Master's Seminar
Advanced Topics in Vision-Language Models

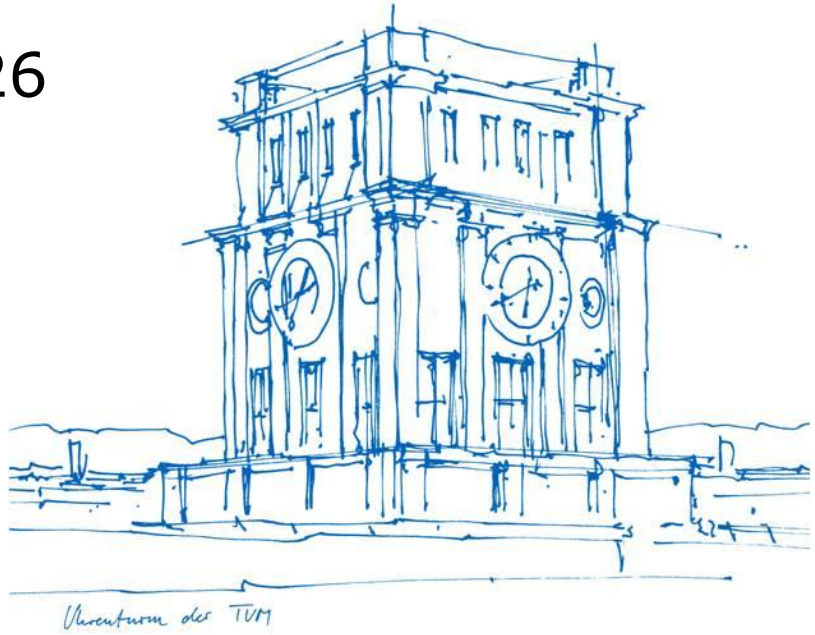
SS 2026

Prof. Dr. Zeynep Akata

Chair for Interpretable and Reliable Machine Learning

TUM School of Computation, Information and Technology

Technische Universität München



What Are the Papers?

In total, we have 16 papers on 4 topics

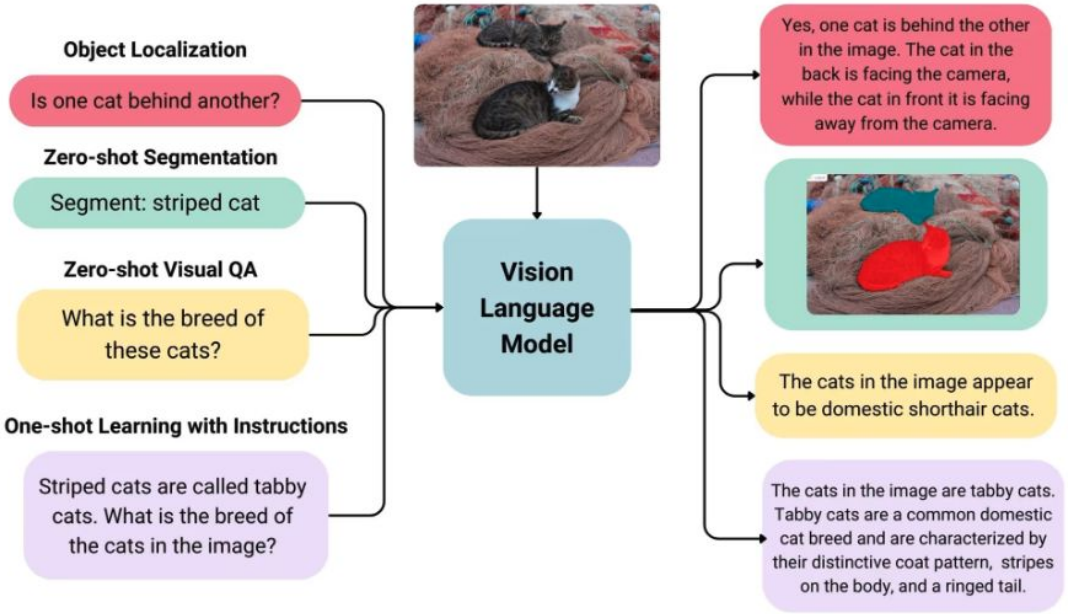
- Foundation Vision-Language Models
- Text-to-Image Models
- Explainability and Mechanistic Interpretability
- Foundation Model Adaptation

1. Foundation Vision-Language Models

1. Foundation Vision-Language Models

Data is often inherently multimodal.

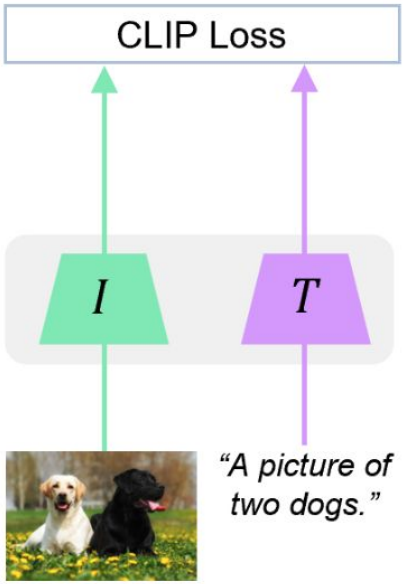
Foundation VLMs: the **large** pretrained models → general-purpose



Source: Joas Pambou, smashingmagazine, 2024

1.1 COSMOS: Cross-Modality Self-Distillation for Vision Language Pre-training

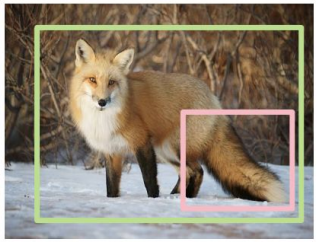
CLIP → Coarse **Alignment**



COSMOS →Image and Text Cropping Strategy

Global

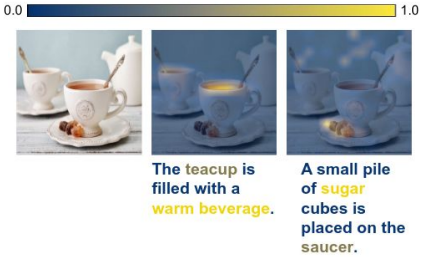
Local



The image captures a red fox, standing alone in a snowy forest. The fox with vibrant red fur is the main subject of the image. The fox is facing the camera. The background of the image is filled with trees and bushes. The snow covers the ground and the trees.

Introduced Loss:

- Local-to-global (img & txt) augmentations correspondence (DINO)
- Cross-modal student-teacher distillation



Much less data →Improves on retrieval, classification, and segmentation

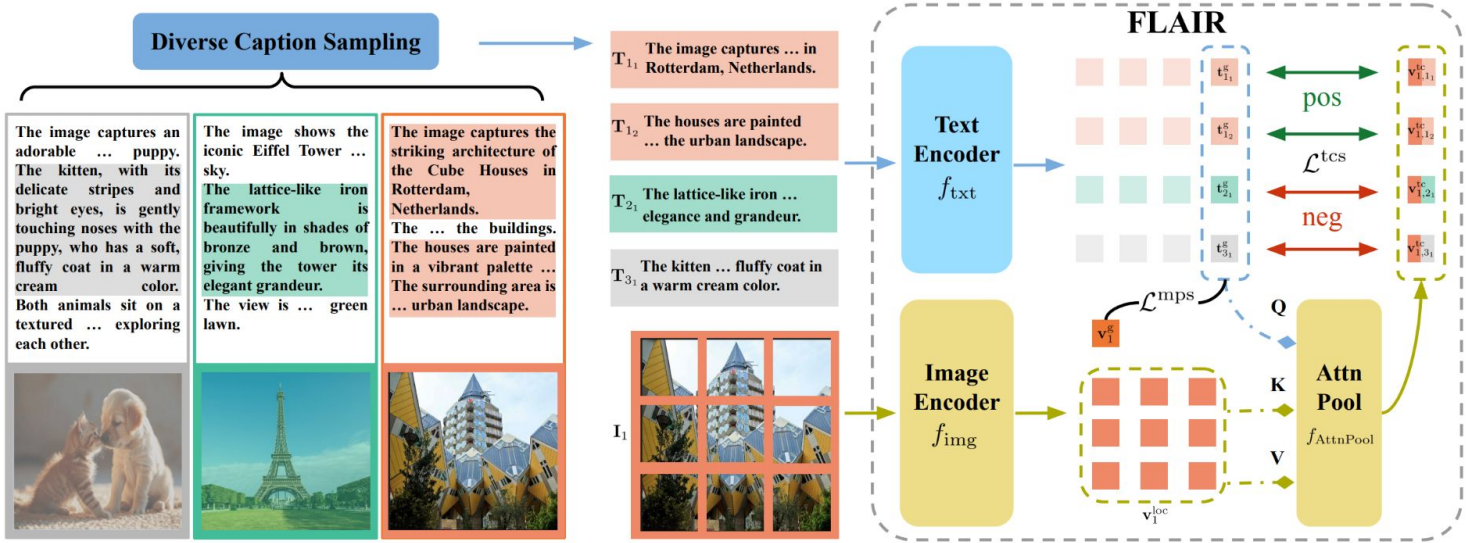
1.2 FLAIR: VLM with Fine-grained Language-informed Image Representations

CLIP → single global image embedding

FLAIR → **text-specific** image embedding

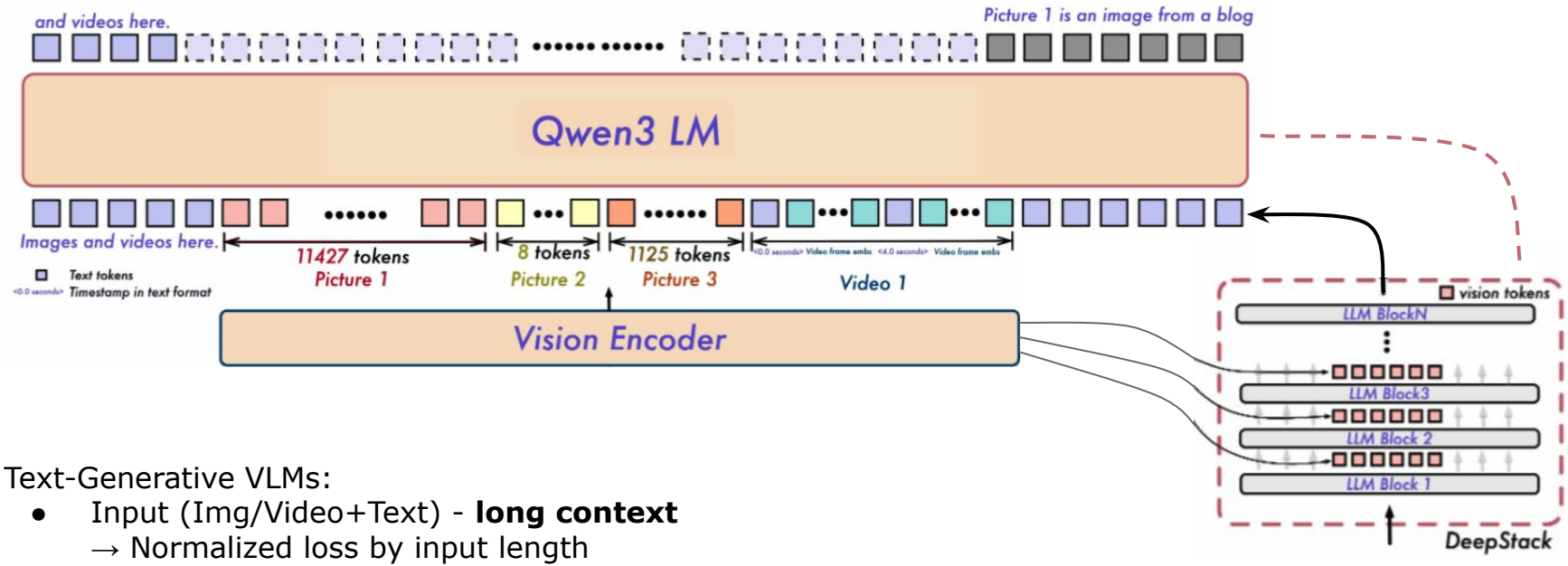
Loss:

- Text-Conditioned Sigmoid loss
- Multi-Positive Sigmoid loss



Improves with less data → fine-grained understanding, zero-shot segmentation

1.3 Qwen3-VL Technical Report



Text-Generative VLMs:

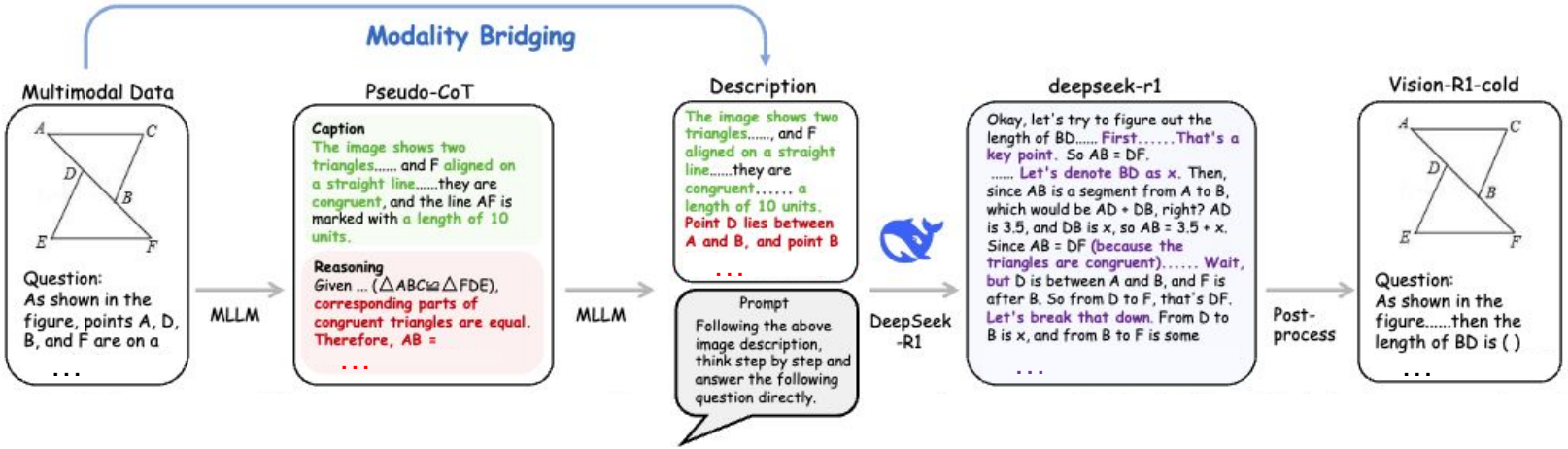
- Input (Img/Video+Text) - **long context**
  → Normalized loss by input length
  → Interleaved MRoPE (know **where** things are in space and time)
  (Multidimensional Rotary Position Embedding)
- DeepStack → **Injects** visual features from **different layers of the ViT** to the LM
- Train using two variants: "thinking" (long CoT) and "non-thinking" (direct answer)

1.4 Vision-R1: Incentivizing Reasoning Capability in Multimodal Large Language Models

Goal: human-like cognitive processes
→ questioning, reflection, and self-inspection.

Problem: direct R1-style RL fails for MLLMs
1. No large-scale high-quality multimodal CoT dataset
2. GRPO training hits the "overthinking" instability → correct reasoning ~ shorter CoT

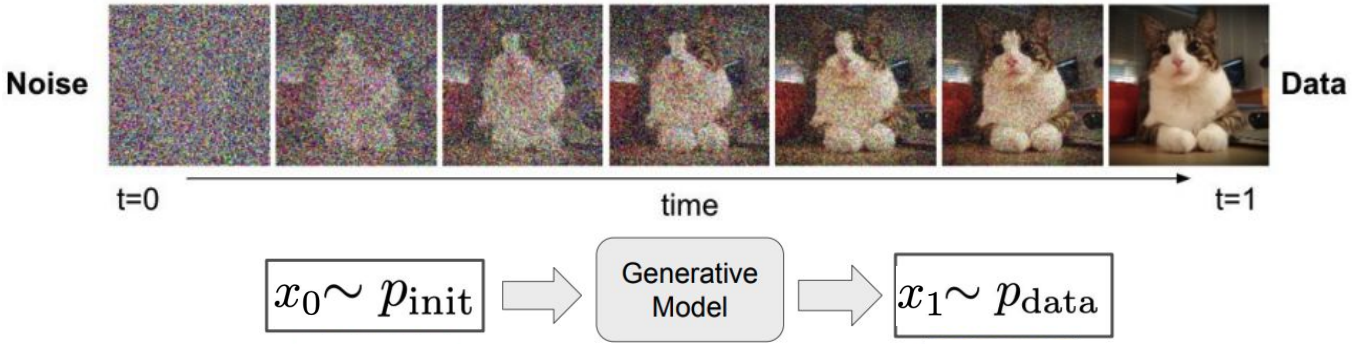
Key: Data curation + training curriculum design



2. Text-to-Image Models

2. Text-to-Image Models

Flow Matching: learn a **velocity field** that transports noise to image along straight paths (ODE)



Source: Peter Holderrieth, MIT, Generative AI with SDE, 2026

Interpolation Path

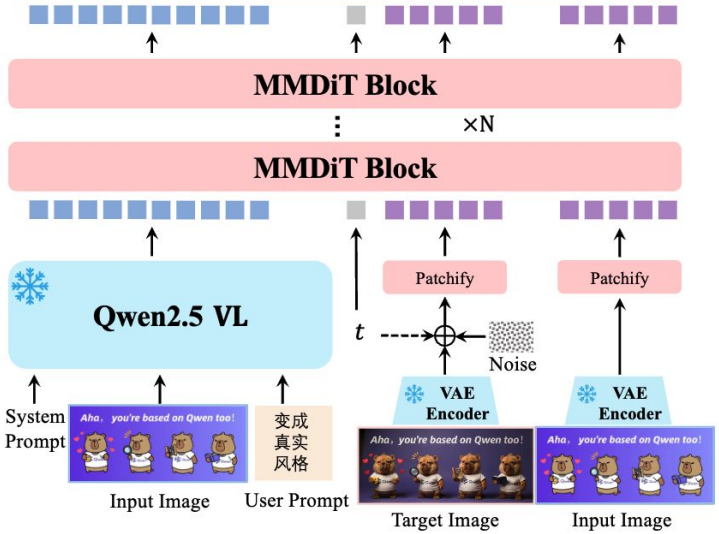
$$x_t = (1-t)\cdot x_0 + t \cdot x_1$$

$$\mathcal{L}(\theta) = \mathbb{E}\left[\| \underbrace{v_\theta(x_t, t)}_{\text{flow model}} - (x_1 - x_0)\|^2\right]$$

Flow matching models can already generate stunning images in 20 to 50 steps.
Now we need more → accurate, fast, and controllable

2.1 Qwen-Image: Multimodal Understanding and Generation

Flow Matching + dual text encoding → unified generation & understanding



Architecture:

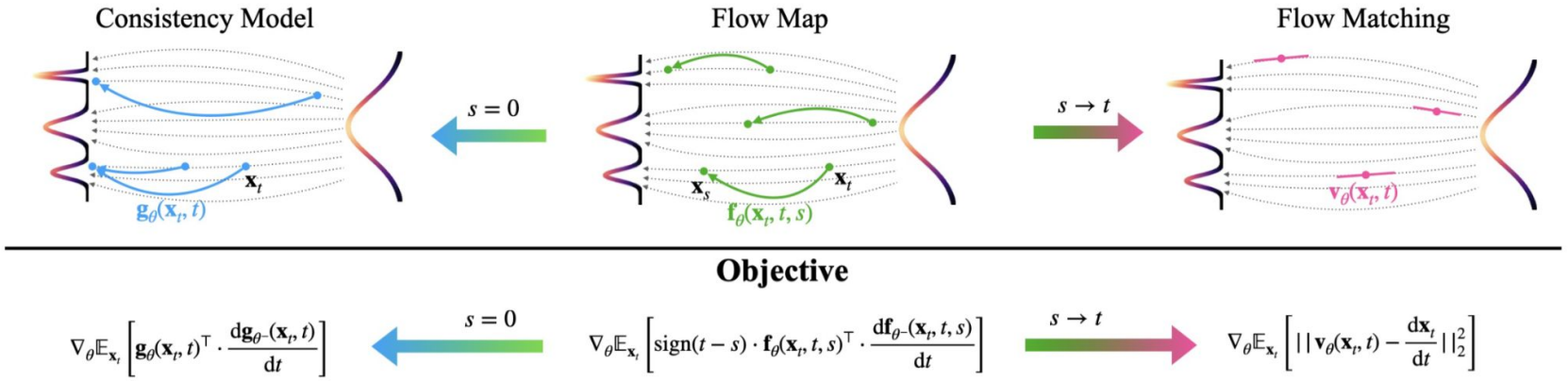
- Decoder: MMDiT (Multimodal Diffusion Transformer)
- Text Encoder: Qwen2.5-VL
- Dual vision-encoding: Qwen2.5-VL + VAE

Multi-task training: T2I, TI2I, I2I

Large-scale data collection + Progressive training
non-text → simple text → complex text → paragraph-level descriptions.

A **twisted** pine trunk leans over the cliff edge, a climber woman **grips the trunk** with two hands, and her partner reaches up, holding onto the woman's **safety belt around her waist.**

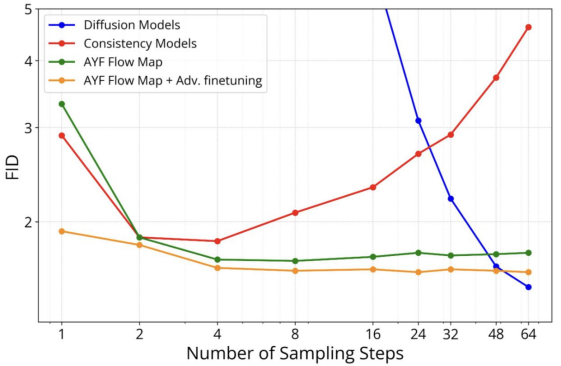
2.2 Align Your Flow: Scaling Continuous-Time Flow Map Distillation



Flow Maps: connect any two noise levels (s,t) in a single step
→ unification of CM and flow matching objectives

Training objectives:
- Eulerian consistency (t=beginning of interval)
- Lagrangian consistency (t=end of interval)

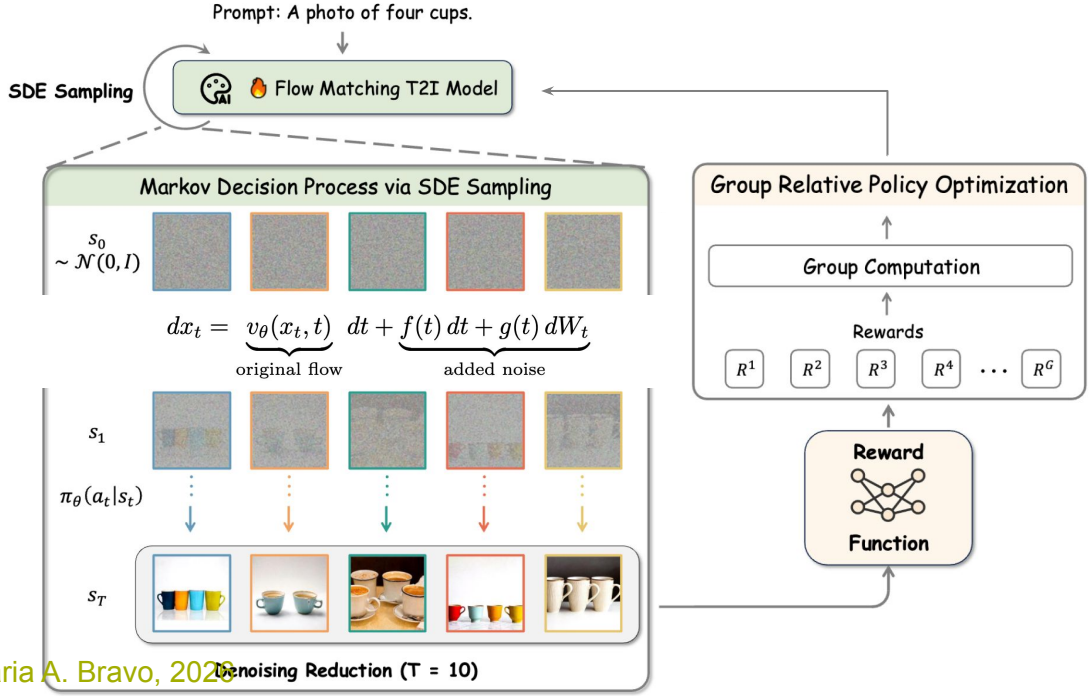
+ Techniques:
- Autoguidance (distillation w/ weak model ↑ sample quality)
- Adversarial finetuning (↑ sample sharpness)

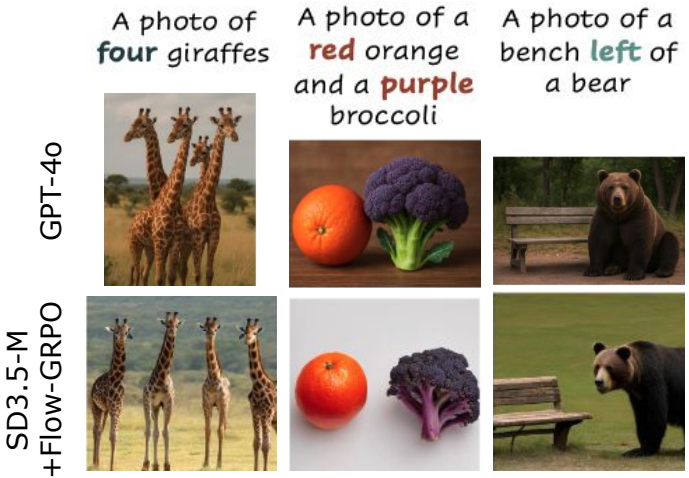


2.3 Flow-GRPO: Training Flow Matching Models via Online RL

Objective: Align with **human preference** → Use **Reward** models and **RL**

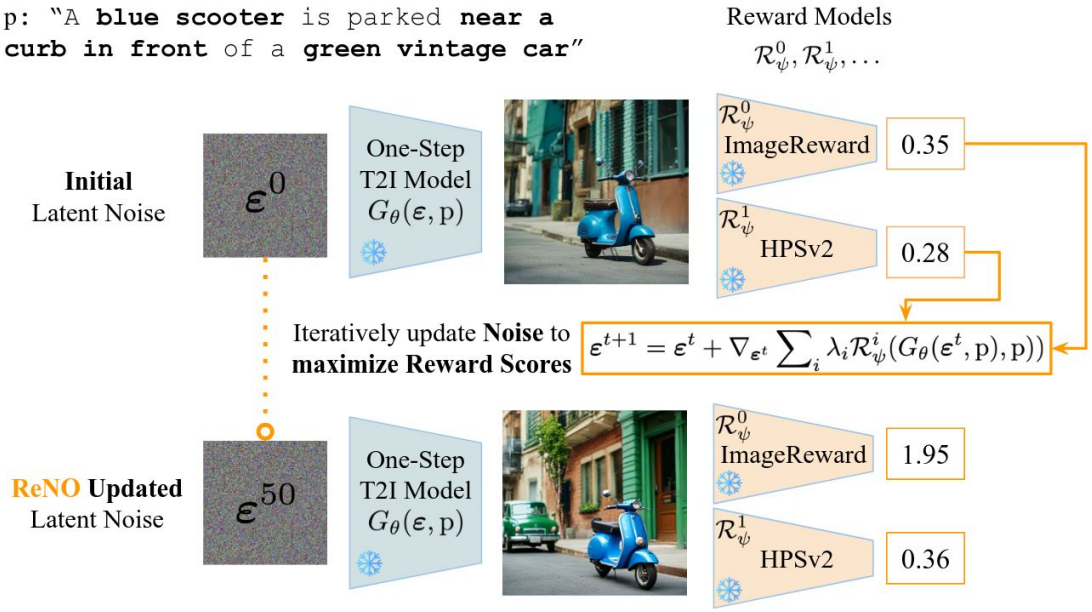
For RL: we need to sample **multiple answers** → Non-deterministic needed
Convert the Flow Matching **ODE to SDE**





SOTA results without compromising quality and diversity
Counting, **Attribute Binding**, **Position**

2.4 ReNO: Enhancing One-step Text-to-Image Models through Reward-based Noise Optimization



Text-to-Image (T2I) models struggle with complex, detailed prompts.

Goal: ↑ Image quality
↑ Prompt alignment
NO retraining

Optimize the initial noise with **reward models**.

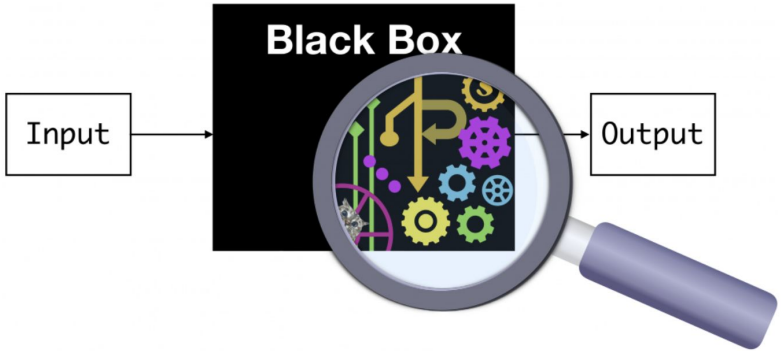
Works with any **one-step T2I model**
→ is differentiable end-to-end, backprop reward signal into noise

3. Explainability and Mechanistic Interpretability

3. Explainability and Mechanistic Interpretability

Can we understand what neural networks are actually computing?

Not just what they output?



Source: Seojin Bang, CMU, 2019

What this brings: **Understanding**

- Superposition → Polysemanticity
- SAEs as a tool → scales, modality
- Interpretability → safety, control

🐍 Python (snake)

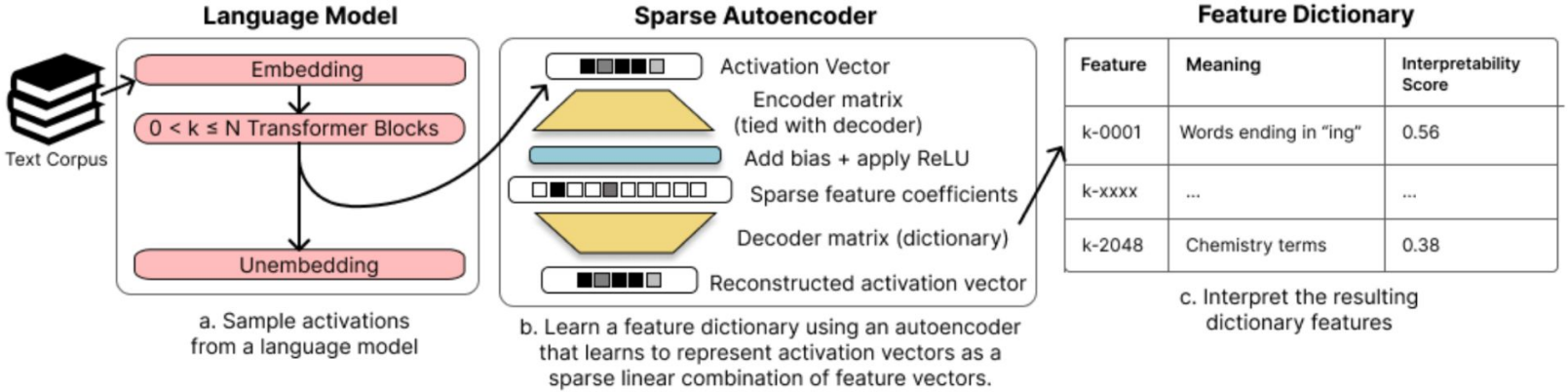
↑

💻 Python (code) ← [NEURON] → 🎉 Rio Carnival

3.1 Sparse Autoencoders Find Highly Interpretable Features in Language Models

If you try to understand the model neuron by neuron, you get **entangled concepts**.

Sparse Autoencoders: decompose neurons activations into clean, **monosemantic** features

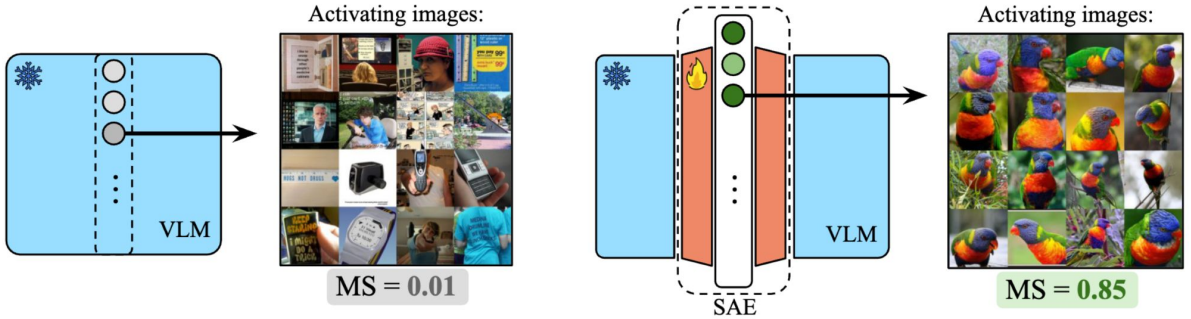


$$\mathbf{c} = \mathrm{ReLU}(M\mathbf{x} + \mathbf{b})$$
$$\hat{\mathbf{x}} = M^T\mathbf{c}$$

$$\mathcal{L}(\mathbf{x}) = \underbrace{||\mathbf{x} - \hat{\mathbf{x}}||_2^2}_{\text{Reconstruction loss}} + \underbrace{\alpha||\mathbf{c}||_1}_{\text{Sparsity loss}}$$

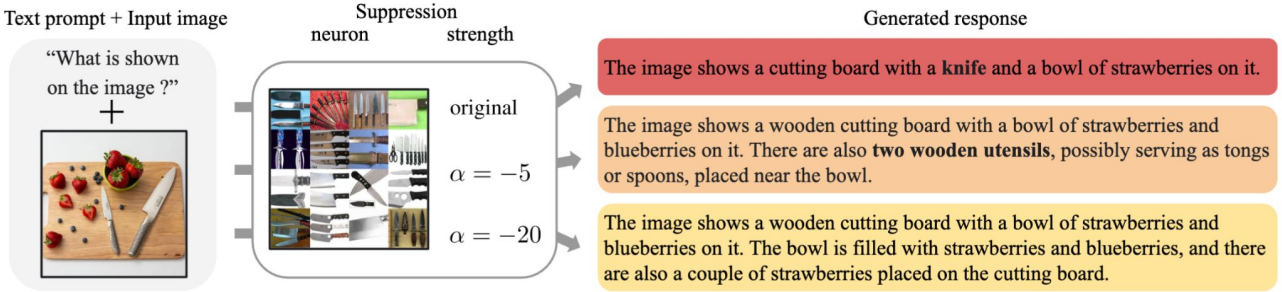
3.2 Sparse Autoencoders Learn Monosemantic Features in Vision-Language Models (VLMs)

SAEs from language to a VLM's (CLIP) vision encoder



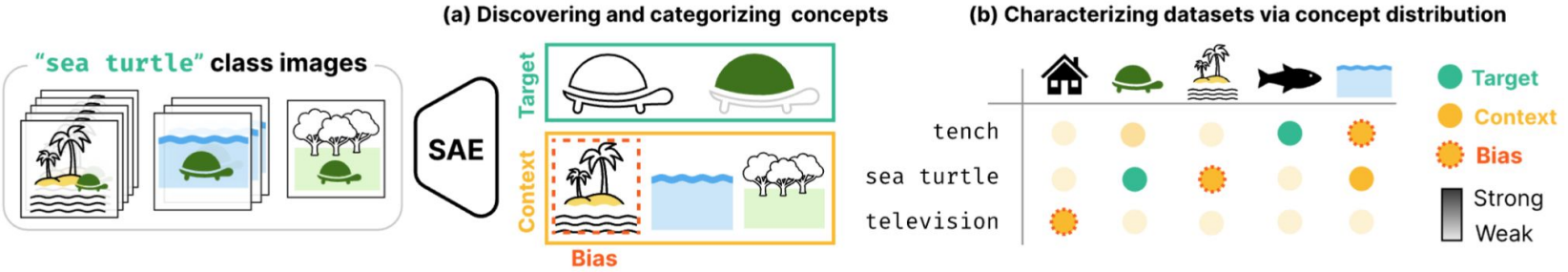
MonoSemanticity score

Steerability → inserting or suppressing concepts



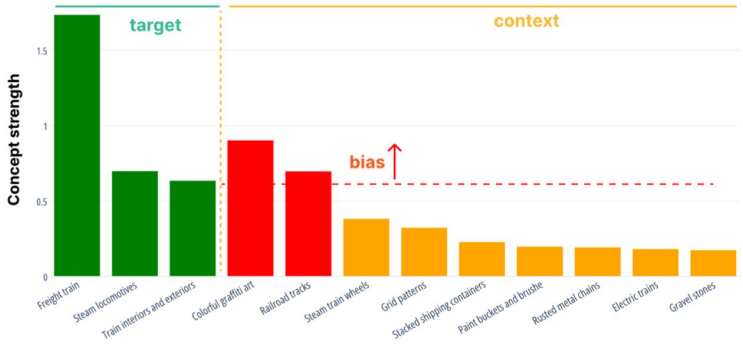
3.3 ConceptScope: Characterizing Dataset Bias via Disentangled Visual Concepts

Uses SAEs to audit **datasets**



→ finds concepts per class:

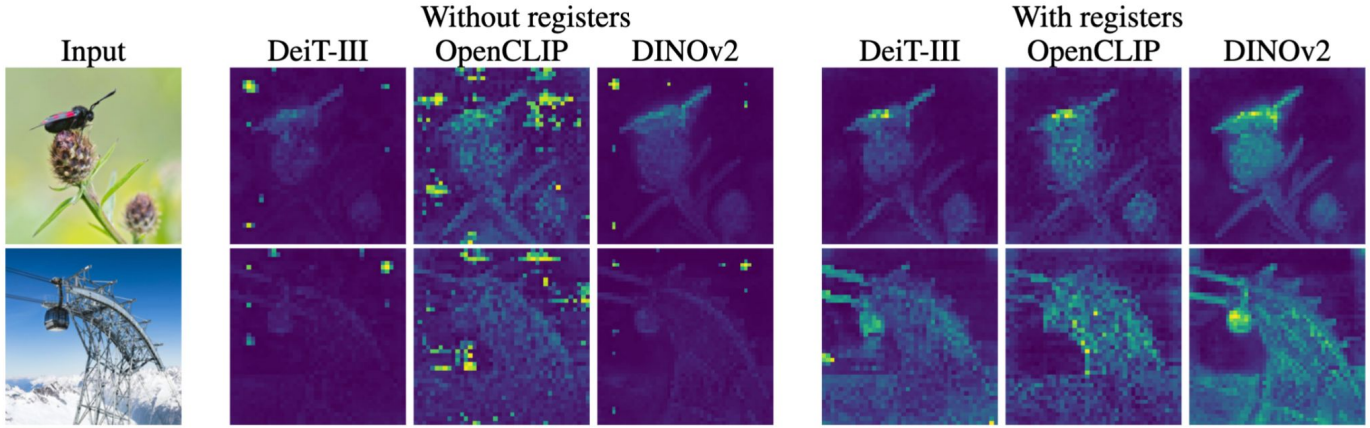
- Target (essential)
- Context (co-occurring but non-essential)
- Bias (context concepts with disproportionate class correlation)



3.4 Vision Transformers Need Registers

ViTs show high-norm uninformative background patches → global memory
Artifact tokens corrupt attention maps and break object discovery algorithms

Solution: Add **register tokens** to the input sequence → absorb the global information and free patch tokens



DINOv2 + registers:

- SOTA on self-supervised dense prediction tasks (depth estimation, segmentation)
- Allow object discovery

4. Foundation Model Adaptation

4. Foundation Model Adaptation

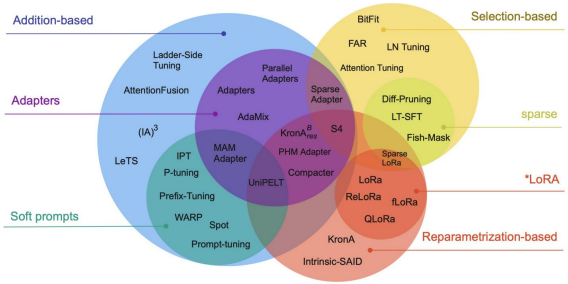
Training a **foundation model** from scratch $$$
→ Take existing: how do you make it yours?

Adaptation: update a pretrained model for your

- Task
- Data
- Hardware

} **without forgetting!**

Parameter-efficient fine-tuning (PEFT)



Source: Lialin et al., The PEFT Taxonomy, 2023

Which parameters do we train?

Memory-efficient training

Adam Optimizer

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{m_2} + \epsilon} \cdot m_1$$

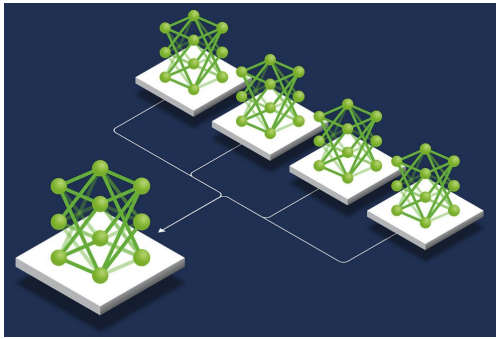
Weights

1st moment

2nd moment

If the gradients are low-rank
→ train in a compact subspace

Model merging

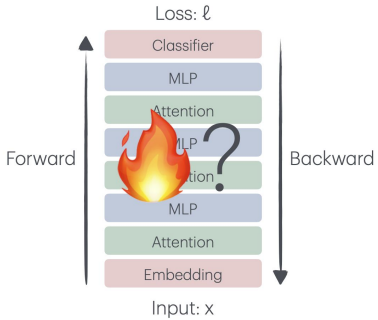


Source: Nvidia, An Intr. to MM for LLMs, 2024

How to merge multiple fine-tuned experts?

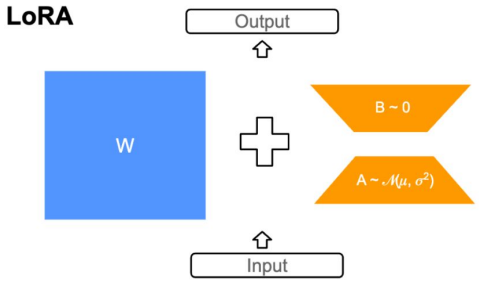
4.1 DeLoRA: Decoupling Angles and Strength in Low-rank Adaptation

Which parameters do we train?



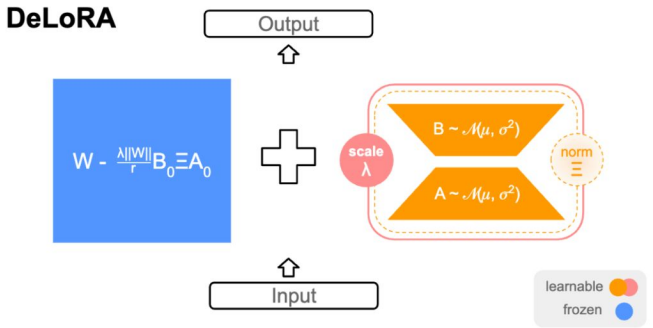
Source: Philipp Krähenbühl, UT Austin, 2025

LoRA: Low-Rank Adaptation



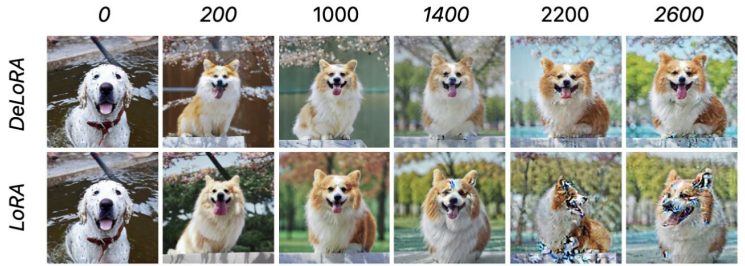
$$Wx + b \rightarrow Wx + b + \frac{\alpha}{R} ABx$$

Sensitive to rank and scaling factor



Adds:

- normalization Ξ: bounds the weights
- scaling factor λ: controls the magnitude of the update



4.2 GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection

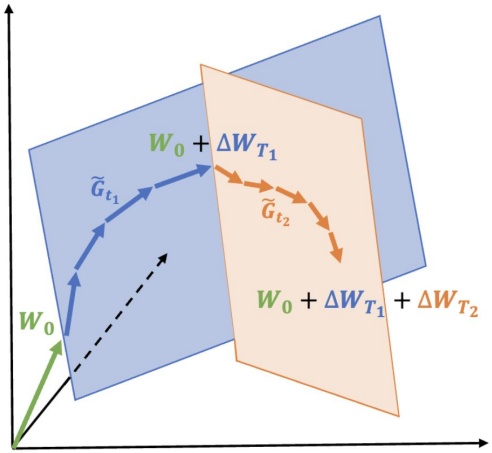
The key idea is to leverage the **slowly changing low-rank structure of the gradient** G of the weight matrix W, rather than attempting to approximate the weight matrix itself as low-rank.

Algorithm 1: GaLore, PyTorch-like

```
for weight in model.parameters():
   grad = weight.grad
   # original space -> compact space
   lor_grad = project(grad)
   # update by Adam, Adafactor, etc.
   lor_update = update(lor_grad)
   # compact space -> original space
   update = project_back(lor_update)
   weight.data += update
```

	60M	130M	350M	1B
Full-Rank	34.06 (0.36G)	25.08 (0.76G)	18.80 (2.06G)	15.56 (7.80G)
GaLore	**34.88** (0.24G)	**25.36** (0.52G)	**18.95** (1.22G)	**15.64** (4.38G)
Low-Rank	78.18 (0.26G)	45.51 (0.54G)	37.41 (1.08G)	142.53 (3.57G)
LoRA	34.99 (0.36G)	33.92 (0.80G)	25.58 (1.76G)	19.21 (6.17G)
ReLoRA	37.04 (0.36G)	29.37 (0.80G)	29.08 (1.76G)	18.33 (6.17G)
r/d_{model}	128 / 256	256 / 768	256 / 1024	512 / 2048
Training Tokens	1.1B	2.2B	6.4B	13.1B

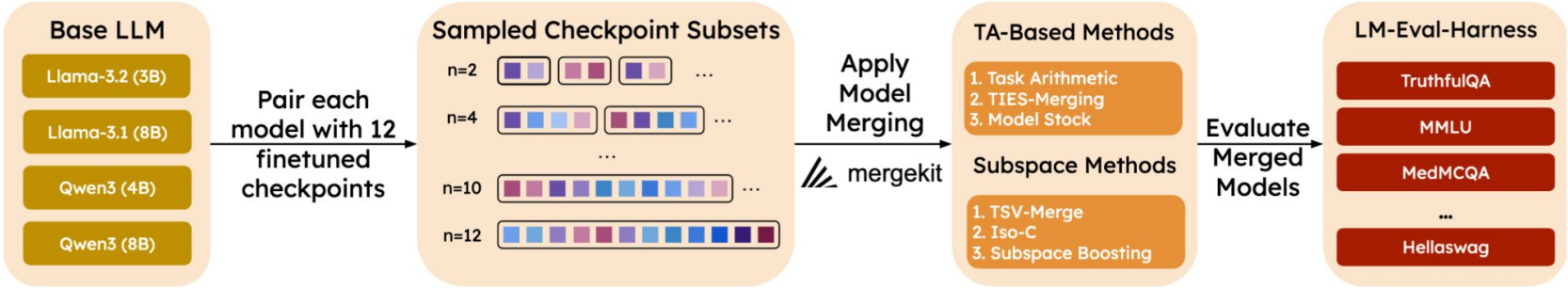
Learning through low-rank subspaces by projecting the gradients using the **Singular Value Decomposition**

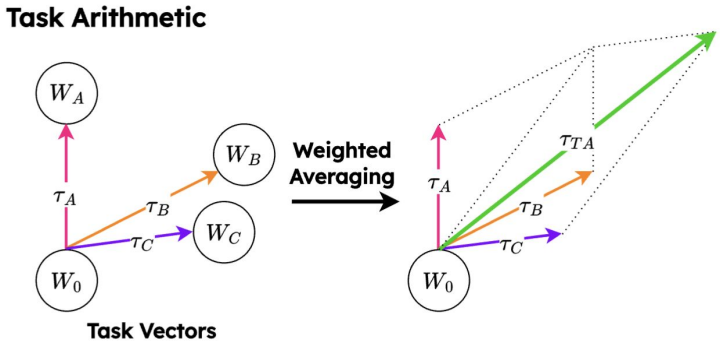


$$W_t = W_0 + \Delta W_{T_1} + \Delta W_{T_2} + \ldots + \Delta W_{T_n},$$

4.3 A Systematic Study of Model Merging Techniques in Large Language Models

The first large-scale systematic evaluation of model merging on modern LLMs (Qwen3, Llama 3)

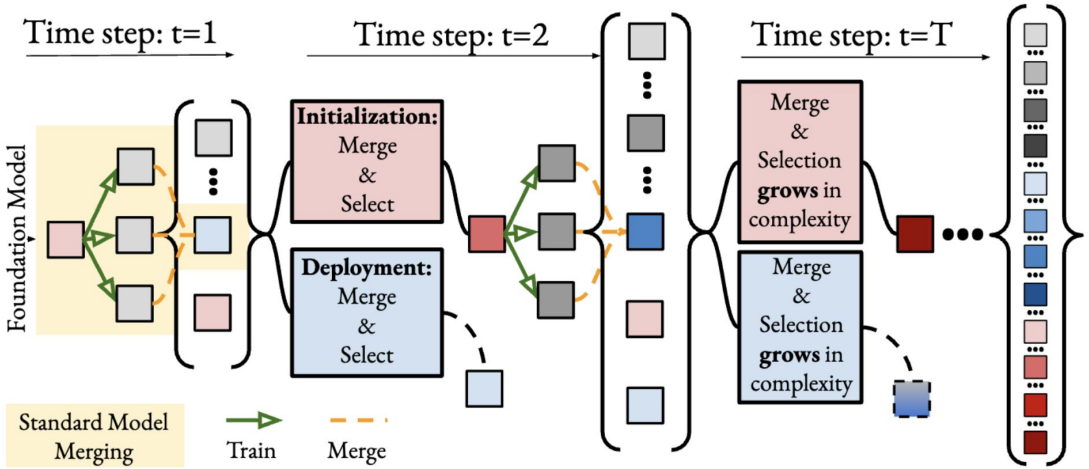




Finding: **Task Arithmetic** (the simplest method) reliably achieves constructive interference (merged model > best individual model)

4.4 How to Merge Your Multimodal Models Over Time?

TIME framework for temporal model merging



1. **Initialization**: how to start training a new expert
   → **accounting for time** is crucial and offline merging techniques do NOT generalize
2. **Merging Technique**: which weight combination method to apply.
   → simple weighted averaging is competitive
3. **Deployment**: what model to serve at each time step
   → **Initialization** and **deployment** are critical and temporal model merging **scales well**

TODO: Paper selection

Please

- Rank all the 4 topic groups
- You will be assigned one main paper (hopefully) from your favourite topic group
- You will be assigned two secondary papers

The deadline is: 8pm today

We'll release the paper assignment on Friday