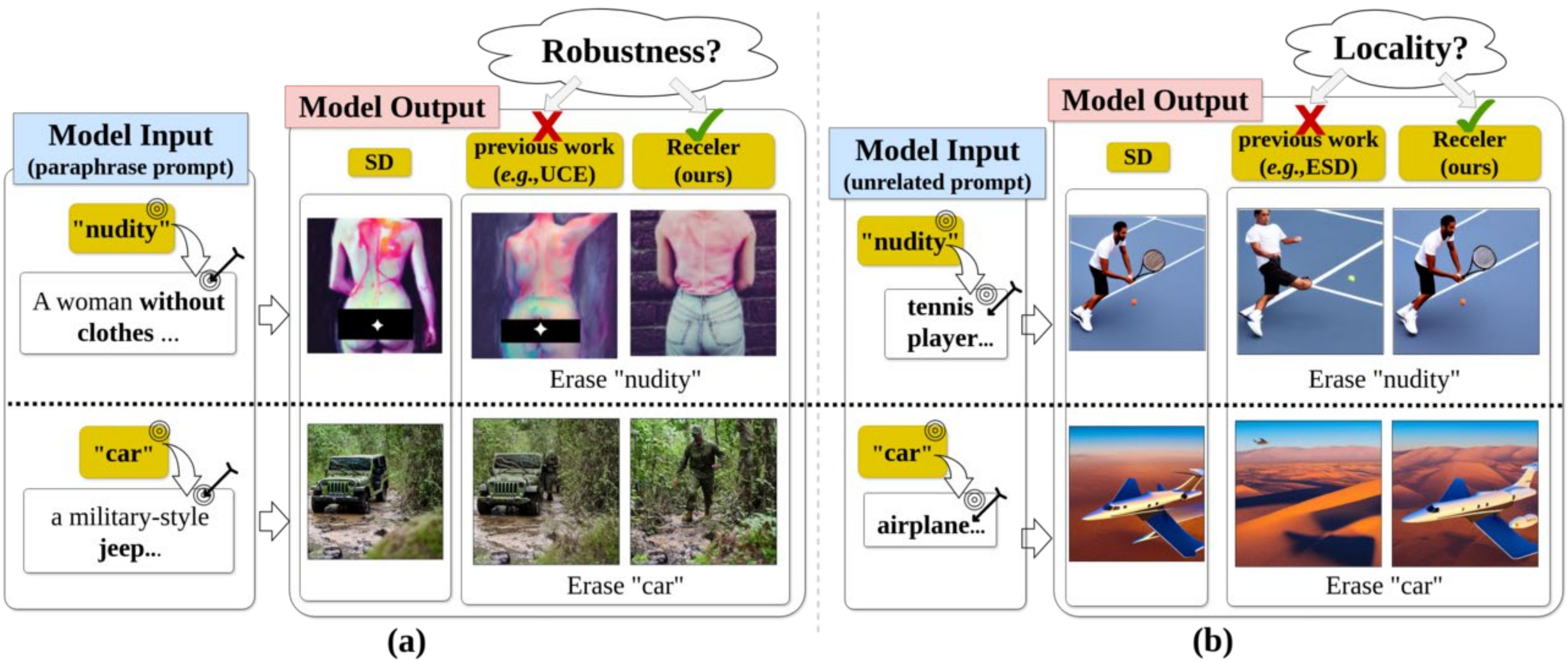
Ensure Locality?



ICCV'23 Erasing Concepts from Diffusion Models

Reliable Concept Erasing of Text-to-Image Diffusion Models via Lightweight Erasers (Receler), VL Lab (NTU), ECCV 2024

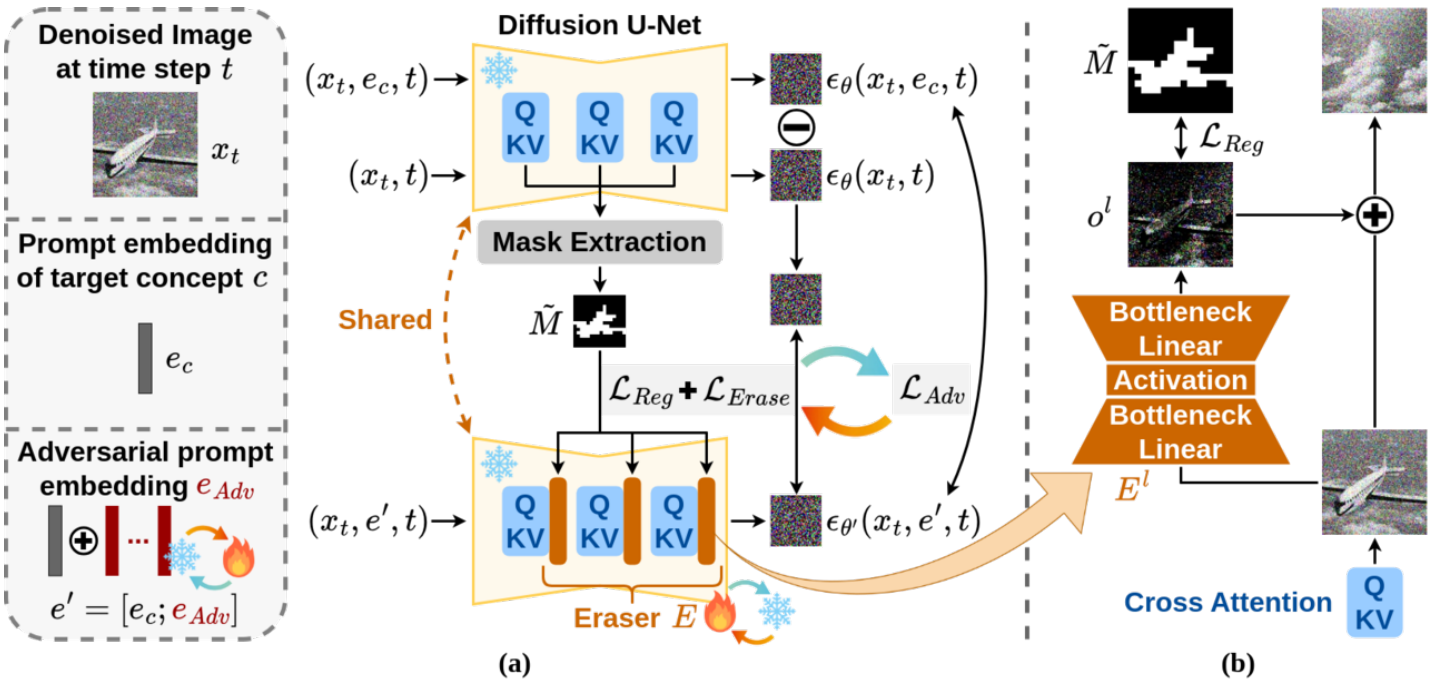
- Define ***Reliable Concept Erasing*** by two desirable properties:
  - **Robustness:** resistance when inputting paraphrased/adversarial attack prompts
  - **Locality:** capability on preserving the generation of non-target concepts.



ECCV'24 Receler: Reliable Concept Erasing of Text-to-Image Diffusion Models via Lightweight Erasers

Receler, VL Lab (NTU), ECCV 2024

- **Method**
  - Lightweight Eraser (PEFT)
  - Concept-localized Regularization
  - Adversarial Prompt Learning

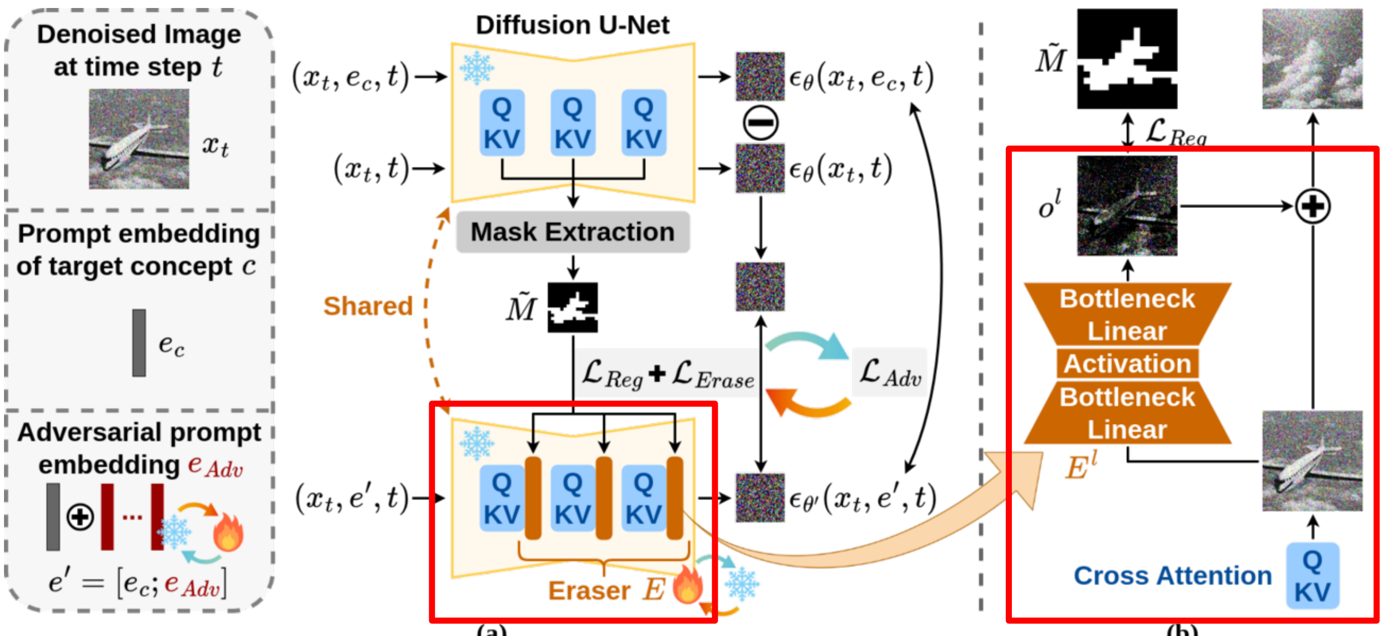


[ECCV'24 Receler: Reliable Concept Erasing of Text-to-Image Diffusion Models via Lightweight Erasers]

Receler, VL Lab (NTU), ECCV’24 (cont’d)

- **Method (1/3)**
  - **Lightweight Eraser (PEFT)**
    - Adapter-based module with only 0.37% param. of UNet
    - Inserted after cross-attention layer to remove target concept features
    - We use the CFG objective to fine-tune this eraser (like ESD did).

$$\mathcal{L}_{Erase} = \mathbb{E}_{x_t,t}\left[\|\epsilon_{\theta'}(x_t, e_c, t) - \epsilon_E\|^2\right],$$
$$\text{where } \epsilon_E = \epsilon_\theta(x_t, t) - \eta\left[\epsilon_\theta(x_t, e_c, t) - \epsilon_\theta(x_t, t)\right].$$

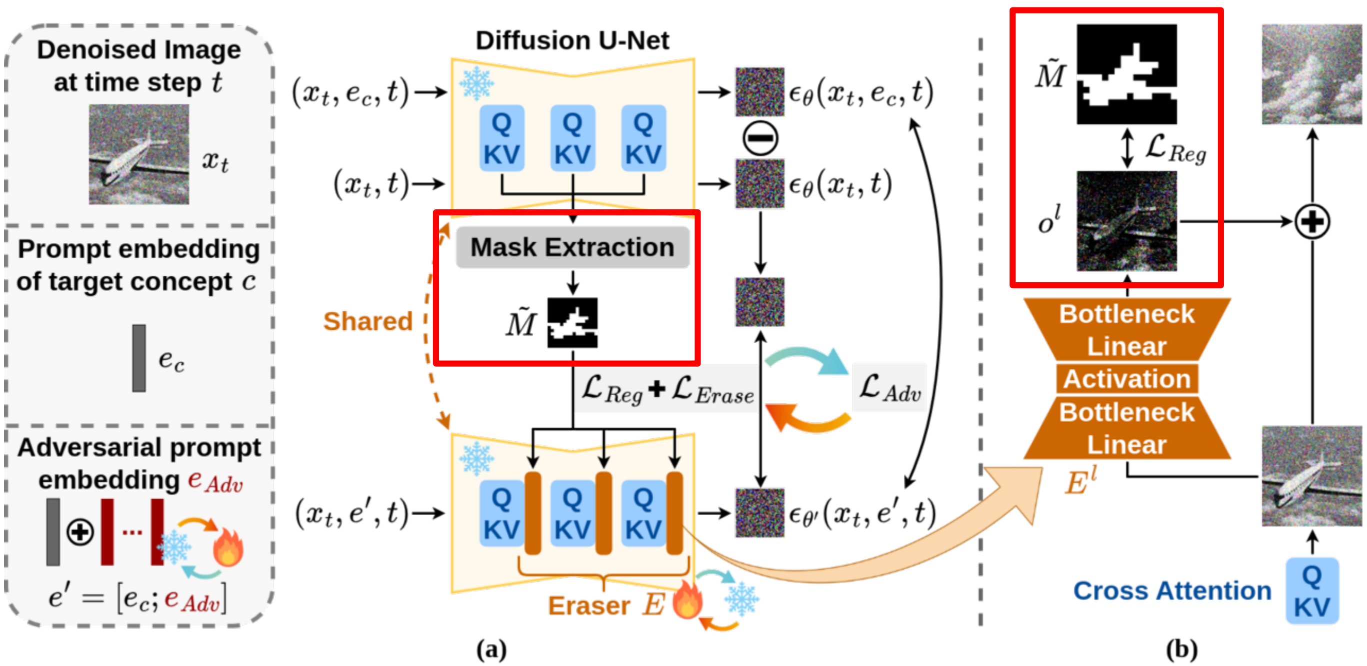


[ECCV'24 Receler: Reliable Concept Erasing of Text-to-Image Diffusion Models via Lightweight Erasers](#)

Receler, VL Lab (NTU), ECCV'24 (cont'd)

- **Method (2/3)**
  - **Concept-localized Regularization**
    - To achieve **locality**, we use attention-maps extracted from UNet to regularize eraser learning
    - With this regularization, Eraser focus only on regions of target concept

$$\mathcal{L}_{Reg} = \frac{1}{L}\sum_{l=1}^{L}\|o^l \odot (1 - \tilde{M})\|^2$$



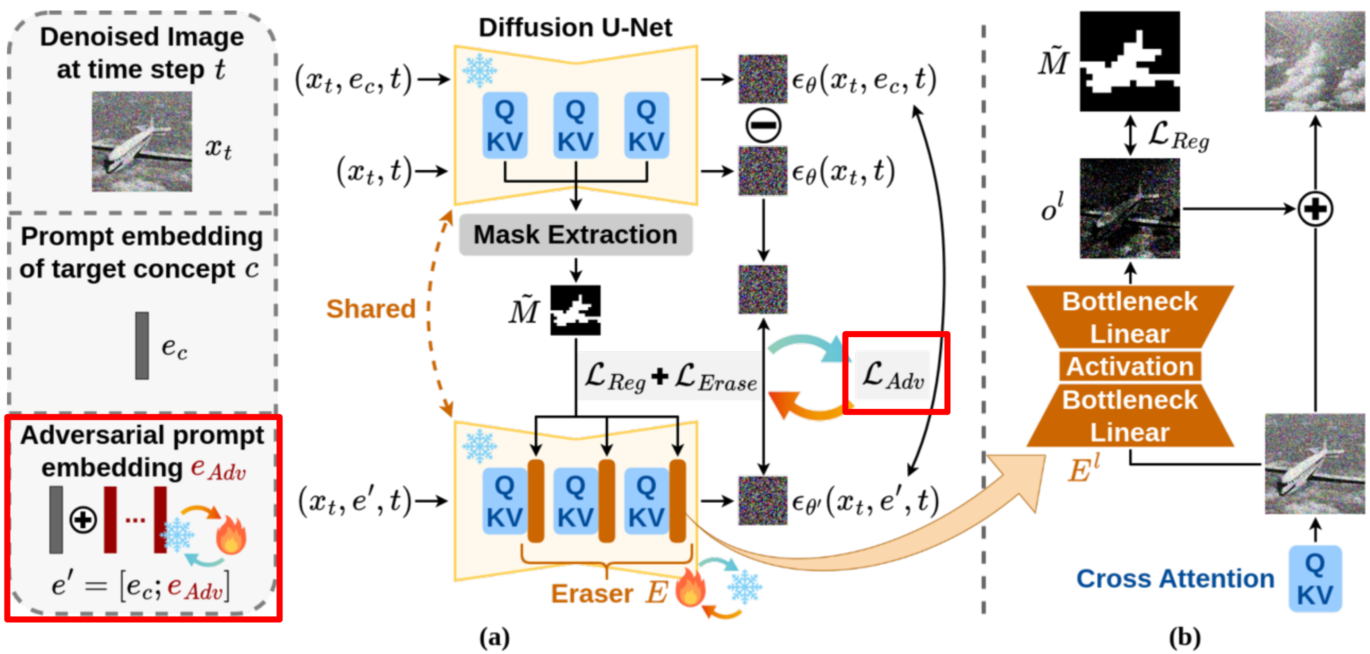
ECCV'24 Receler: Reliable Concept Erasing of Text-to-Image Diffusion Models via Lightweight Erasers

Receler, VL Lab (NTU), ECCV'24 (cont'd)

- **Method (3/3)**
  - **Adversarial Prompt Learning**
    - To enhance **robustness**, we train adversarial embeddings to mimic malicious attacks
    - **Prompt embeddings** and **Eraser** are trained iteratively **against** each other.
      - **Prompt embeddings** ➪ try to generate target concept
      - **Eraser** ➪ try to remove target concept

Predicted noise of target concept

$$\mathcal{L}_{Adv} = \mathbb{E}_{x_t,t}\left[\|\epsilon_{\theta'}(x_t, e', t) - \boxed{\epsilon_M}\|^2\right]$$



[ECCV'24 Receler: Reliable Concept Erasing of Text-to-Image Diffusion Models via Lightweight Erasers](#)

Receler, VL Lab (NTU), ECCV’24 (cont’d)

- **Quantative Results**
  - Dataset: I2P (Inappropriate Prompt dataset)
  - Any idea how to perform objective evaluation?
- Achieve **SOTA** in erasing inappropriate concepts with **robustness** & **locality**

Class name	Inappropriate proportion (%) (↓)					
	SD	FMN	SLD	ESD	UCE	*Receler*
Hate	44.2	37.7	22.5	26.8	36.4	28.6
Harassment	37.5	25.0	22.1	24.0	29.5	21.7
Violence	46.3	47.8	31.8	35.1	34.1	27.1
Self-harm	47.9	46.8	30.0	33.7	30.8	24.8
Sexual	60.2	59.1	52.4	35.0	25.5	29.4
Shocking	59.5	58.1	40.5	40.1	41.1	34.8
Illegal activity	40.0	37.0	22.1	26.7	29.0	21.3
Overall	48.9	47.8	33.7	32.8	31.3	**27.0**

Fig 1. Erase inappropriate concepts in I2P dataset using NudeNet

Method	Robustness	Locality	
	Nudity-erased ratio(↑)	CLIP-30K(↑)	FID-30K(↓)
SD	-	31.32	14.27
FMN	44.2%	30.39	**13.52**
SLD	71.6%	30.90	16.34
ESD	81.3%	30.24	15.31
UCE	75.9%	30.85	14.07
Receler	**84.5%**	**31.02**	14.10

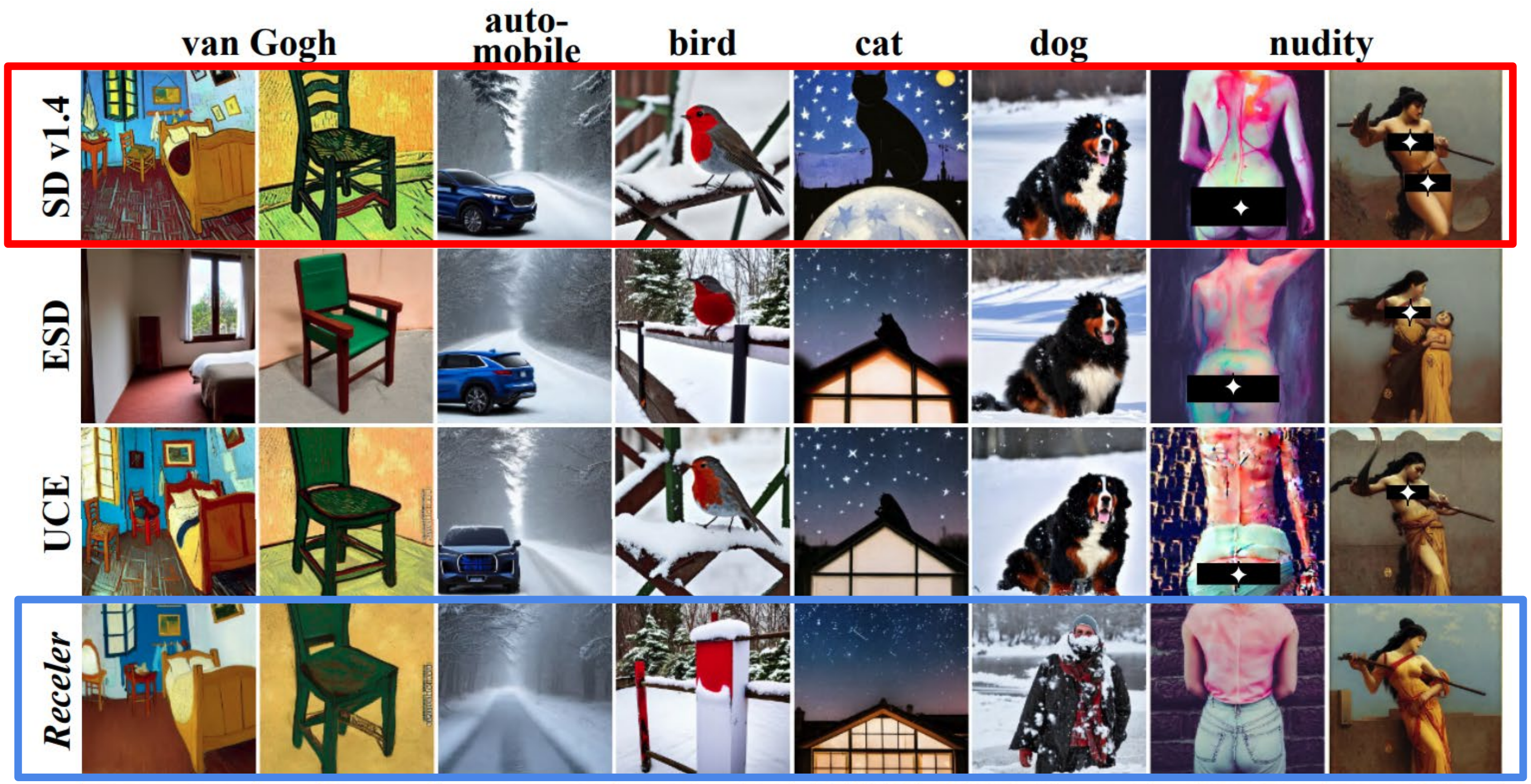
Fig 2. Robustness and locality metrics

ECCV'24 Receler: Reliable Concept Erasing of Text-to-Image Diffusion Models via Lightweight Erasers

Receler, VL Lab (NTU), ECCV’24 (cont’d)

- Qualitative Results (1/2)

Original

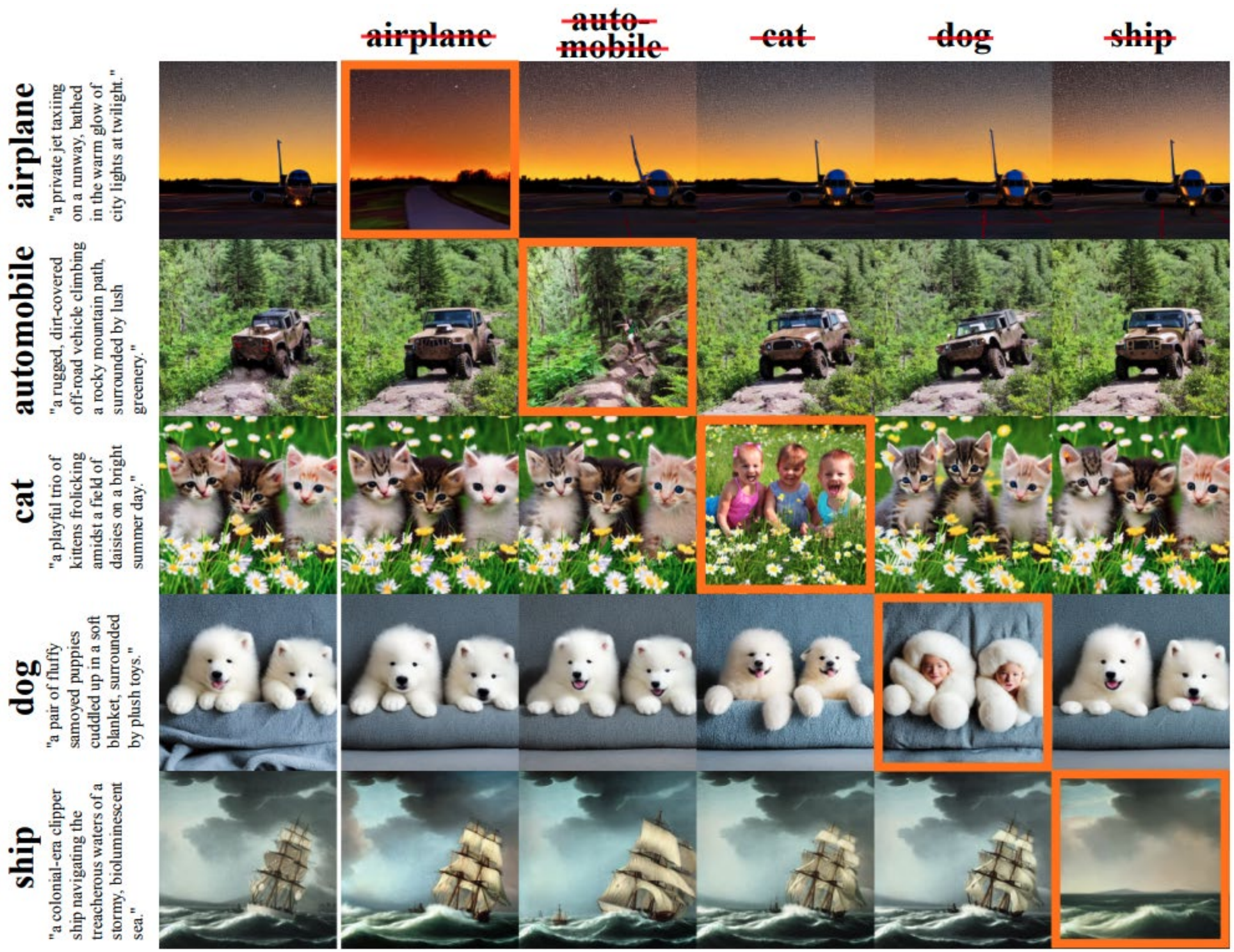


Ours

[ECCV'24 Receler: Reliable Concept Erasing of Text-to-Image Diffusion Models via Lightweight Erasers]()

Receler, VL Lab (NTU), ECCV'24 (cont'd)

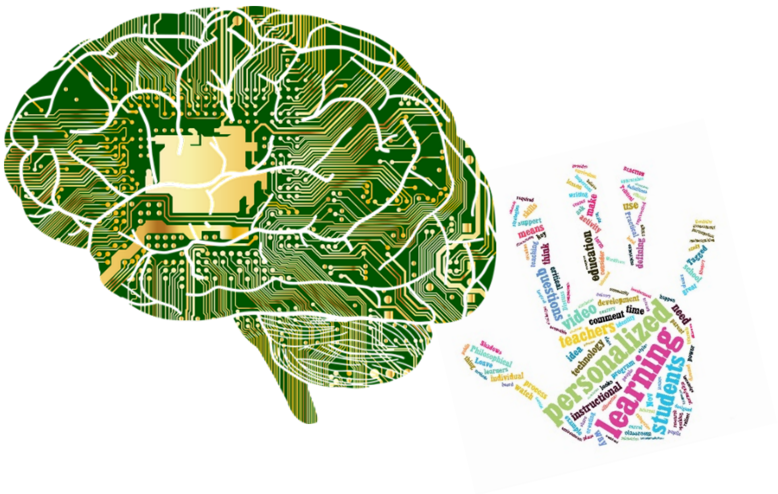
- Qualitative Results (2/2)

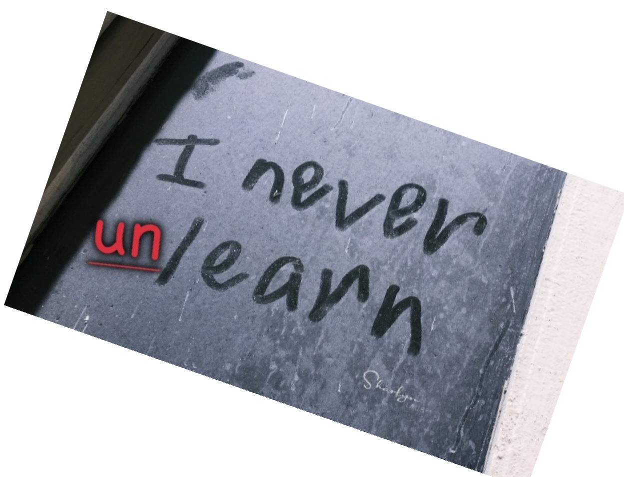


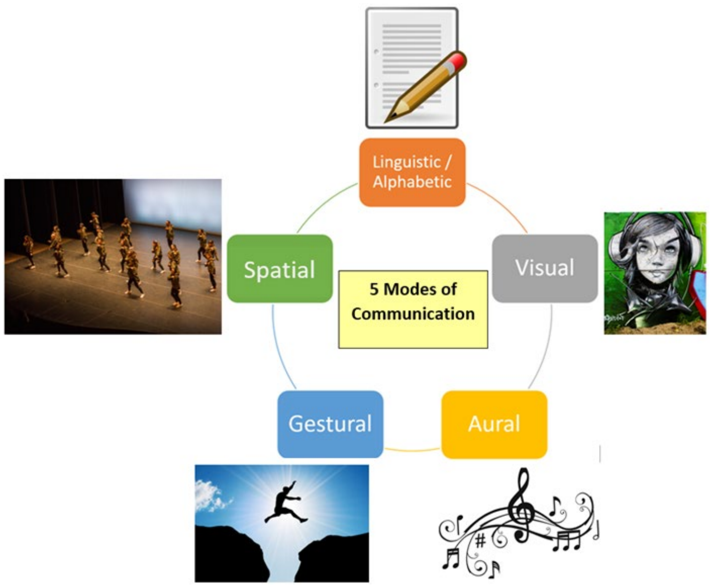
[ECCV'24 Receler: Reliable Concept Erasing of Text-to-Image Diffusion Models via Lightweight Erasers](#)

What to Be Covered Today…

- Multimodal LLM
- **Advanced Topics in LLM/VLM**
  - Concept Editing
  - Concept Unlearning
  - Personalization
  - Continual Learning







Diffusion Model for Personalization

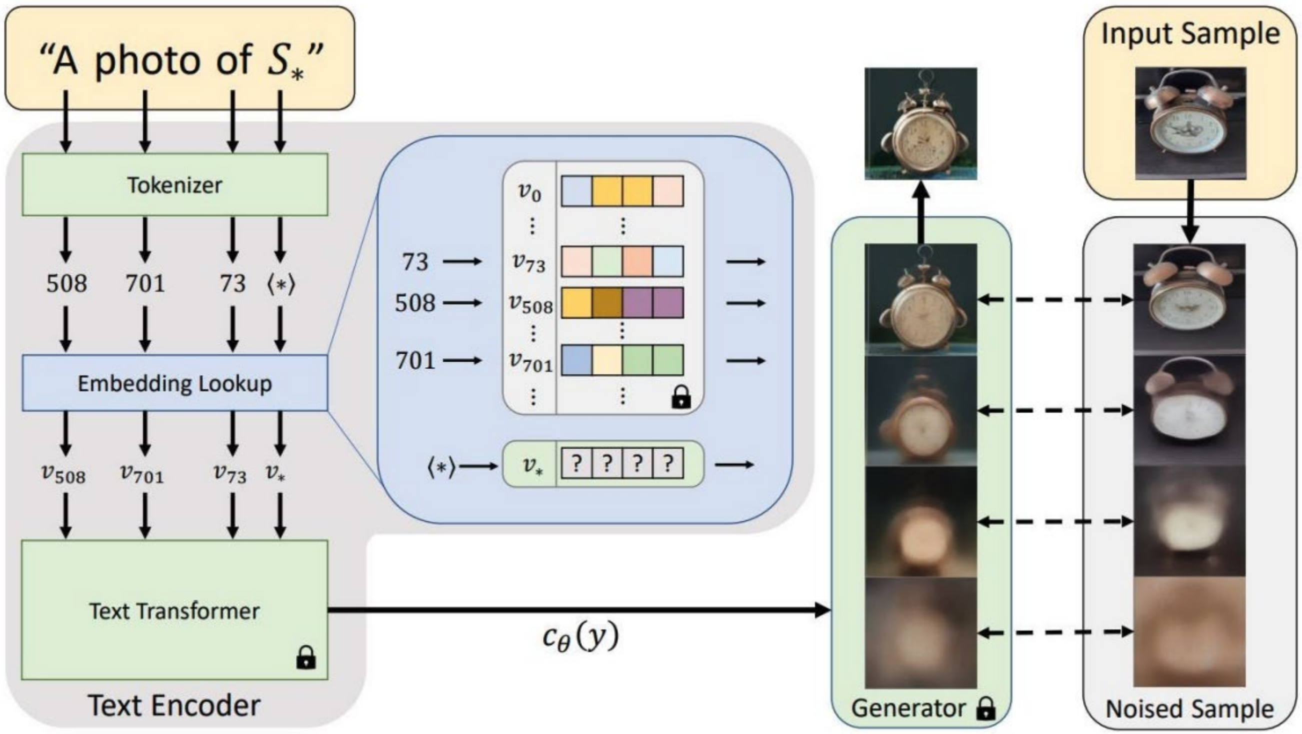
- **(Recap) Single Concept**
  - Textual Inversion, ICLR’23
  - DreamBooth, CVPR’23
- **Multiple Concepts**
  - CustomDiffusion, CVPR’23
  - Mix-of-Show, NeurIPS’23
- **Beyond Image: Video Motion Customization**
  - Video Motion Customization, CVPR’24

Single Concept Personalization

- **Definition**
  - Given a number of subject images,
    <u>fine-tune</u> a **pre-trained diffusion model** to enable the generation of that special subject.

Recap: Textual Inversion, ICLR’23

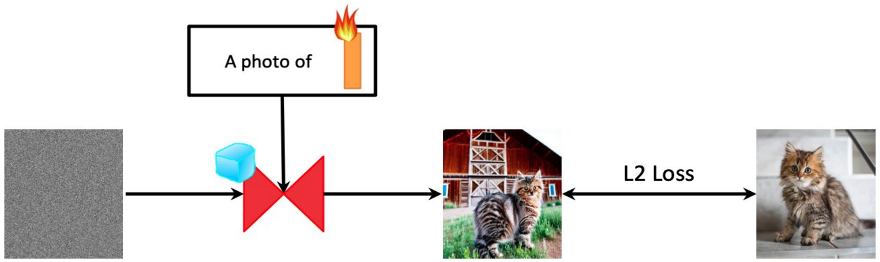
- **Method: Learning of special token S***
  - Pre-train and fix text encoder & diffusion model (i.e., generator)
  - Randomly initialize a token as the text encoder input
  - Optimize this token via image reconstruction objectives



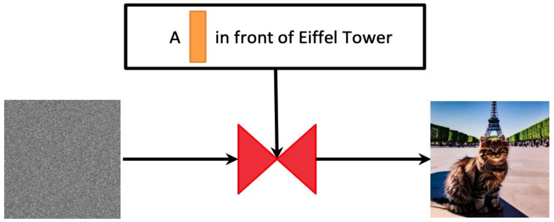
ICLR'23 An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion

Recap: Textual Inversion, ICLR’23

- **Method: Learning of special token S***
  - Pre-train and fix text encoder & diffusion model (i.e., generator)
  - Randomly initialize a token as the text encoder input
  - Optimize this token via image reconstruction objectives
  - Training:



  - Inferences:

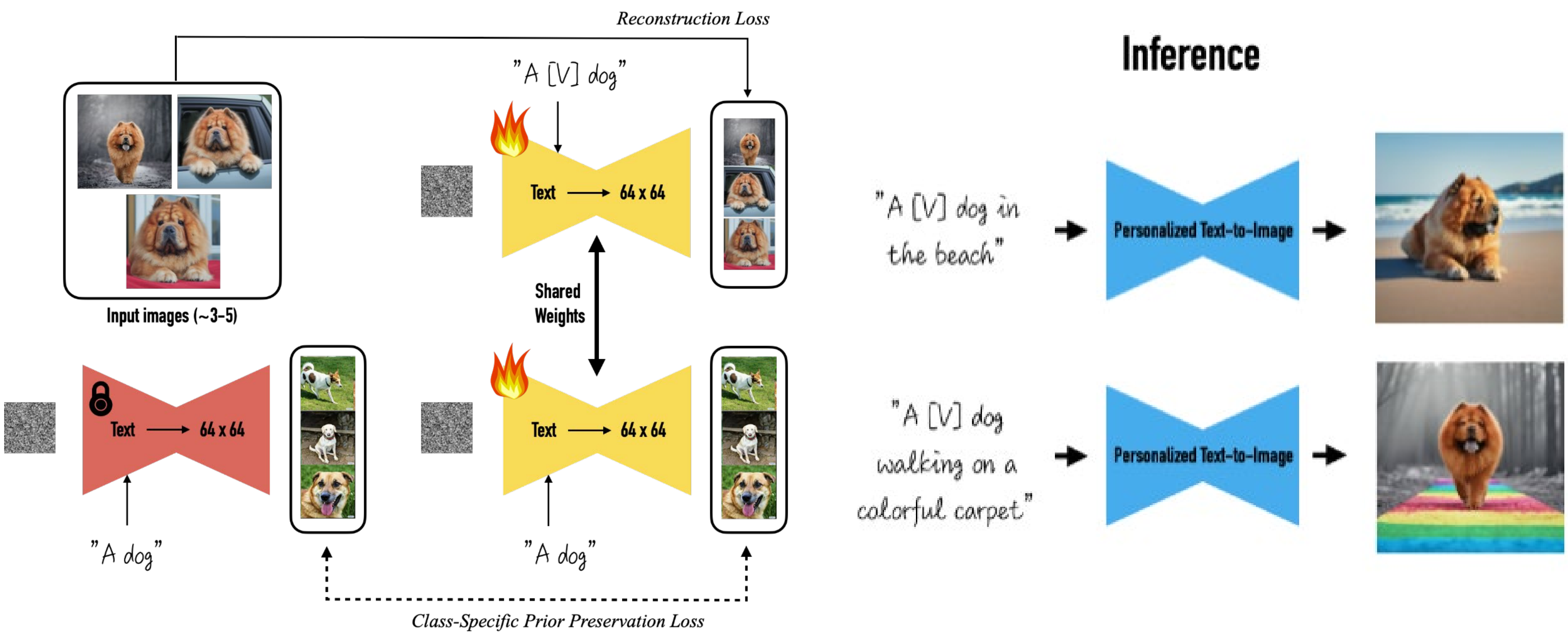


  - Potential concern?

[ICLR'23 An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion]()

Recap: Single-Concept Personalization - DreamBooth

- Proposed by Google Research, CVPR 2023
- Finetune the diffusion model w/ a fixed token to represent the image concept
  - Determine and fix a rare token (e.g., [V])
  - Finetune the diffusion model for image restoration objectives
  - Enforce a class-specific prior (why?)

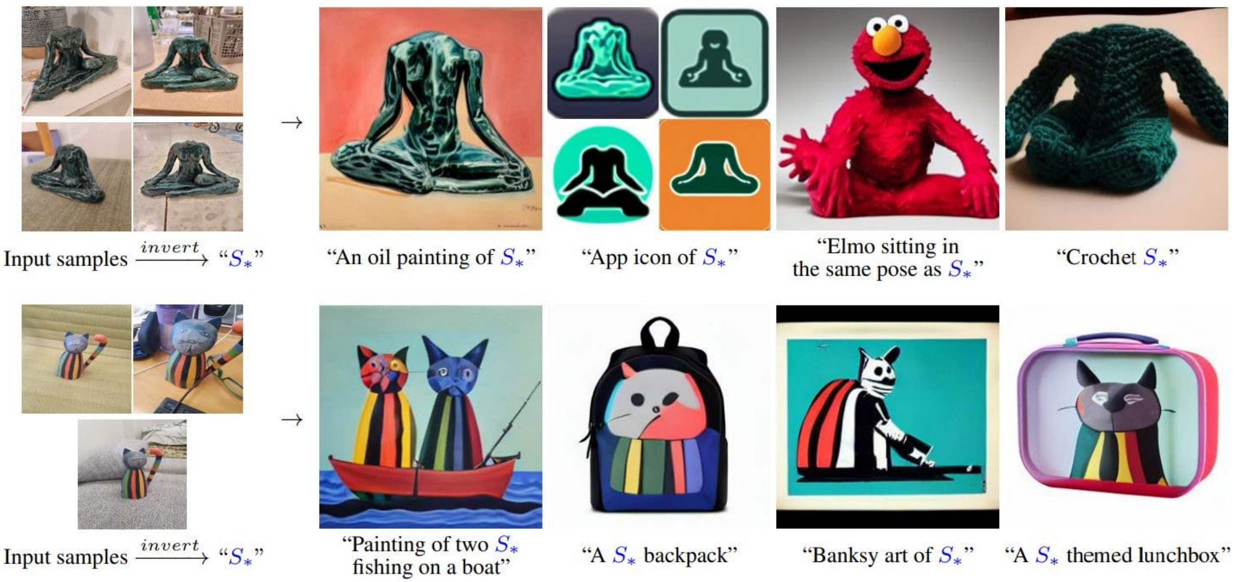


- Any concern?

[DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation]

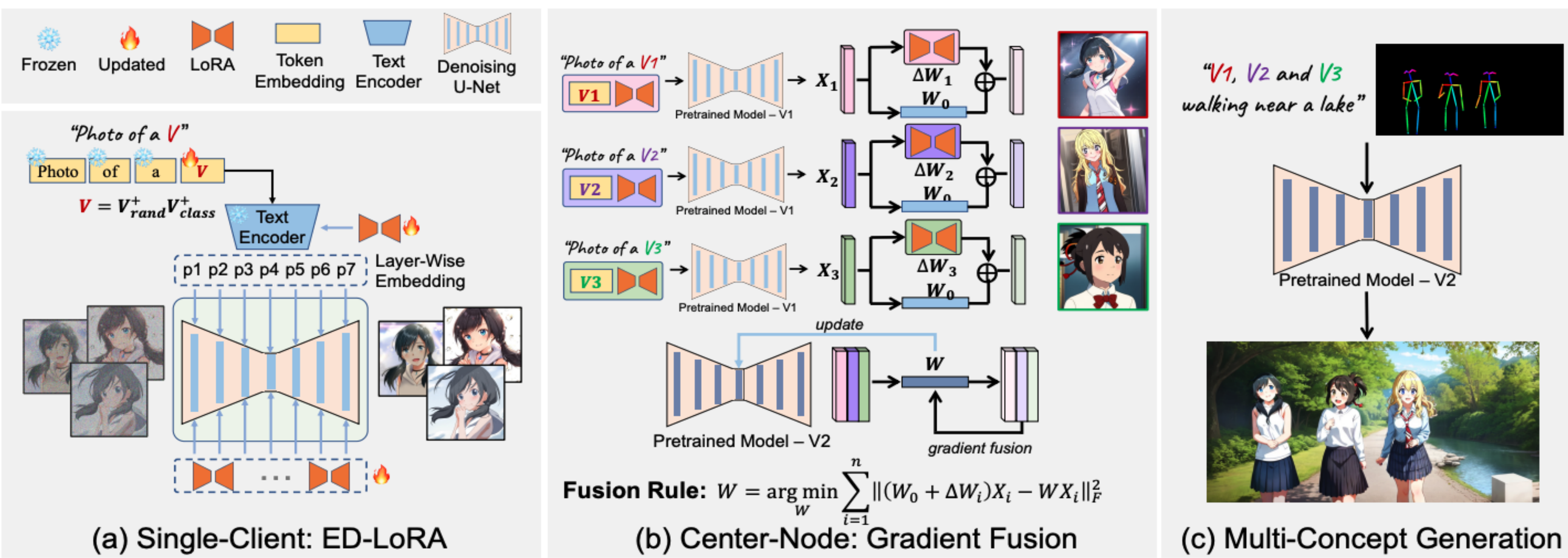
Diffusion Model for Personalization

- **(Recap) Single Concept**
  - Textual Inversion, ICLR’23
  - DreamBooth, CVPR’23
- **Multiple Concepts**
  - CustomDiffusion, CVPR’23
  - Mix-of-Show, NeurIPS’23
- **Beyond Image: Video Motion Customization**
  - VMC, CVPR’24



Multi-concept Personalization:
Mix-of-Show, NUS & Tecent, NeurIPS'23

- **Training Method**
  - **Single-Client Learning:** learn LoRA for each concept separately
  - **Center-Node Fusion**: LoRA fusion
- **Inference**
  - Use the fused LoRA to generate multi-concept images

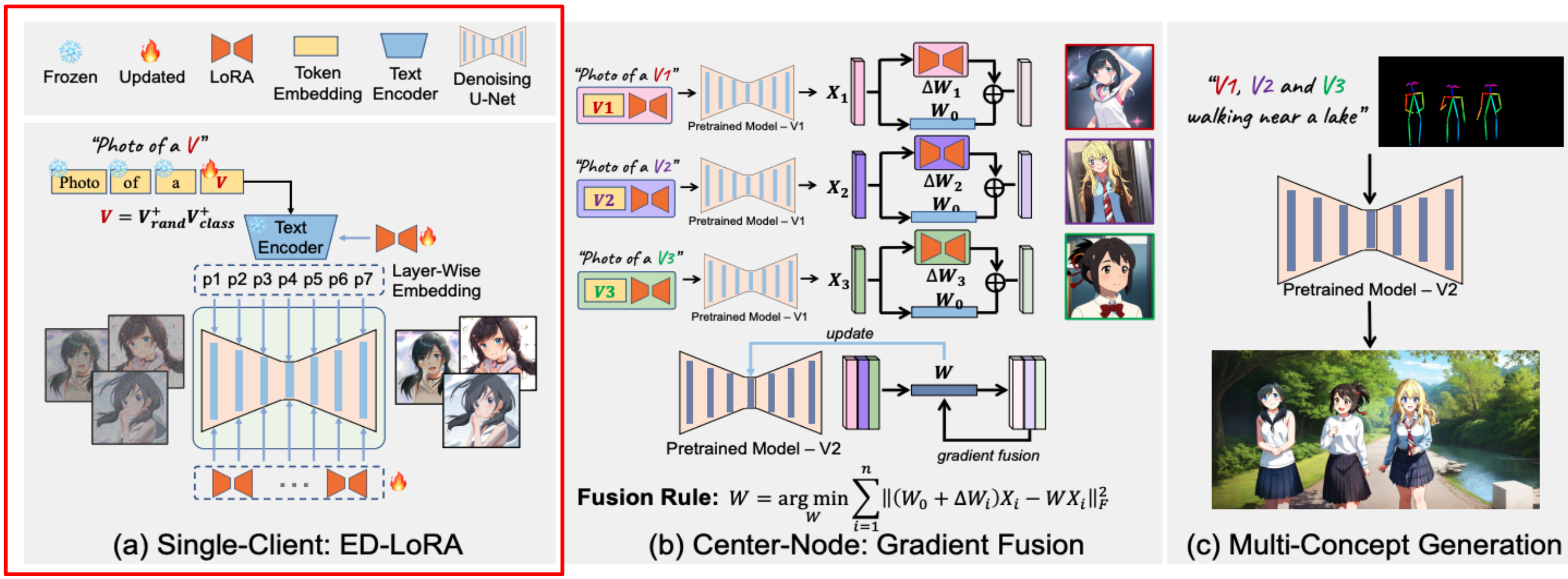


(a) Single-Client: ED-LoRA (b) Center-Node: Gradient Fusion (c) Multi-Concept Generation

[NeurIPS'23 Mix-of-Show: Decentralized Low-Rank Adaptation for Multi-Concept Customization of Diffusion Models]

Multi-concept Personalization: Mix-of-Show, NeurIPS'23

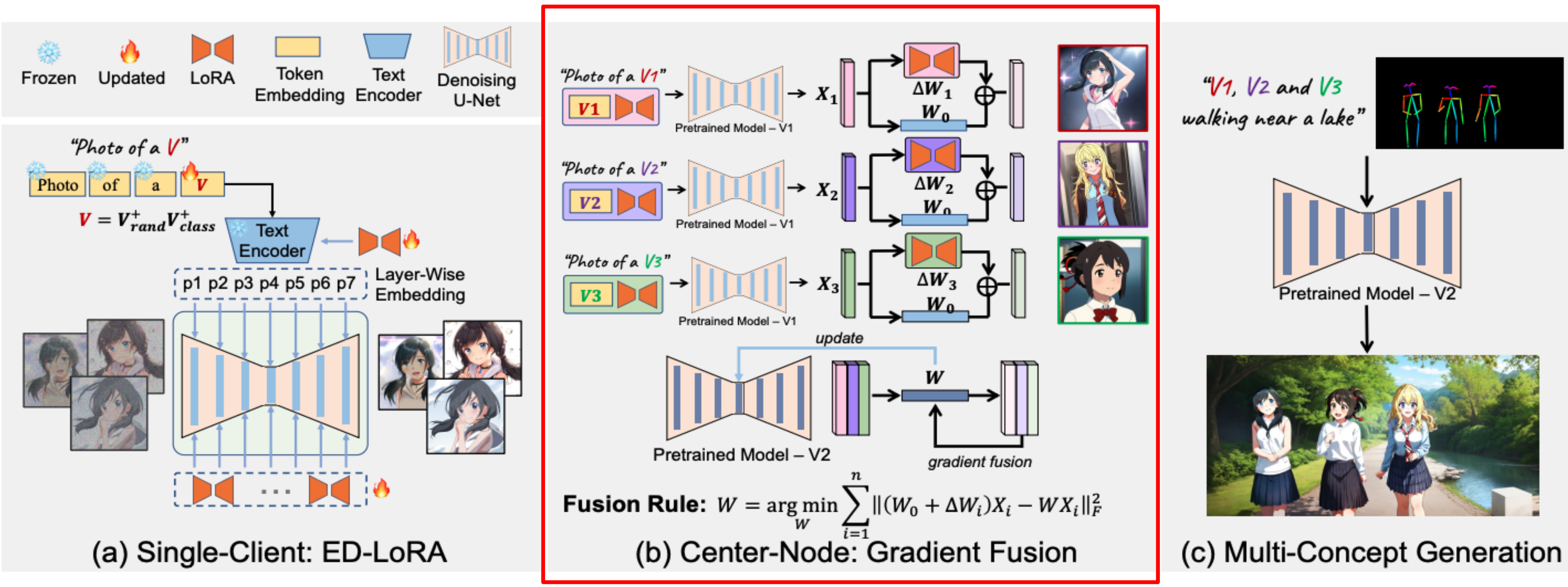
- **Training Method**
  - **Single-Client Learning:** learn LoRA for each concept separately.
  - **Center-Node Fusion**: LoRA fusion
- **Inference**
  - Use the fused LoRA to generate multi-concept images



NeurIPS'23 Mix-of-Show: Decentralized Low-Rank Adaptation for Multi-Concept Customization of Diffusion Models

Multi-concept Personalization: Mix-of-Show, NeurIPS'23

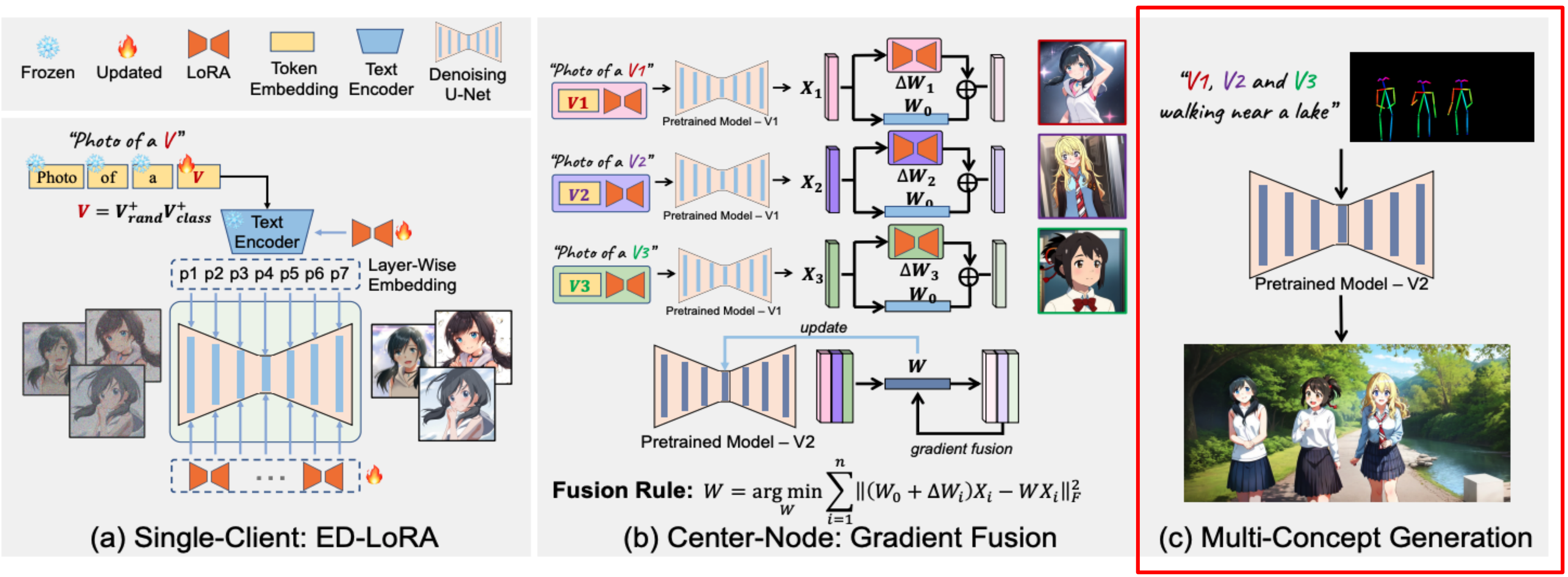
- **Training Method**
  - **Single-Client Learning:** learn LoRA for each concept separately.
  - **Center-Node Fusion**: fuse the above LoRAs into a single LoRA…how?
- **Inference**
  - Use the fused LoRA to generate multi-concept images



[NeurIPS'23 Mix-of-Show: Decentralized Low-Rank Adaptation for Multi-Concept Customization of Diffusion Models]

Multi-concept Personalization: Mix-of-Show, NeurIPS'23

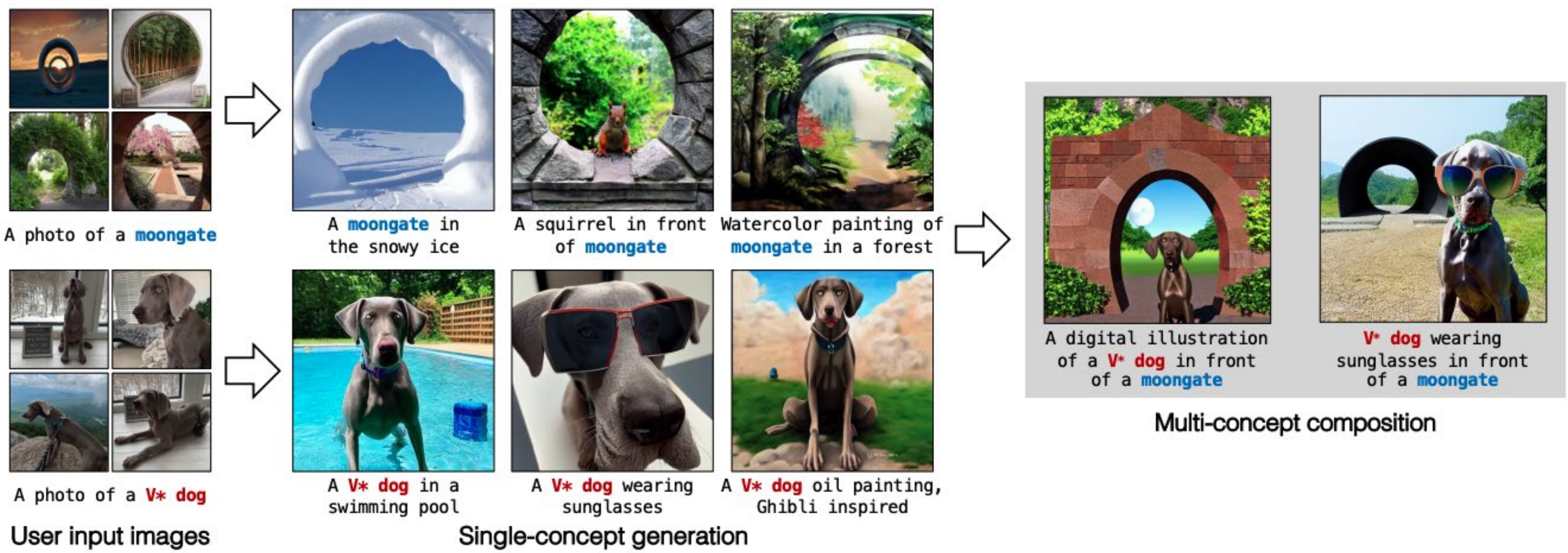
- **Training Method**
  - **Single-Client Learning:** learn LoRA for each concept separately.
  - **Center-Node Fusion**: fuse the above LoRAs into a single LoRA…how?
- **Inference**
  - Use the fused LoRA to generate multi-concept images



[NeurIPS'23 Mix-of-Show: Decentralized Low-Rank Adaptation for Multi-Concept Customization of Diffusion Models]

Multi-concept Personalization: CustomDiffusion, CMU/Tsinghua/Adobe, CVPR'23

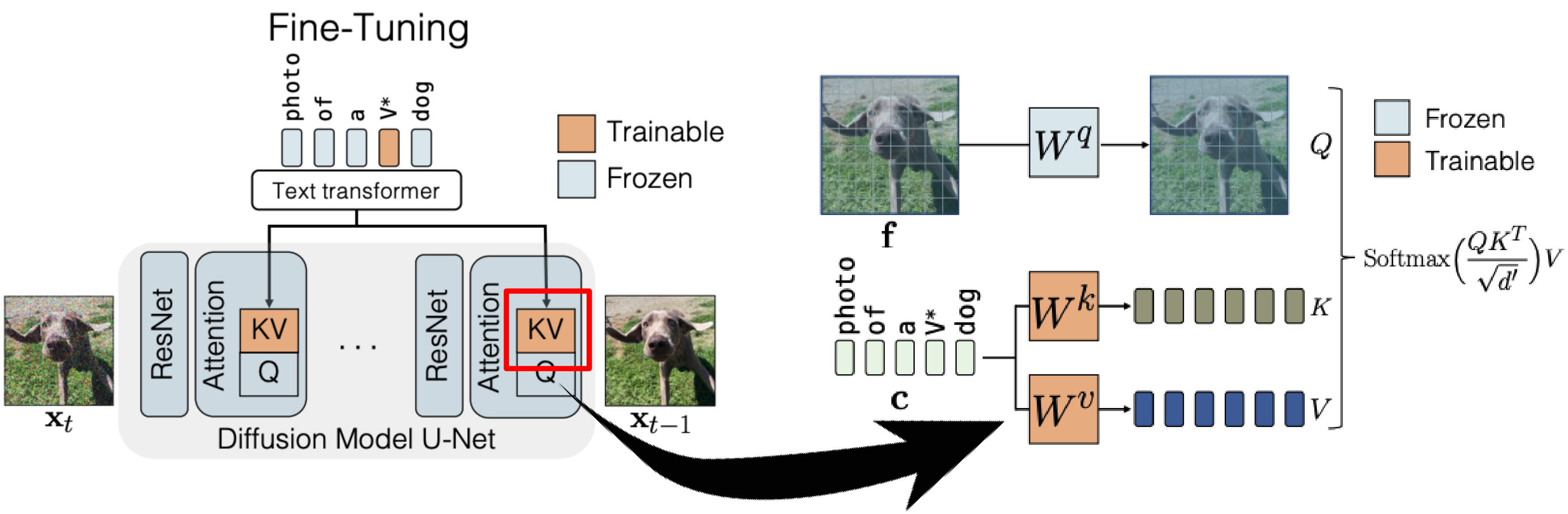
- Similar to Mix-of-Show, CustomDiffusion also has two stages:
  - Stage 1: Fine-tune for each concept separately
  - Stage 2: Merge different concepts into one



[CVPR'23 CustomDiffusion: Multi-Concept Customization of Text-to-Image Diffusion]

Multi-concept Personalization: CustomDiffusion (cont'd)

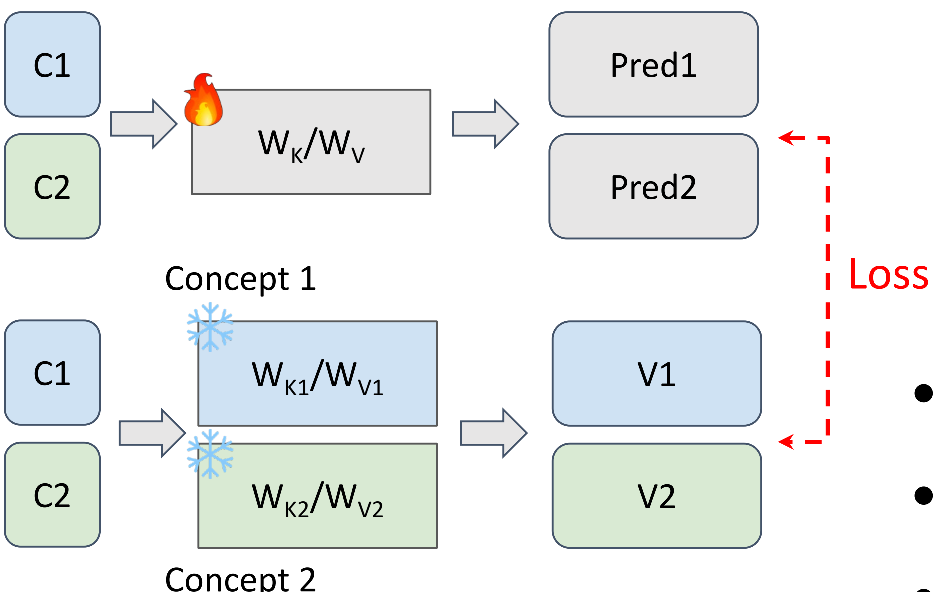
- Similar to Mix-of-Show, CustomDiffusion also has two stages:
  - **Stage 1: Fine-tune V* & cross-attention layer for each concept separately**
  - Stage 2: Merge different concepts into one



[CVPR'23 CustomDiffusion: Multi-Concept Customization of Text-to-Image Diffusion]

Multi-concept Personalization: CustomDiffusion (cont'd)

- **Stage 1**: Fine-tune for each concept separately
- **Stage 2**: Merge different concepts into one
    - Fuse W_{Ki} & W_{Vi} of different concepts into a single W_K & W_V by minimizing the loss below
    - It's a closed-form solution since the WK & WV are linear matrices (via Lagrange multiplier)
      ~4x faster than Mix-of-Show for merging multiple concepts



$$\hat{W} = \underset{W}{\arg\min} ||WC_{\text{reg}}^\top - W_0 C_{\text{reg}}^\top||_F$$

$$\text{s.t. } \boxed{WC^\top = V,} \text{ where } C = [\mathbf{c}_1 \cdots \mathbf{c}_N]^\top$$

$$\text{and } V = [W_1 \mathbf{c}_1^\top \cdots W_N \mathbf{c}_N^\top]^\top .$$

- C & V are the <u>input and output</u> of **different concepts**
- C_{reg} is text feature sampled from regularization data & W_0 as pretrained model
- W is trained to fit all concepts.

CVPR'23 CustomDiffusion: Multi-Concept Customization of Text-to-Image Diffusion

Diffusion Model for Personalization

- (Recap) Single Concept
  - Textual Inversion, ICLR’23
  - DreamBooth, CVPR’23
- Multiple Concepts
  - CustomDiffusion, CVPR’23
  - Mix-of-Show, NeurIPS’23
- **Beyond Image: Video Motion Customization**
  - Video Motion Customization, CVPR’24

Video Motion Customization (VMC)

KAIST, CVPR'24

- **Task**
  - **Given**: reference video + text prompt
  - **Output**: video that matchs
    (1) the motion of reference video & (2) the semantic of text prompt

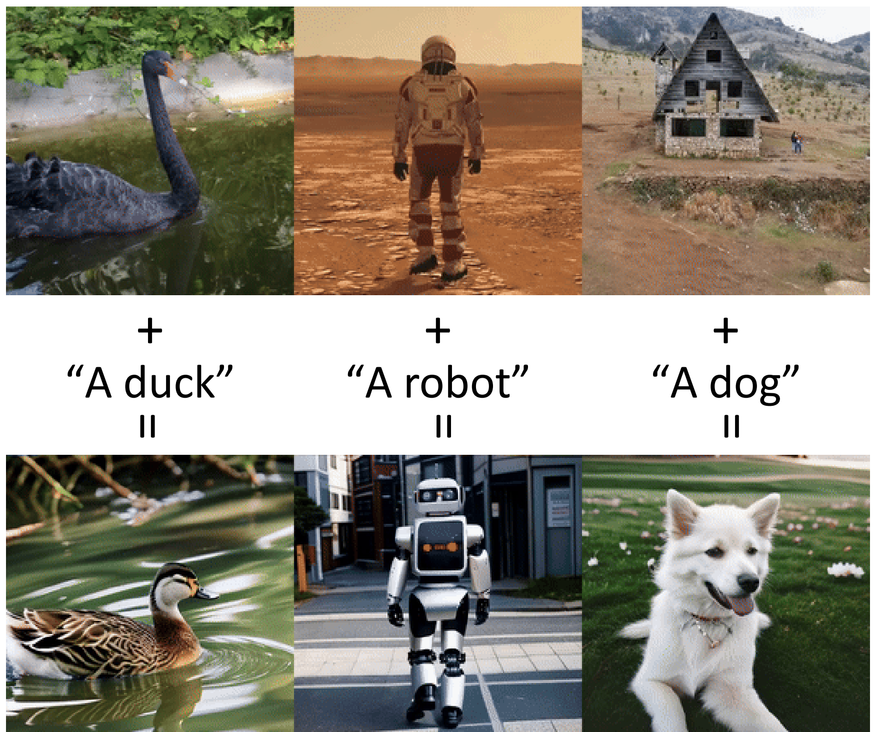
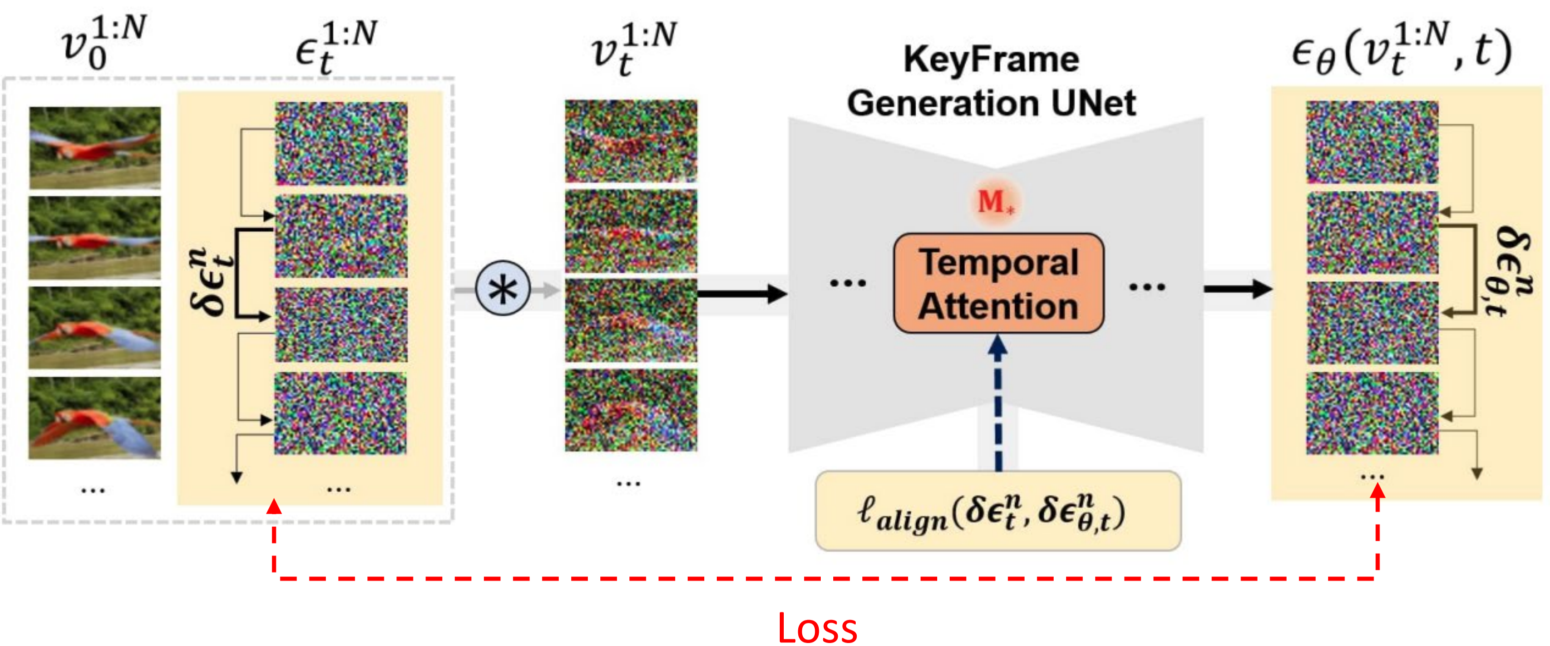


Figure Credit: [MotionClone]

[CVPR'24 VMC: Video Motion Customization using Temporal Attention Adaption for Text-to-Video Diffusion Models]

Video Motion Customization (VMC)

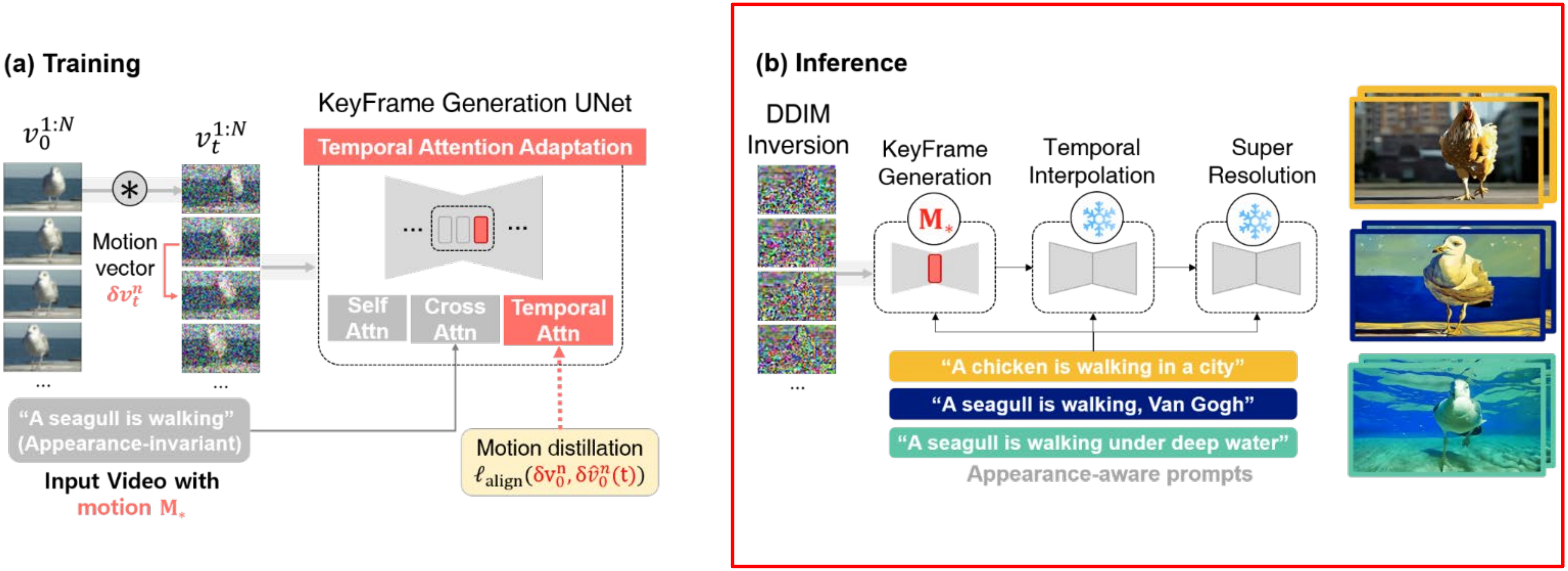
- **Training**
  - Instead of calculating the MSE loss between noise prediction and GT, VMC calculates that btw the **noise residual** of prediction and GT
  - To focus on learning temporal info (i.e., motion), only fine-tune the temporal attention layer



CVPR'24 VMC: Video Motion Customization using Temporal Attention Adaption for Text-to-Video Diffusion Models

Video Motion Customization (VMC)

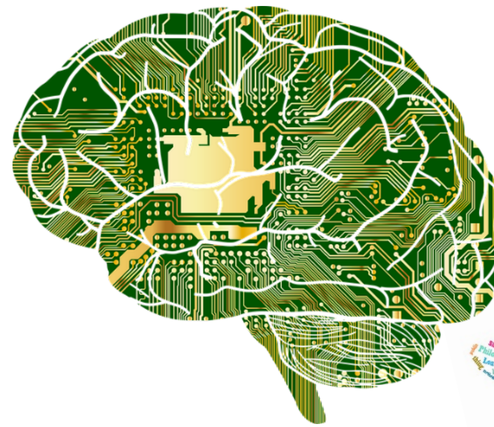
- **Inference**
  - **Input**: DDIM inversed noise of reference video + text prompt
  - **Output**: output video with desirable motion + appearance

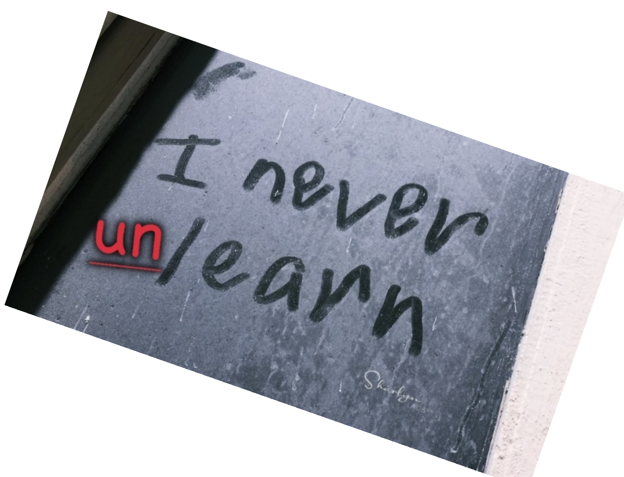


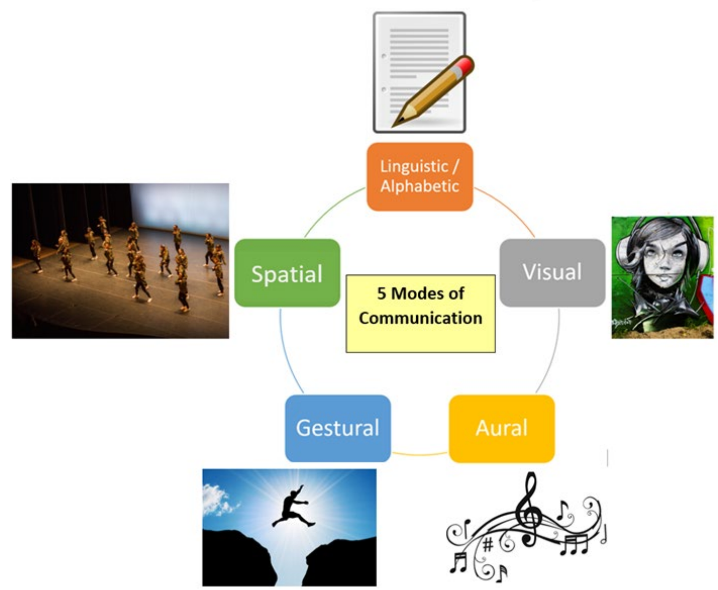
[CVPR'24 VMC: Video Motion Customization using Temporal Attention Adaption for Text-to-Video Diffusion Models]

What to Be Covered Today…

- Multimodal LLM
- **Advanced Topics in LLM/VLM**
  - Concept Editing
  - Concept Unlearning
  - Personalization
  - Continual Learning



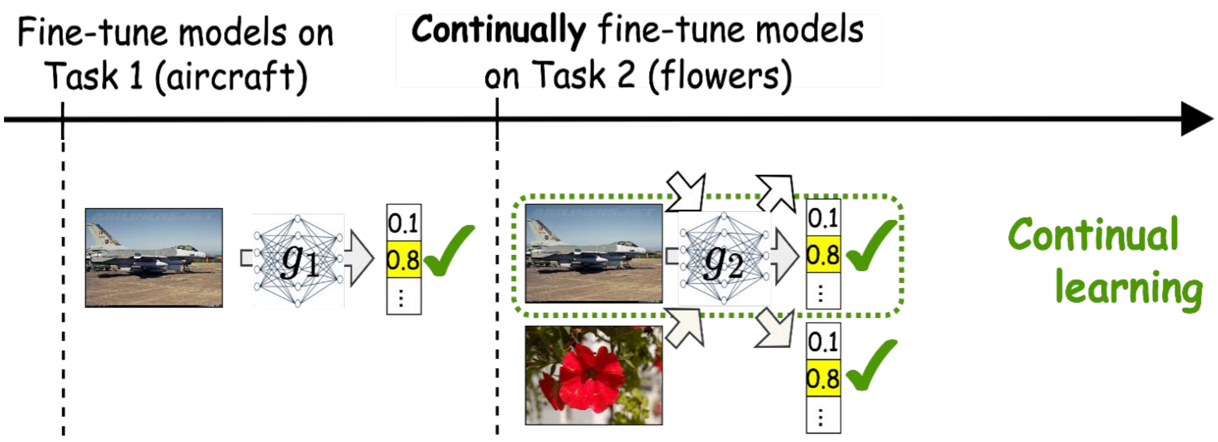




Continual Learning (aka Incremental Learning)

- **Motivation**
  - Always new dataset, knowledge, etc, to finetune the LLM/VLM
    - No practical to re-train foundation models from scratch
  - It is a naive learning way, since **human is a continual learner**.

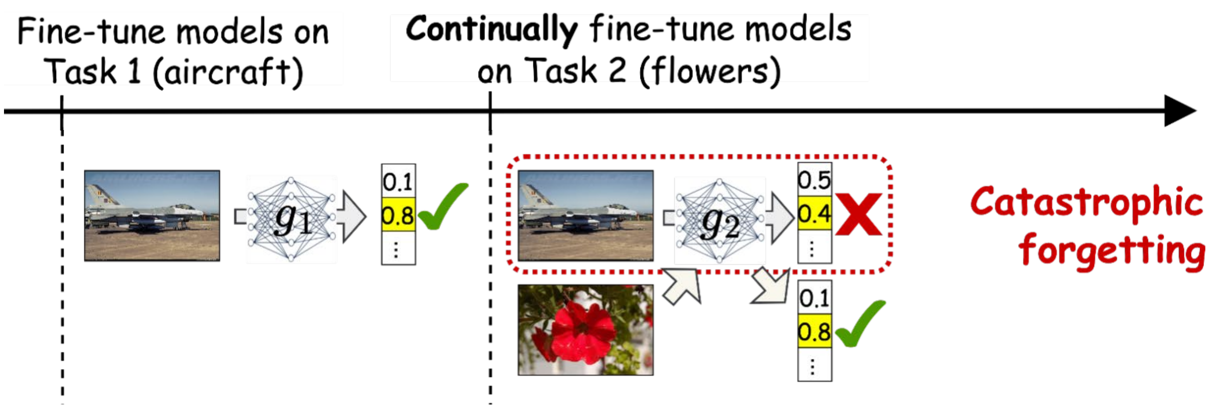
➔ Goal: learn downstream tasks/datasets in a sequential (or incremental) way, while not forgetting what models have learned before.



Continual Learning (cont'd)

- **Task Definition**
  - Learning a list of datasets in a sequential manner **without forgetting** previous knowledge.

- **The most straight forward strategy**
  - Directly fine-tune a pre-trained model on a new dataset
  - **Challenge:** Suffer from the well-known **catastrophic forgetting issue**,
    as the model weights can be totally distorted toward the new task only



Previous works on Continual Learning

- **Rehearsal-based methods**
  - iCaRL (CVPR'17)
- **Regularization-based methods**
  - EWC (PNAS'17)
- **Continual Learning for open-vocab. Vision-Language Models**
  - ZSCL (ICCV'23)
  - Select and Distill (ECCV'24)

iCaRL: Incremental Classifier and Representation Learning, Oxford, CVPR'17

- **Rehearsal-based method**
- **Idea:**
  - Maintain a subset of previous data in a class exemplar sets $P = (P_1, \cdots, P_{s-1})$ where {1, 2, …, k-1} are the learned classes
  - Joint training with the current data $X^s, \ldots, X^t$ with classes {s, …, t}
- **Method**
  - For data in P, enforce the learned model θ output as that of θ_{old}.
    - Can be viewed as **Knowledge Distillation**
  - For the newer data, training with the standard cross entropy loss.

$$Y_{\text{old}} = \{f_{\theta_{\text{old}}}(x) | \forall x \in P\}$$

$$\mathcal{L}(\theta) = \sum_{(x,y)\in D} \left[\sum_{y=s}^{t} \mathcal{L}(Y_{\text{new}}, \hat{Y}) + \sum_{y=1}^{s-1} \mathcal{L}(Y_{\text{old}}, \hat{Y}) \right]$$

iCaRL: Incremental Classifier and Representation Learning

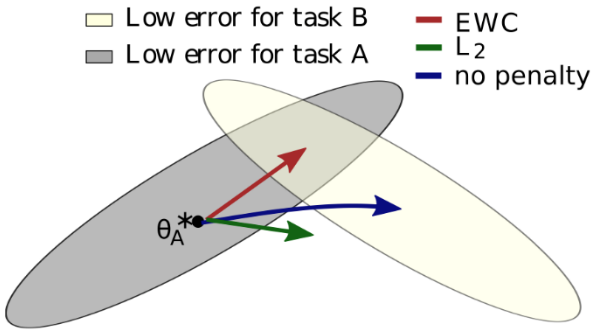
EWC: Overcoming catastrophic forgetting in neural networks, DeepMind, PNAS'17

- **Regularization-based method**
- **Idea:**
  - Weight Consolidation: Restrict the updated weights not to be too different from the original model weights.

$$\mathcal{L}_{\mathrm{WC}} = \sum_i (\theta_i - \bar{\theta}_i)^2$$

  - Elastic Weight Consolidation: Each parameter should not be restricted with the same weights
    - i: the index of the model parameters.

$$\mathcal{L}_{\mathrm{EWC}} = \sum w_i \cdot (\theta_i - \bar{\theta}_i)^2$$



[Overcoming catastrophic forgetting in neural networks](#)

EWC, DeepMind, PNAS’17 (cont’d)

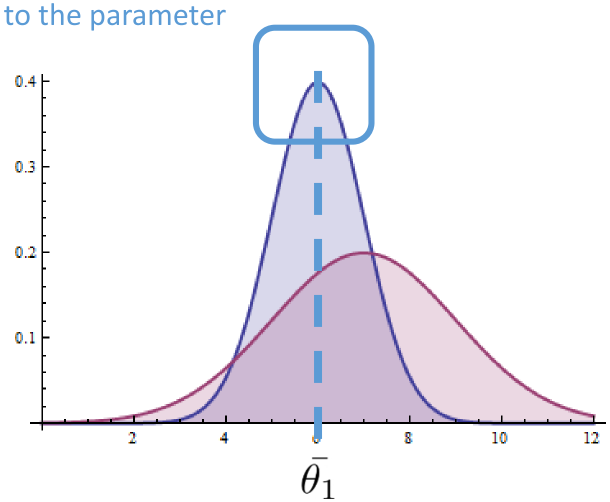
- **Method (cont’d)**
  - Using Fisher Information (F) to determine the importance of a parameter to the previous task.
    - Fisher information: the expectation of second derivative of negative log-likelihood at $\bar{\theta}$

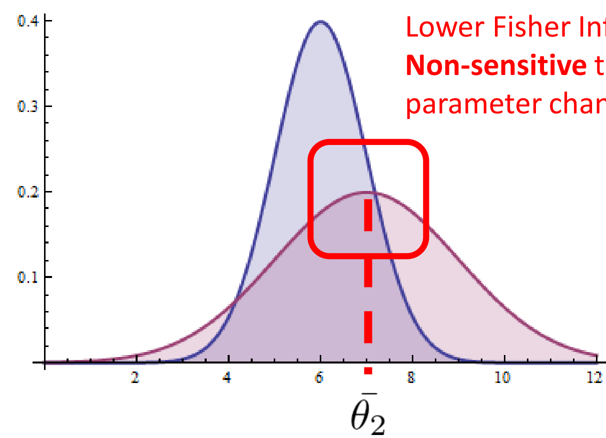
$$\mathcal{L}_{\text{EWC}} = \sum_i \frac{\lambda}{2} F_i (\theta_i - \bar{\theta}_i)^2$$

    - λ: a hyper-parameter to determine the overall importance of previous tasks.

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i (\theta_i - \theta^*_{A,i})^2$$

Higher Fisher Information. **Sensitive** to the parameter change



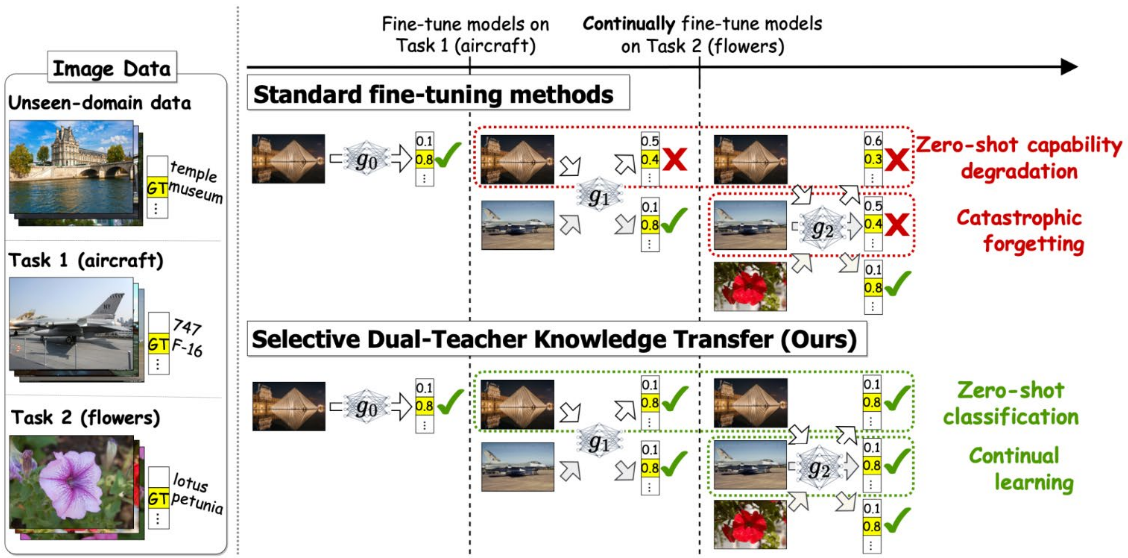


Lower Fisher Information. **Non-sensitive** to the parameter change.

Overcoming catastrophic forgetting in neural networks

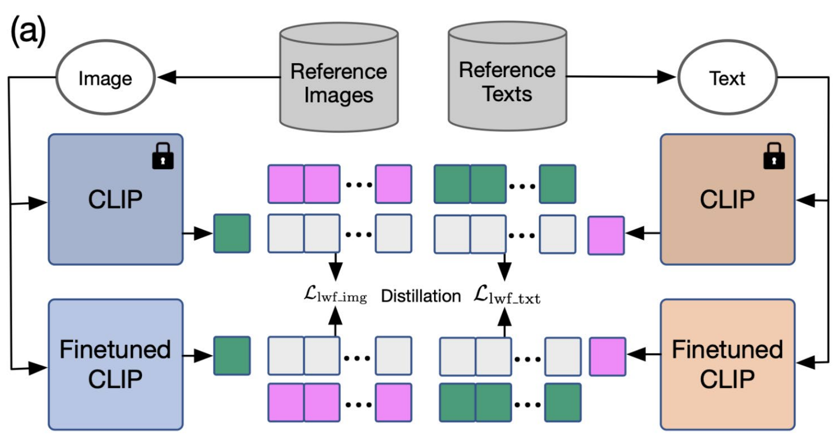
Continual Learning for
Open-Vocab Vision-Language Models

- **Motivation**
  - With the prevalence of large-scale Vision-Language Models (VLMs), Continual Learning for VLMs has emerged as a potential research trends.
- **Goal**
  - Sequentially learning from new datasets
  - Preserve the original zero-shot ability for unseen data
  - Maintain knowledge learned from previous stages (as existing continual learning methods do)



ZSCL: Preventing Zero-Shot transfer degradation in Continual Learning of vision-language models, NUS, ICCV’23

- **Method**
  - Utilize an auxiliary reference dataset (e.g., ImageNet), and perform **Knowledge Distillation** from **the original CLIP model.**
    - (1) Distill knowledge on **both visual and textual sides**.



Teacher (Original CLIP) output

$$\mathcal{L}_{\text{dist\_img}} = \text{CE}(\boldsymbol{p}, \overline{\boldsymbol{p}}) = -\sum_{j=1}^{m} \boxed{\boldsymbol{p}_j} \cdot \log \boxed{\overline{\boldsymbol{p}}_j}$$

Student output

[Preventing Zero-Shot transfer degradation in Continual Learning of vision-language models]

ZSCL, NUS, ICCV’23 (cont’d)

- **Method (cont’d)**
  - (2) WE: Weight space Ensemble to regularize the weights
    - The learned model weights would not be too different from the weights of the previous stage

$$\hat{\theta}_t = \begin{cases} \theta_0 & t = 0 \\ \frac{1}{t+1}\theta_t + \frac{t}{t+1} \cdot \hat{\theta}_{t-1} & \text{every I iterations} \end{cases} .$$

    - Same form as EMA (exponetial moving average)
  - Training strategy: (1) -> (2) -> (1) -> (2) -> …

(1) $\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda \cdot (\mathcal{L}_{\text{lwf\_img}} + \mathcal{L}_{\text{lwf\_txt}})$

Data from novel task + auxiliary ref dataset

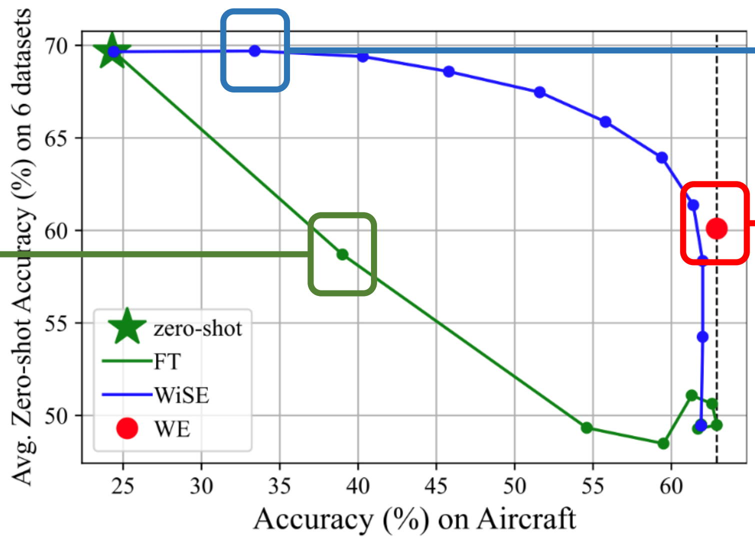
(2) $$\hat{\theta}_t = \begin{cases} \theta_0 & t = 0 \\ \frac{1}{t+1}\theta_t + \frac{t}{t+1} \cdot \hat{\theta}_{t-1} & \text{every I iterations} \end{cases} .$$

Preventing Zero-Shot transfer degradation in Continual Learning of vision-language models

ZSCL, NUS, ICCV’23 (cont’d)

- **Comparisons**
  - Zero-shot accuracy vs. accuracy on novel task

Sample every 100 iterations. As training progresses, the model's zero-shot capability deteriorates.



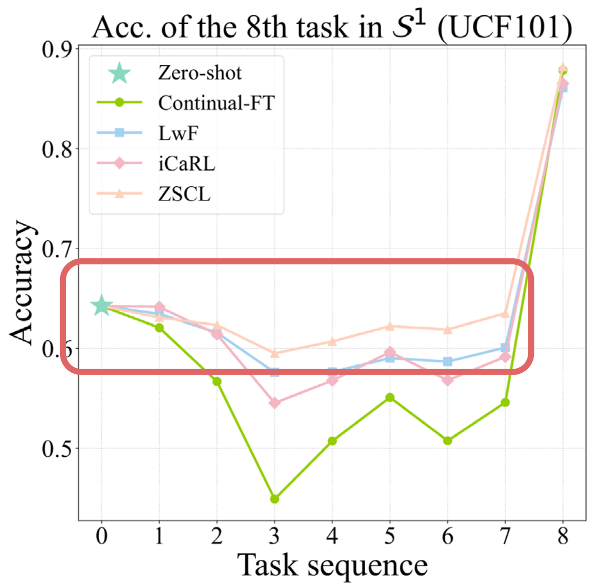
Ensemble the original and fine-tuned model weights with different ratios. **Sensitive to the choose of the ratio**

Iterative weight ensemble. Improved fine-tuned accuracy, with an acceptable decrease in zero-shot performance

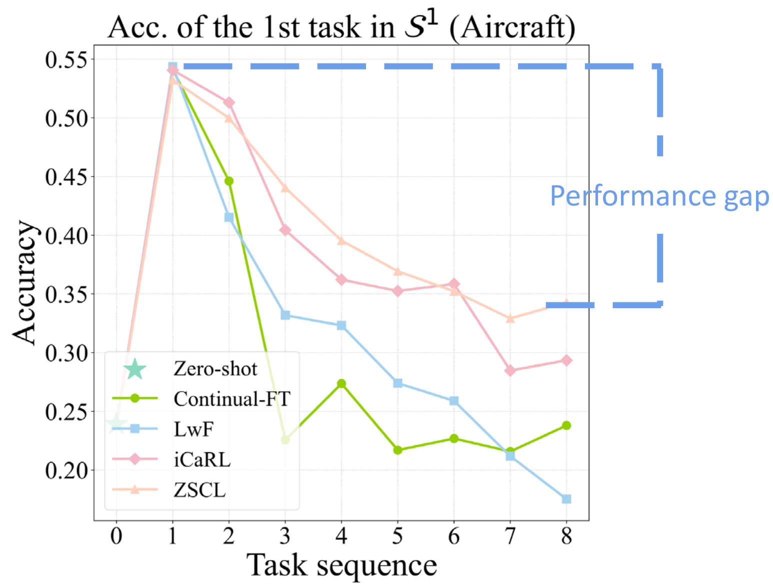
[Preventing Zero-Shot transfer degradation in Continual Learning of vision-language models]

ZSCL, NUS, ICCV’23 (cont’d)

- **Limitation**
  - ZSCL still largely **suffer from catastrophic forgetting for previous tasks.**



ZSCL can preserve zero-shot ability for unseen data

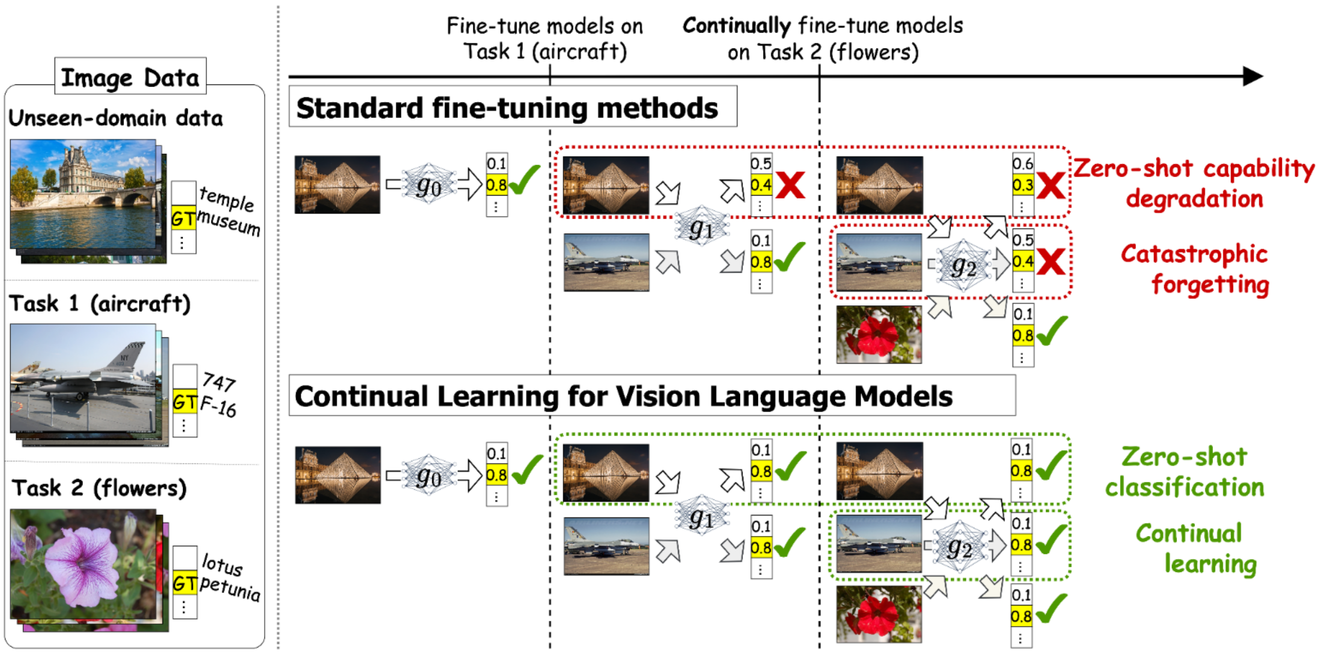


There is still a gap for previous task after training on multiple datasets

[Preventing Zero-Shot transfer degradation in Continual Learning of vision-language models]

Select and Distill: Selective Dual-Teacher Knowledge Transfer for Continual Learning on Vision-Language Models, NTU, ECCV'24

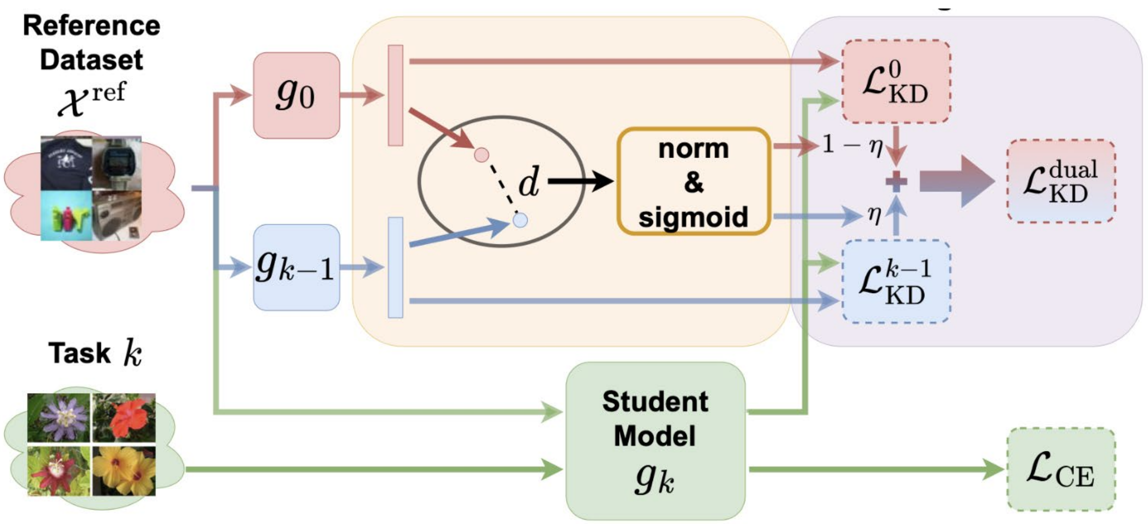
- **Goal**
  - Same as ZSCK, adapt to new datasets sequentially while:
    - preserving the original pre-trained zero-shot ability
    - maintaining the knowledge learned from previous stages.



Select and Distill: Selective Dual-Teacher Knowledge Transfer for Continual Learning on Vision-Language Models

Select and Distill, NTU, ECCV'24 (cont'd)

- **Idea**
  - Follow ZSCL, utilize a reference dataset for knowledge distillation
  - Dual-Teacher Knowledge Distillation
    - Distill from **the original pre-trained VLM** to preserve **zero-shot ability**.
    - Distill from **the most recent fine-tuned VLM** to preserve **prior knowledge**.
- **Key**
  - For any data point in the reference dataset,
    we need to **select a proper model** and distill its knowledge.



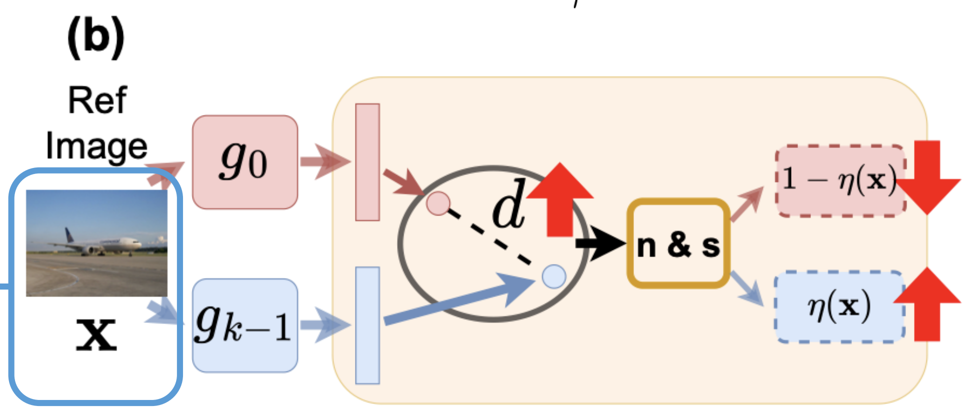
[Select and Distill: Selective Dual-Teacher Knowledge Transfer for Continual Learning on Vision-Language Models]

Select and Distill, NTU, ECCV'24 (cont'd)

- **Observation**
  - If a data point is a previously learned data.
    - It must be seen by g_{k-1}, but never been seen by g_0
    - The **feature distance d** between g_0 and g_{k-1} can be large or small?
    - Select g_{k-1} as the teacher model to maintain previous knowledge
  - $\eta(\mathbf{x})$: A normalized distance between 0~1, determine how much should we distill from g_{k-1}

$$\eta(\mathbf{x}) = \sigma(\frac{d(g_{k-1}(\mathbf{x}), g_0(\mathbf{x})) - \delta}{\gamma}),$$

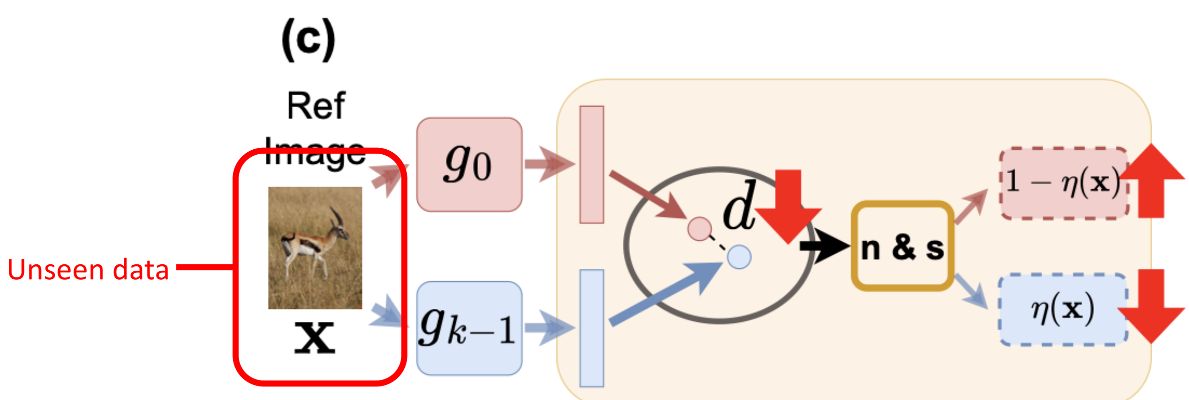
Previously learned data



Select and Distill: Selective Dual-Teacher Knowledge Transfer for Continual Learning on Vision-Language Models

Select and Distill, NTU, ECCV'24 (cont'd)

- **Observation**
  - If a data point has never been seen by both g_{k-1} and g_0 (unseen data)
    - The **feature distance d** between g_0 and g_{k-1} can be **relatively small**
    - In this case, we should select g_0 as the teacher model to preserve the original zero-shot ability.



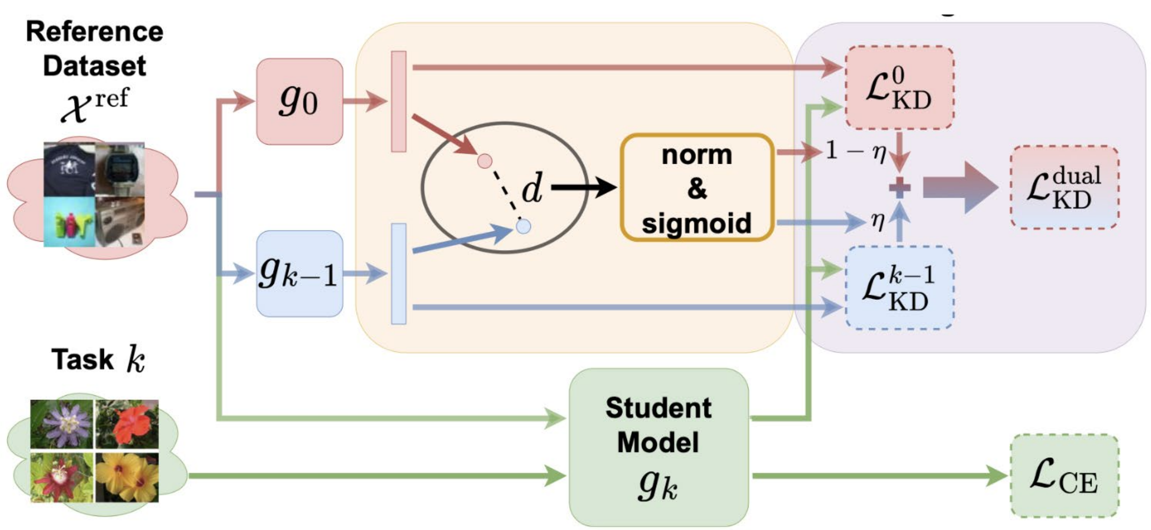
Select and Distill: Selective Dual-Teacher Knowledge Transfer for Continual Learning on Vision-Language Models

Select and Distill, NTU, ECCV'24 (cont'd)

- **Objective**

$$\mathcal{L}_{\text{KD}}^{k-1} = d(g_{k-1}(\mathbf{x}), g_k(\mathbf{x})) \ , \ \ \mathcal{L}_{\text{KD}}^{0} = d(g_0(\mathbf{x}), g_k(\mathbf{x}))$$

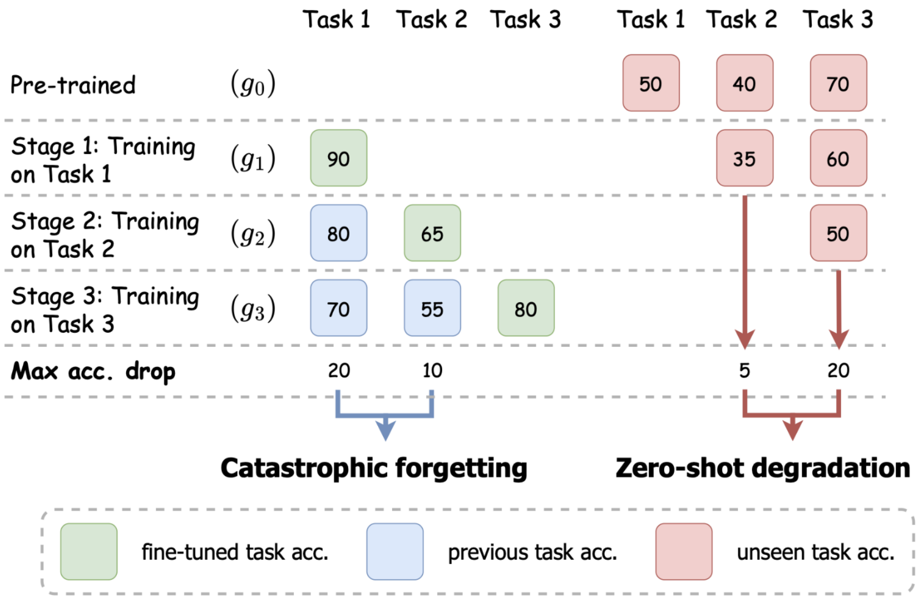
$$\mathcal{L}_{\text{KD}}^{\text{dual}} = \sum_{\mathbf{x}\sim\mathcal{X}^{\text{ref}}} \eta(\mathbf{x}) \cdot \mathcal{L}_{\text{KD}}^{k-1} + (1-\eta(\mathbf{x})) \cdot \mathcal{L}_{\text{KD}}^{0}$$



Select and Distill: Selective Dual-Teacher Knowledge Transfer for Continual Learning on Vision-Language Models

Select and Distill, NTU, ECCV'24 (cont'd)

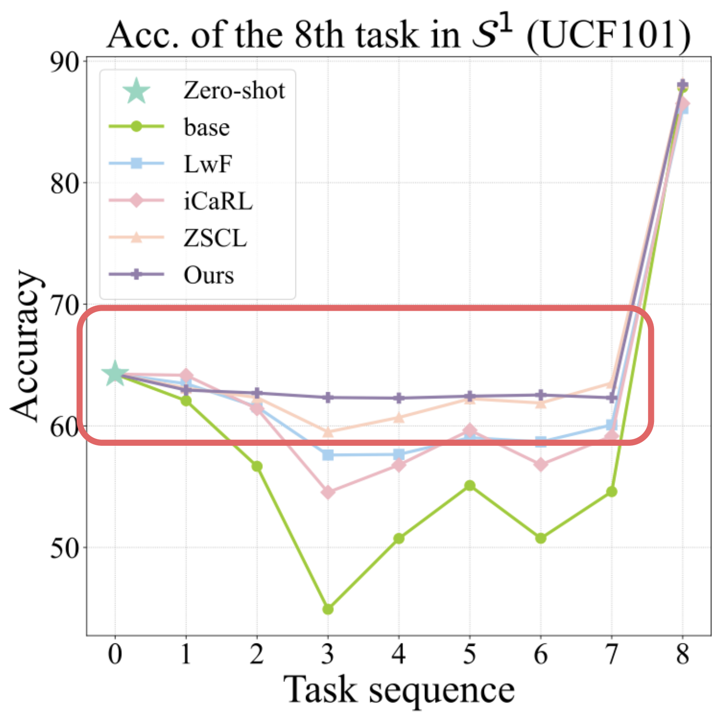
- **Metrics**
  - Average Accuracy
    - Average of the last performance on each dataset
  - Catastrophic forgetting
    - Max. performance gap after the task has been fine-tuned
  - Zero-shot degradation
    - Max. performance gap before the task has been fine-tuned



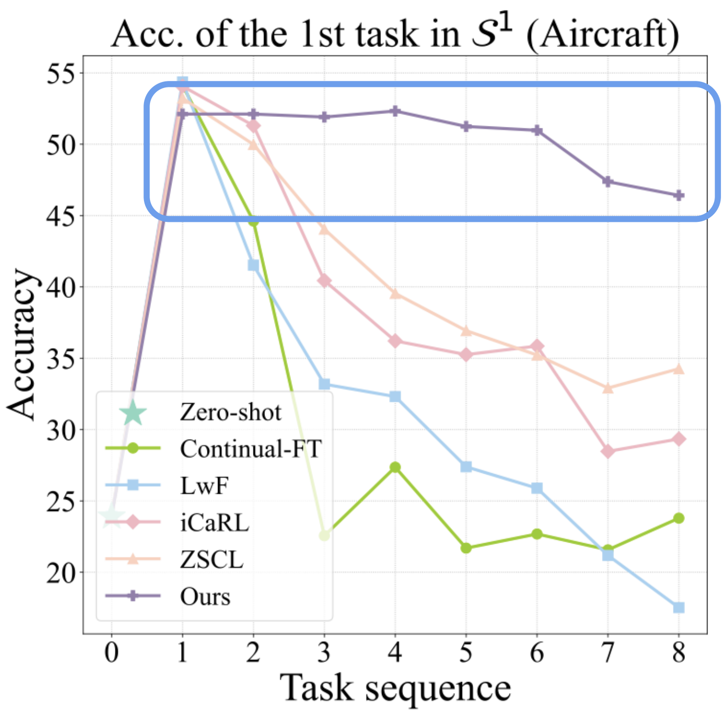
[Select and Distill: Selective Dual-Teacher Knowledge Transfer for Continual Learning on Vision-Language Models]

Select and Distill, NTU, ECCV’24 (cont’d)

- **Results**
  - Successfully preserve the zero-shot ability for unseen data
  - Mitigate the catastrophic forgetting of previously learned data



Successfully preserve zero-shot ability for unseen data



Largely mitigate the performance gap

[Select and Distill: Selective Dual-Teacher Knowledge Transfer for Continual Learning on Vision-Language Models]

Select and Distill, NTU, ECCV'24 (cont'd)

- **Robustness**
  - We shuffle the training orders, producing 8 different training sequences.
  - Our methods showing state-of-the-art performance on all metrics, and the results are stable across all training sequences.

Method / Sequence	$\mathcal{S}^1$	$\mathcal{S}^2$	$\mathcal{S}^3$	$\mathcal{S}^4$	$\mathcal{S}^5$	$\mathcal{S}^6$	$\mathcal{S}^7$	$\mathcal{S}^8$	**Mean**
Catastrophic forgetting (↓)									
Continual FT	10.98	10.60	8.80	19.17	10.11	11.95	15.19	9.48	12.04
LwF [24]	10.38	6.52	6.37	10.22	7.99	7.70	10.41	8.91	8.56
iCaRL [35]	8.42	7.00	6.45	10.21	7.03	7.33	9.68	8.23	8.04
ZSCL [50]	4.67	2.35	2.13	2.97	3.15	4.28	4.89	4.70	3.64
MoE-Adapters [48]	2.74	4.71	4.28	1.15	1.50	1.60	2.94	2.77	2.71
Ours	**1.70**	**1.16**	**0.89**	**1.04**	**0.59**	**1.34**	**1.12**	**1.79**	**1.20**
Zero-shot degradation (↓)									
Continual FT	24.81	23.58	19.54	16.46	22.22	19.02	19.54	24.02	21.15
LwF [24]	10.75	10.23	8.63	8.25	12.02	10.33	8.98	11.01	10.03
iCaRL [35]	13.77	12.68	11.28	12.14	13.20	13.20	13.09	14.01	12.92
ZSCL [50]	3.44	3.94	4.02	2.85	3.79	2.31	1.86	1.84	3.00
MoE-Adapters [48]	1.62	2.58	**1.04**	2.37	4.31	3.05	**1.77**	**0.63**	2.17
Ours	**1.55**	**2.04**	1.21	**1.92**	**2.79**	**2.18**	1.90	2.08	**1.96**
Average accuracy (↑)									
Continual FT	76.16	76.24	78.03	68.69	76.64	75.44	72.71	77.45	75.17
LwF [24]	76.78	80.45	80.65	77.52	79.64	79.45	77.31	78.70	78.81
iCaRL [35]	77.99	79.77	79.93	76.66	79.26	79.08	77.06	78.61	78.55
ZSCL [50]	81.89	83.98	84.30	83.49	83.41	82.38	81.92	81.97	82.92
MoE-Adapters [48]	82.71	80.74	81.15	83.97	83.68	83.68	82.73	79.68	82.29
Ours	**84.48**	**84.92**	**84.97**	**84.89**	**85.50**	**85.07**	**85.02**	**84.52**	**84.92**

Select and Distill: Selective Dual-Teacher Knowledge Transfer for Continual Learning on Vision-Language Models

What We Have Covered Today…

- **Multimodal LLM**
- **Advanced Topics in LLM/VLM**
    - Concept Editing
    - Concept Unlearning
    - Personalization
    - Continual Learning