ETH zürich ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA Qualcomm

ICML International Conference On Machine Learning

Project page: github.com/ForeverBlue816/GRACE

Gated Relational Alignment via Confidence-based Distillation for Efficient VLMs

Integrated Systems Laboratory (ETH Zürich)

Yanlong Chen yanlchen@student-ethz.ch

Amirhossein Habibian

Supervisors

Prof. Dr. Yawei Li, Prof. Dr. Luca Benini

PULP Platform
Open Source Hardware, the way it should be!

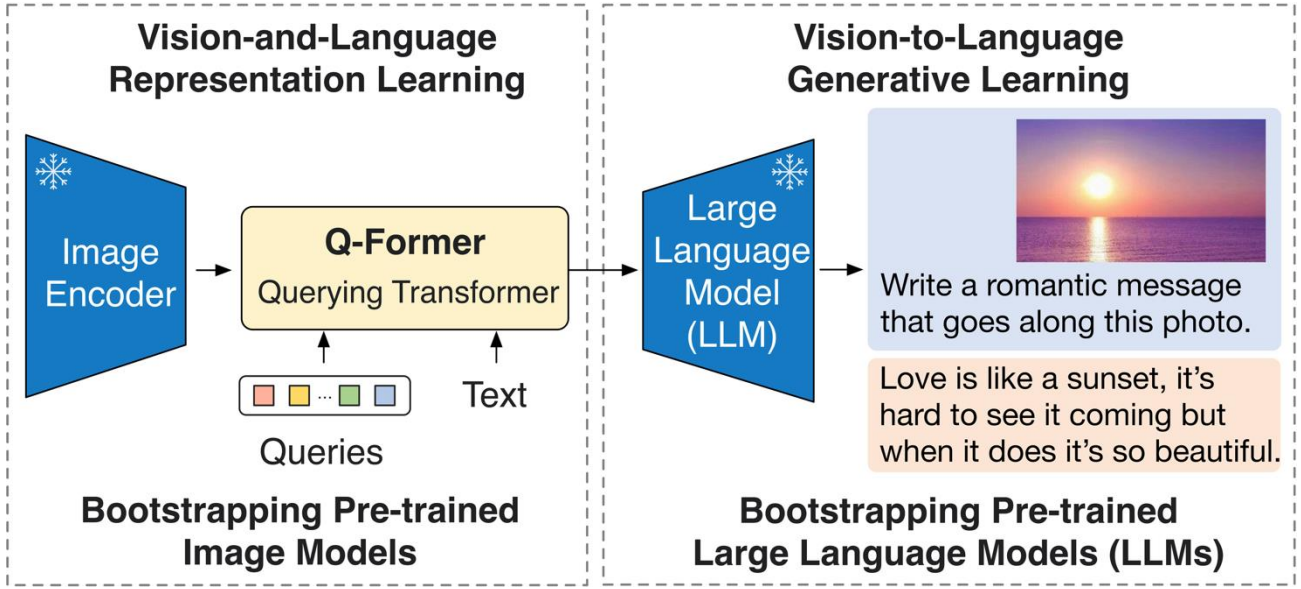
@pulp_platform

pulp-platform.org

youtube.com/pulp_platform

Background & Motivation

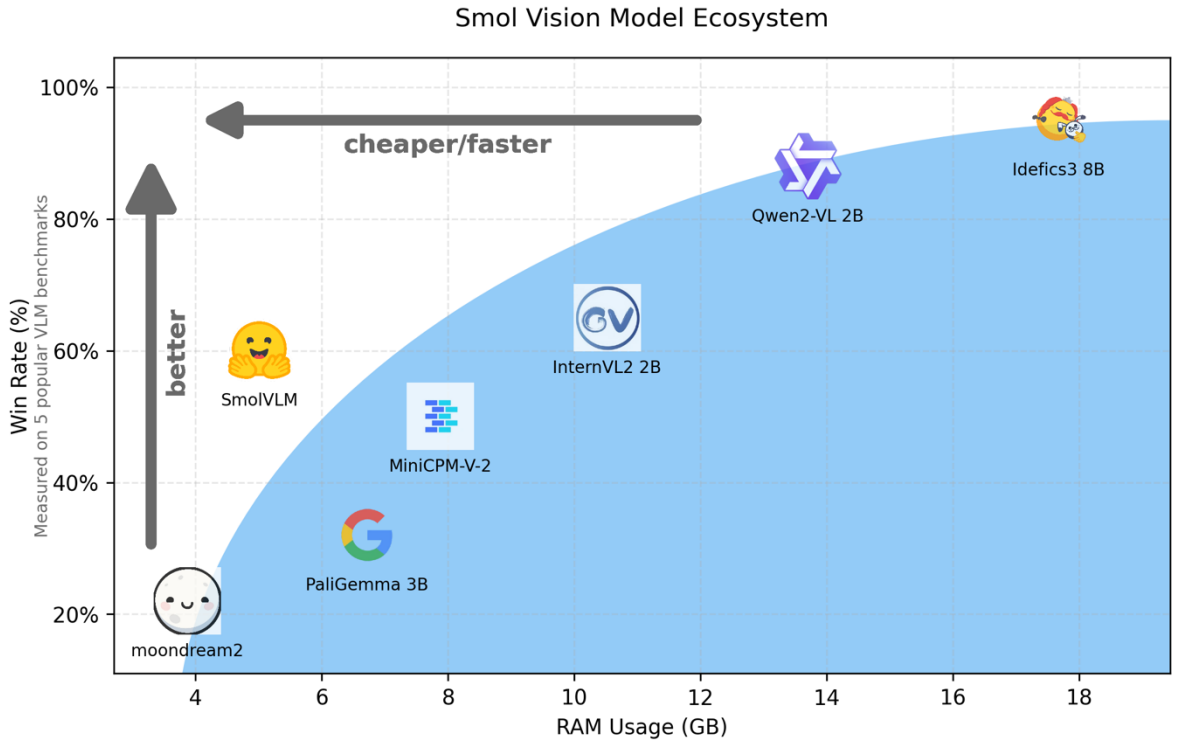
- **The Rise of Vision-Language Models (VLMs)**
  - Multimodal Prowess: VLMs bridge visual perception and linguistic reasoning, enabling tasks like Visual Question Answering (VQA) and Scene Understanding.
  - Embodied AI: Serving as the core for Vision-Language-Action (VLA) robotic systems.



Typical VLMs Architecture

Background & Motivation

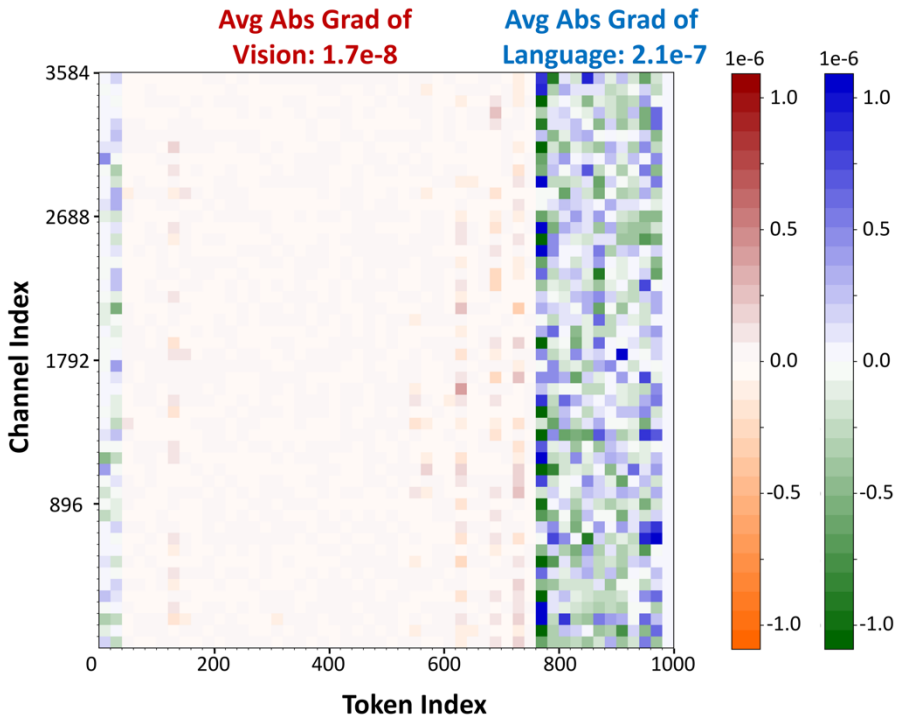
- **The Deployment Bottleneck**
  - Typically require billions of parameters.
  - Consume **massive memory and compute resources**.
  - Extremely challenging to deploy on resource-constrained or edge devices.



Model Size vs GPU Memory Trade-off

Challenges of VLM Quantization

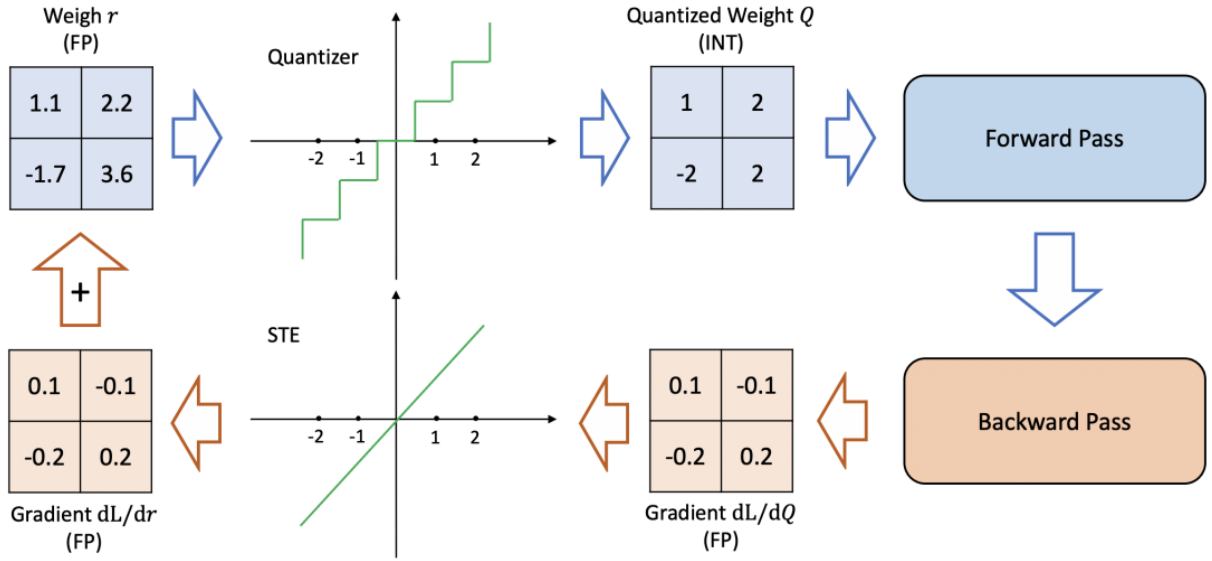
- **Post-Training Quantization (PTQ)**
  - PTQ reduces model precision after training by quantizing weights and activations using small calibration data.
  - Often leads to **significant accuracy degradation** in VLMs due to complex cross-modal interactions and heterogeneous feature distributions.



The gradients of loss function with resepct to the token features in LLaVA-1.5 7B VLM

Challenges of VLM Quantization

- **QAT (Quantization-Aware Training)**
  - QAT simulates low-precision quantization errors during the training phase, allowing the model to adapt to and compensate for information loss.
  - Spends huge resources merely to **recover the original accuracy**, bounded by the full-precision baseline.



QAT pipeline

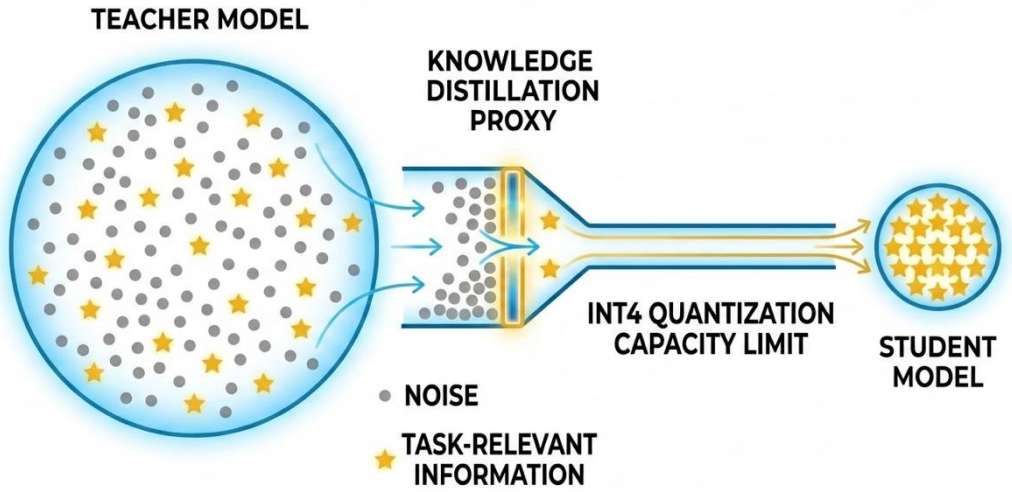
Is it possible to design a quantization framework that enables a low-bit (INT4) VLM to actually surpass its full-precision baseline?

Proposed Framework

Gated Relational Alignment via Confidence-based Distillation for Efficient VLMs.

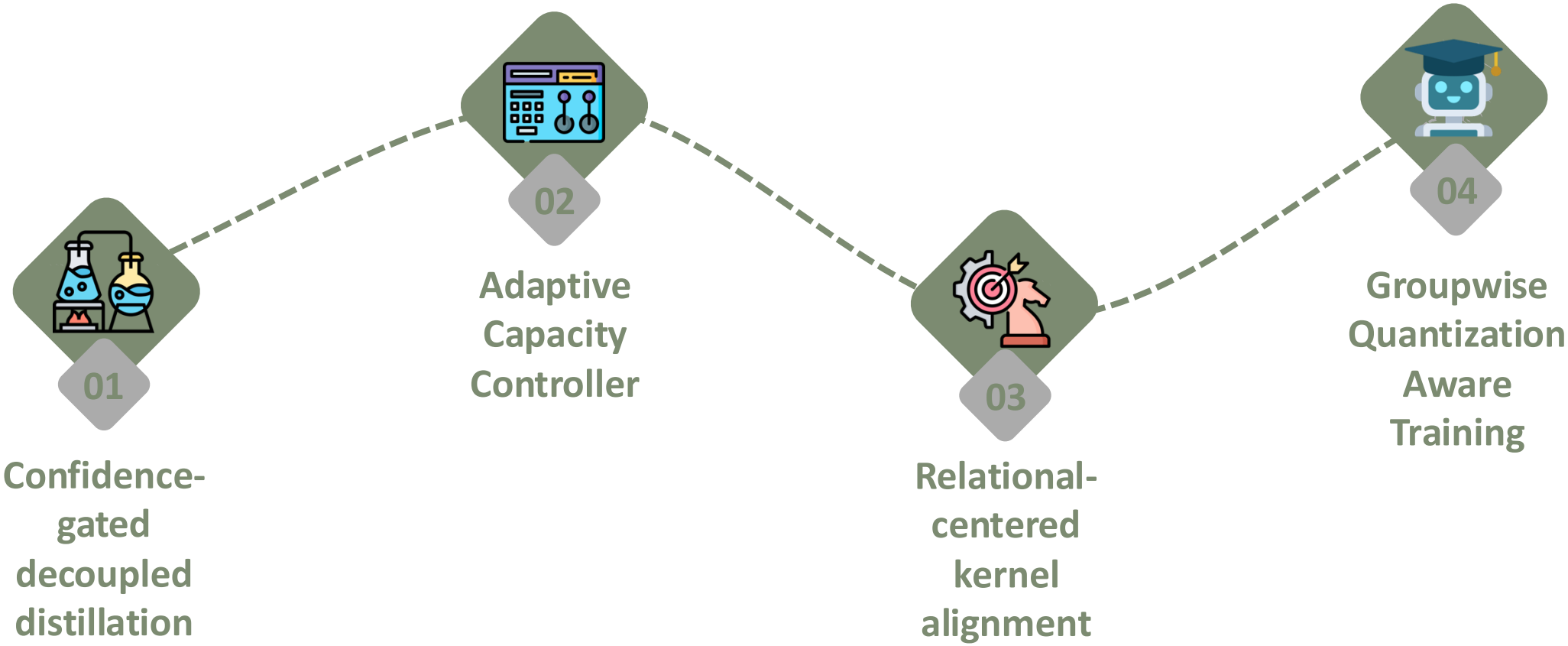
Core Philosophy: Unifying **Knowledge Distillation** and **QAT** under the **Information Bottleneck** principle.

- Quantization limits the student's information capacity.
- Distillation serves as a proxy to guide what task-relevant information to preserve within this strict budget.

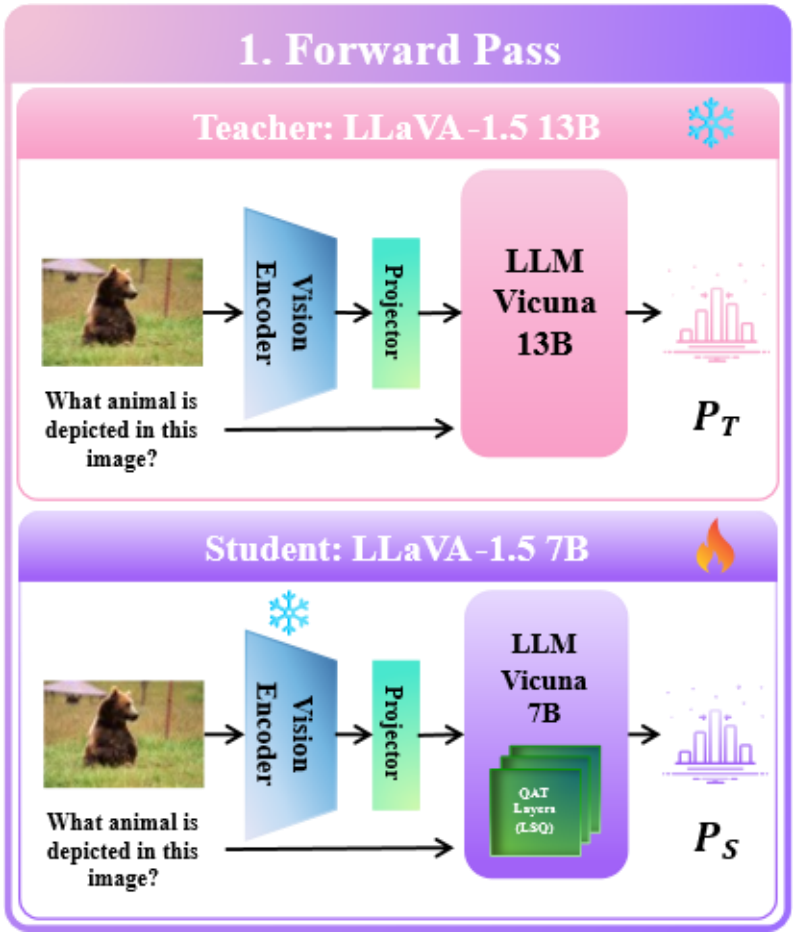


Proposed Framework

Gated Relational Alignment via Confidence-based Distillation for Efficient VLMs.

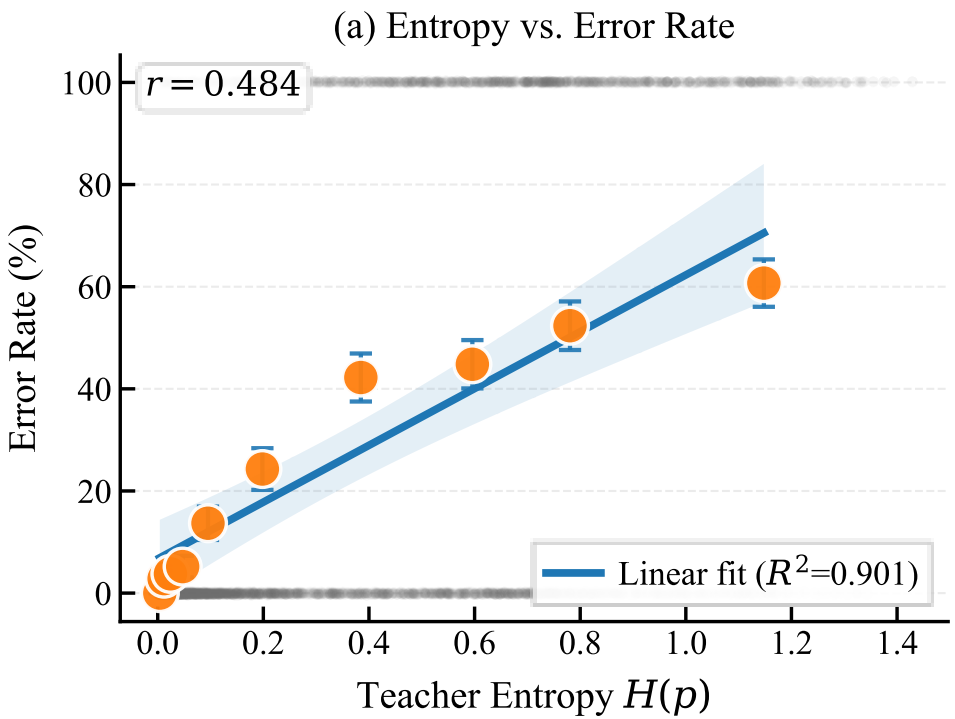


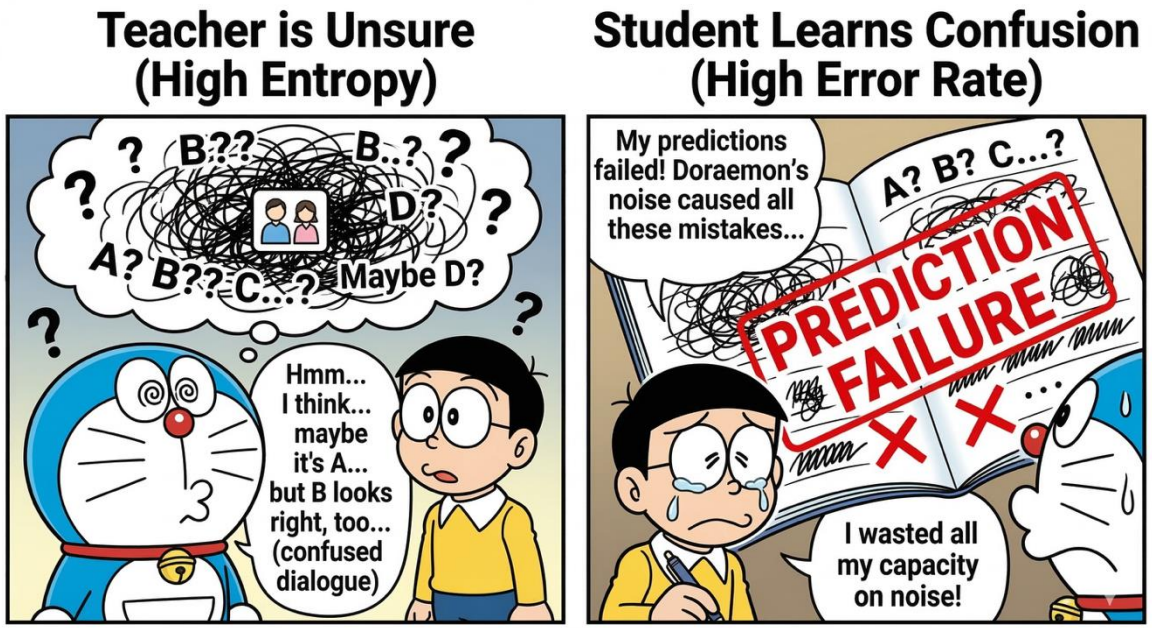
Confidence-Gated Decoupled Distillation



$$Uncertainty:\ H(P_T) = -\sum_{v} P_T(v) \log P_T(v)$$

Confidence-Gated Decoupled Distillation



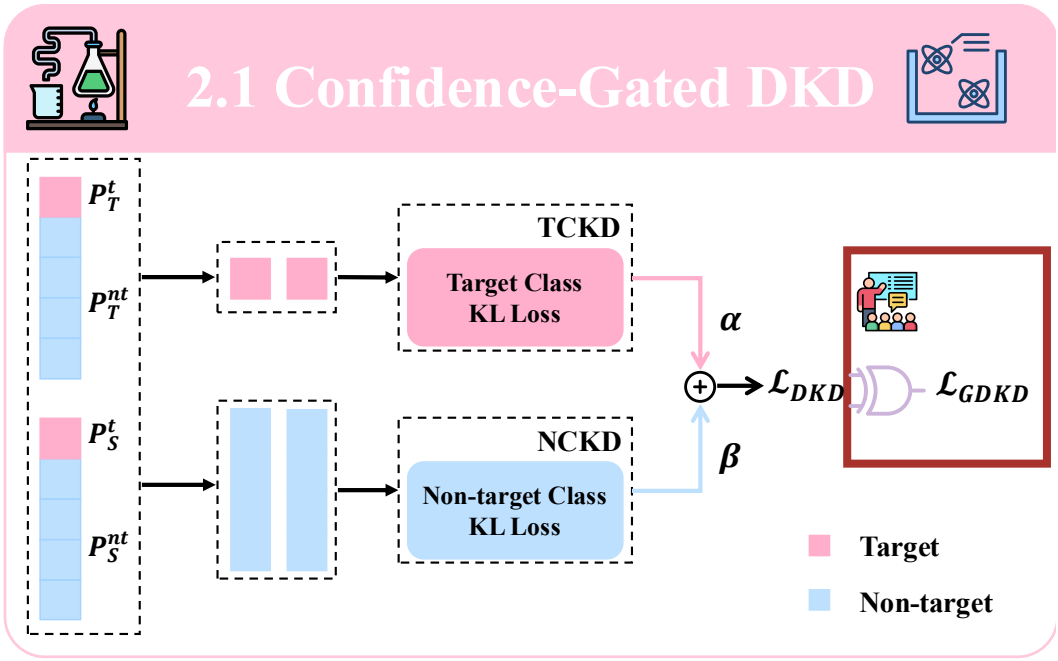


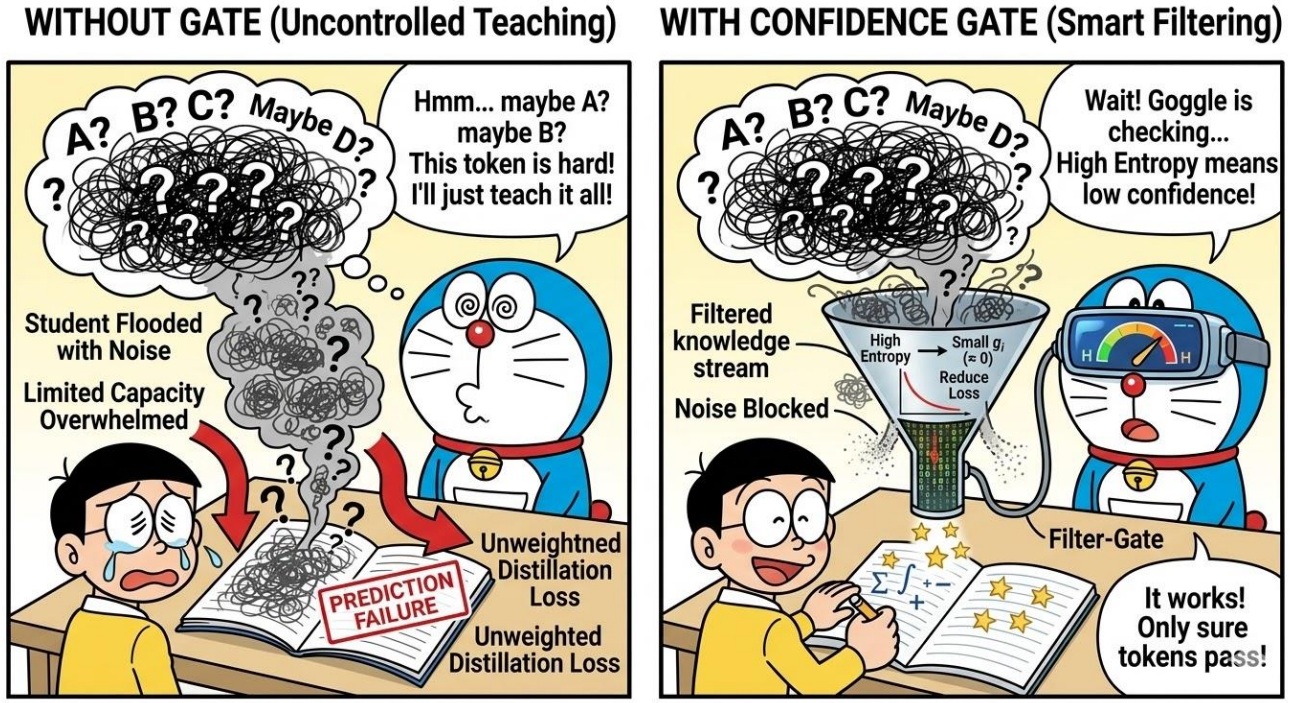
Correlation between teacher entropy and error rate

Teacher **entropy** exhibits a strong correlation with **prediction errors**.

Teachers can produce **uncertain or noisy predictions**, forcing the low-capacity student to learn irrelevant noise.

Confidence-Gated Decoupled Distillation

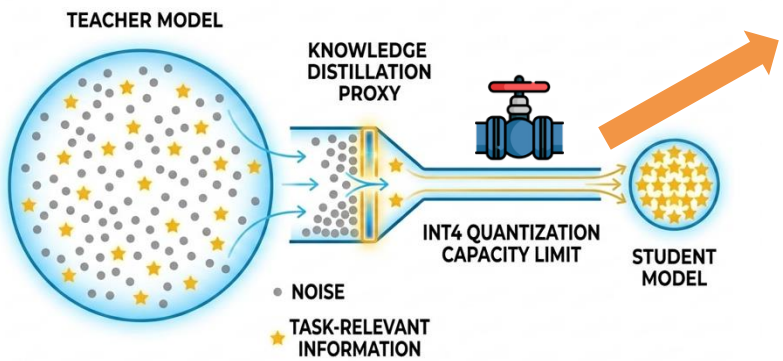




$\mathcal{L}_{DKD}$ captures the correct answer and the relationships among the wrong answers.

The student only learns from the tokens that the teacher is sure about.

Adaptive Controller for Capacity Balancing



Since our student has a capacity limit, we cannot just force-feed it knowledge all the time.

2.3 Adaptive Information Bottleneck

Monitor $\hat{\mathcal{L}}_{\text{GDKD}}$

- **If $\hat{\mathcal{L}}_{\text{GDKD}} < \tau$: decrease β**
- **If $\hat{\mathcal{L}}_{\text{GDKD}} > \tau$: increase β**

Dynamically adjusts distillation weight β

The adaptive controller continuously monitors the distillation loss against a predefined tolerance threshold τ.

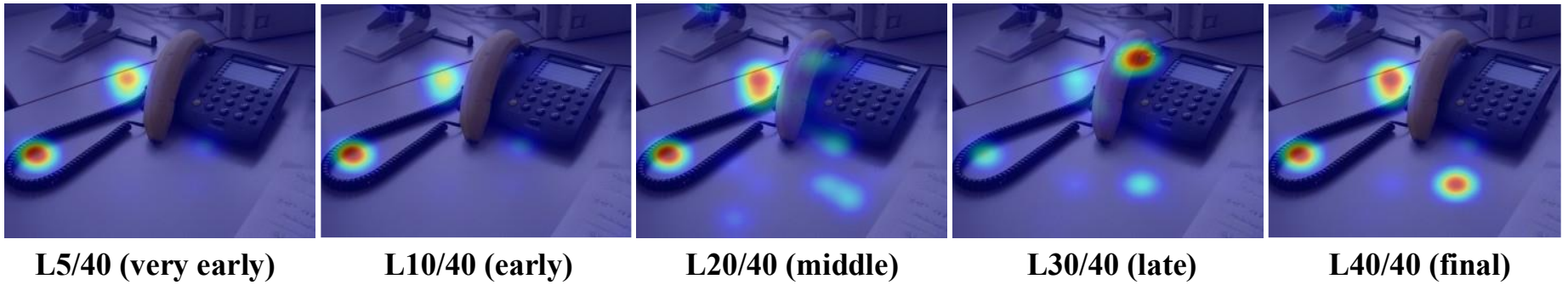
This automatically balances the knowledge transfer without overwhelming the student.

Final distillation loss is $\boldsymbol{\beta} * \mathcal{L}_{GDKD}$.

Relational-Centered Kernel Alignment

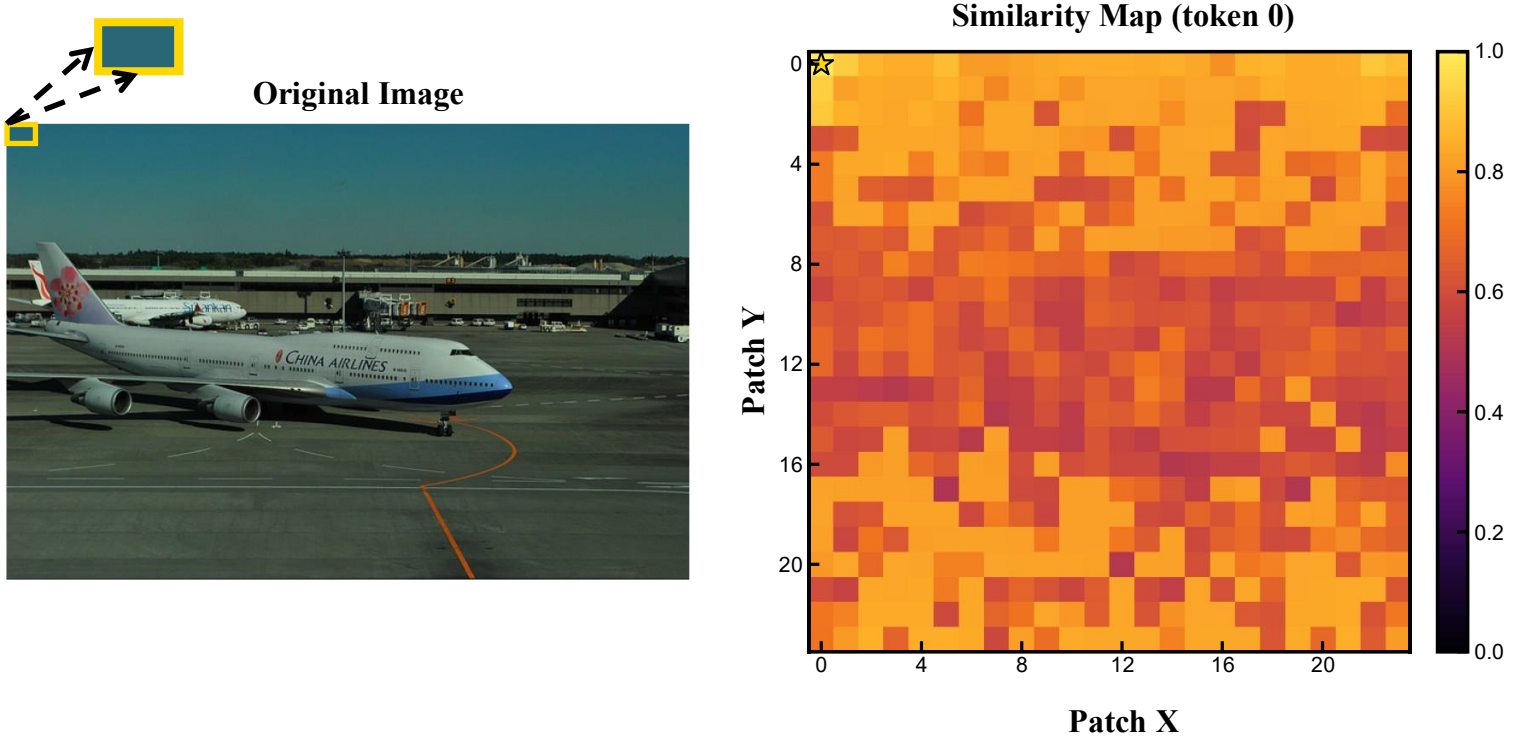
While logit-based distillation transfers output predictions, larger teacher models may possess superior **visual reasoning capabilities** not captured at the output level.

What object is being used as the telephone receiver?



Multi-layer attention visualization of LLaVA-1.5 13B (top) and 7B (bottom).

Relational-Centered Kernel Alignment

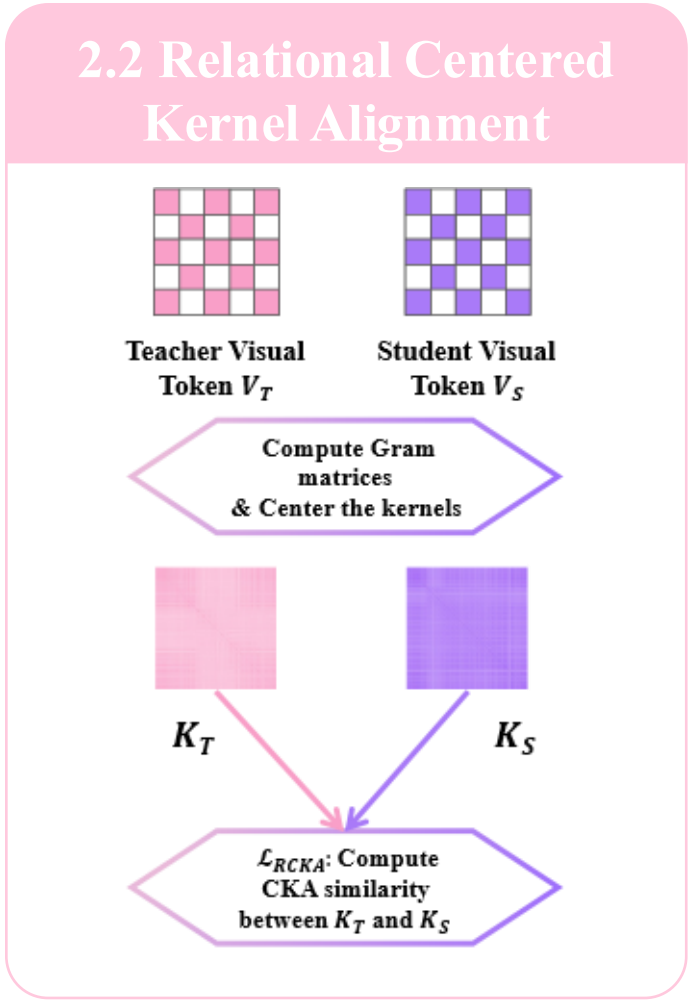


How to **transfer this relational knowledge** to the student?

Visualization of pairwise visual token similarity from LLaVA-1.5 13B.

The teacher model does not just process **isolated patches**; it inherently captures **semantic relationships and geometric structures**.

Relational-Centered Kernel Alignment



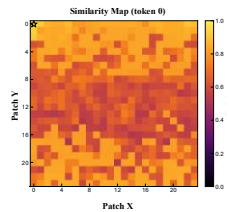
Instead of aligning individual visual tokens, we construct Gram matrices to capture the internal geometric structure between tokens.

- Visual token features from the Teacher $V_T \in \mathbb{R}^{N \times D_T}$ and the Student $V_S \in \mathbb{R}^{N \times D_S}$

The heatmap is just **one row** of this global structure!

- Constructing the Gram Matrices

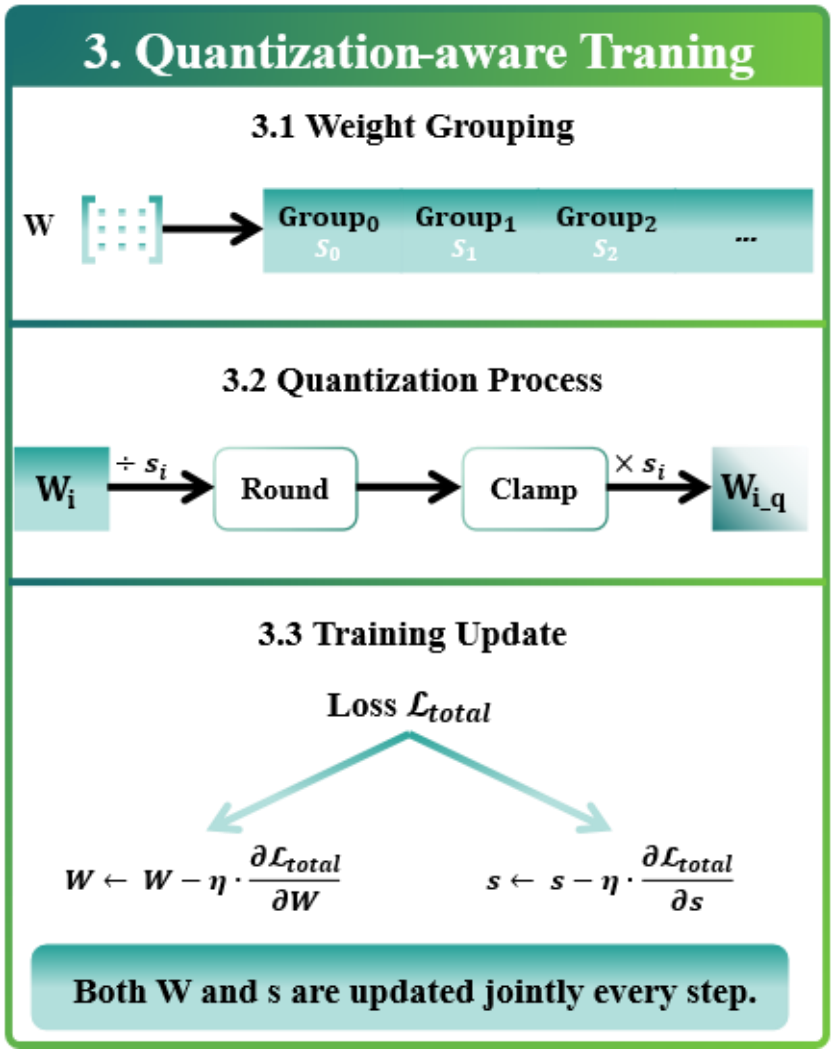
$$K_T = V_T V_T^\top \in \mathbb{R}^{N \times N}, \quad K_S = V_S V_S^\top \in \mathbb{R}^{N \times N}$$



- $\mathcal{L}_{\text{RCKA}} = 1 - \dfrac{\text{Tr}(K_T' K_S')}{||K_T'||_F ||K_S'||_F}$ → Similarity

Force the student to learn **how objects related to each other**.

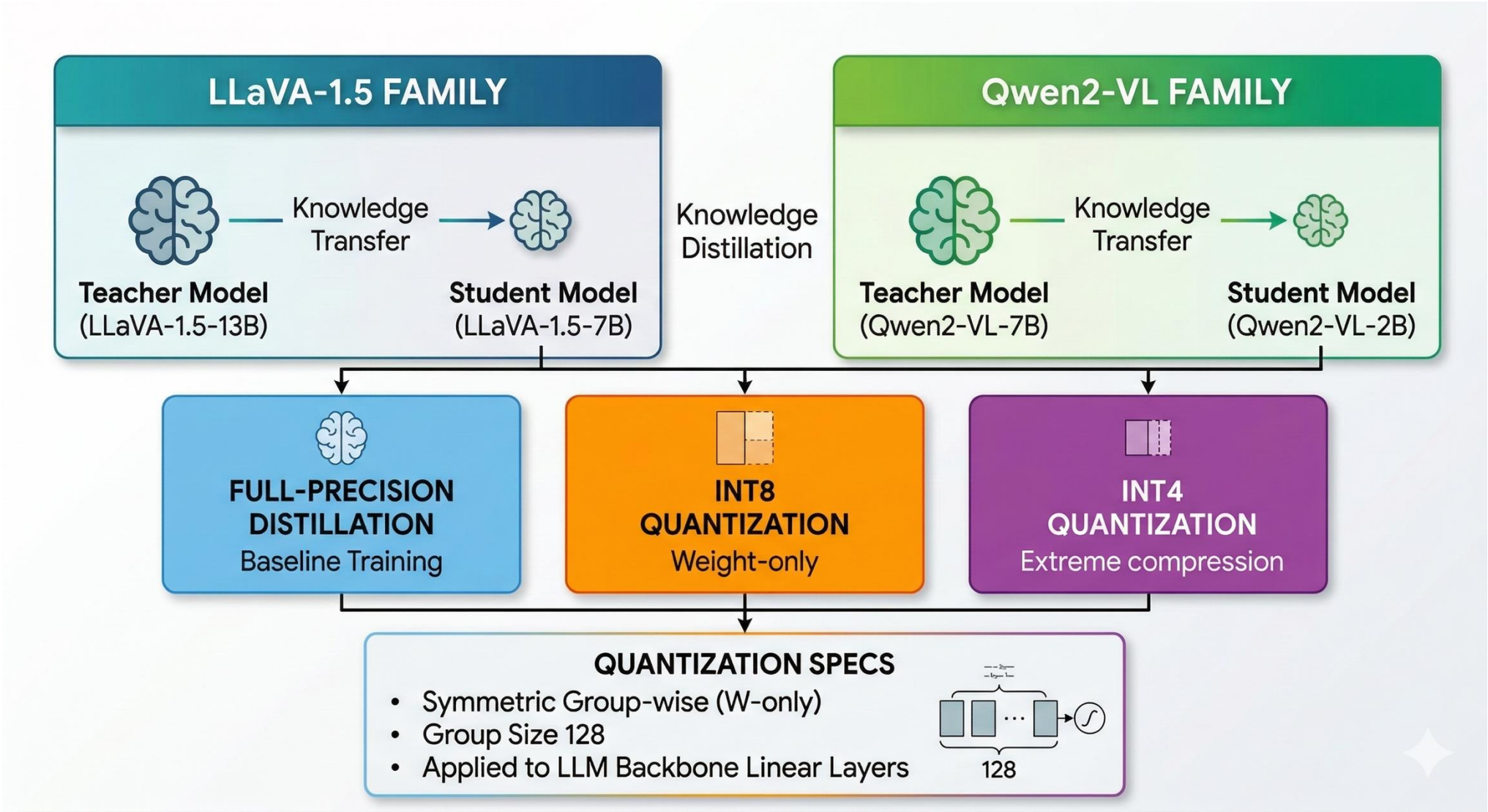
Groupwise Quantization Aware Training



$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \beta * \mathcal{L}_{GDKD} + \omega * \mathcal{L}_{RCKA}$$

- Group Size: 128
- Each group maintains its own independent scaling factor s_i to effectively isolate weight outliers.
- Both the weights and the scaling factors are fully **learnable**, which are jointly optimized at every step to minimize the total loss $\mathcal{L}_{total}$.

Experimental Setup



Experimental Setup

Training Dataset:

- **Dataset Used:** ShareGPT4V dataset
- **Purpose:** 1.2M high-quality image-text pairs generated by GPT-4V

Hardware & Computational Cost:

- **GPU Infrastructure:** Trained on 8 NVIDIA H100 GPUs
- **Training Time:** 12 hours of training for LLaVA-1.5 and 8 hours for Qwen2-VL

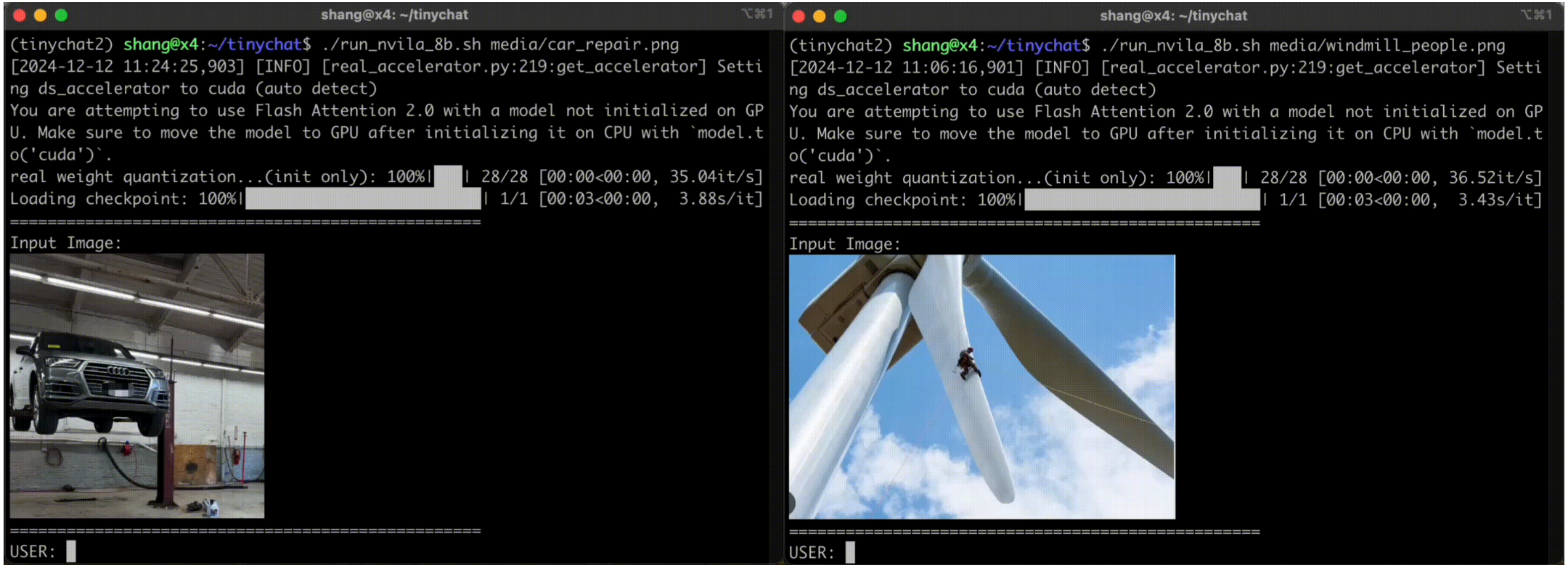
ShareGPT4V Data Collection

Experimental Results

Bitwidth	Method	MMB	MMStar	MMMU	Hallusion	AI2D	OCR	SEED	SQA	Avg.
BF16	Qwen2-VL-7B (Teacher)	80.7	60.7	54.1	50.6	83.0	84.5	76.9	83.3	71.7
BF16	Qwen2-VL-2B (Baseline)	72.6	50.4	45.4	42.9	73.1	81.0	72.7	73.7	64.0
BF16	**GRACE (Ours)**	77.9	55.5	52.0	48.3	80.0	83.1	75.7	81.0	**69.2** ↑5.2%
8-bit	**GRACE (Ours)**	77.4	55.2	51.8	47.9	79.7	82.9	75.5	80.8	**68.9** ↑4.9%
4-bit	RTN	<u>70.98</u>	45.07	37.22	<u>42.74</u>	<u>71.15</u>	78.8	72.45	70.67	61.1 ↓3.9%
4-bit	GPTQ [ICLR'23]	70.51	46.33	36.22	39.62	70.53	79.5	72.62	72.10	60.9 ↓3.1%
4-bit	AWQ [MLSys'24]	68.89	44.80	37.33	39.55	70.08	78.9	71.83	<u>72.15</u>	60.4 ↓3.6%
4-bit	MBQ [CVPR'25]	70.55	44.53	38.22	39.89	70.21	<u>80.9</u>	71.86	71.72	61.0 ↓3.0%
4-bit	SPEED-Q [arXiv'25]	69.85	<u>50.87</u>	<u>42.0</u>	41.71	70.92	76.5	<u>74.40</u>	72.01	<u>62.3</u> ↓1.7%
4-bit	**GRACE (Ours)**	**76.9**	**54.3**	**51.1**	**46.5**	**78.6**	**81.4**	**75.6**	**79.1**	**68.0** ↑4.0%

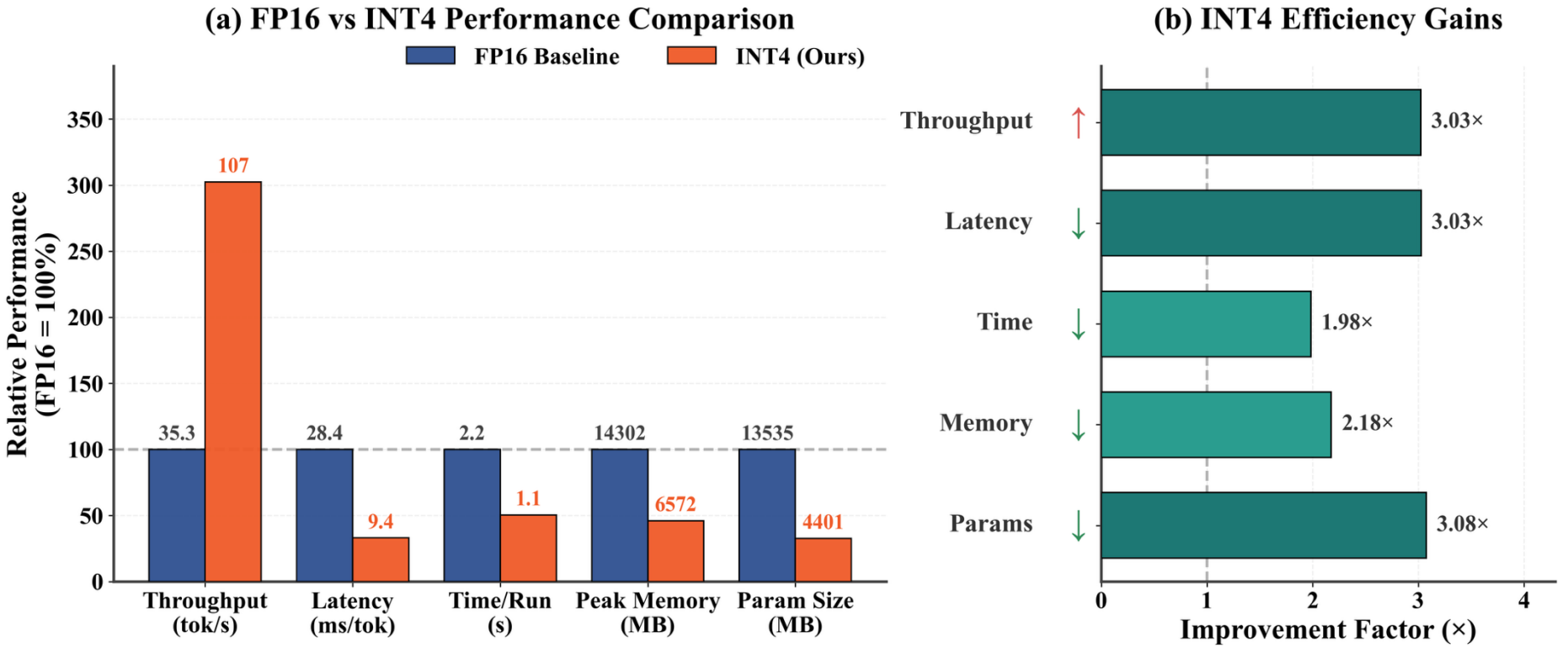
Comparison with quantization methods on Qwen2-VL-2B. Best results among 4-bit models are **bolded**, second best are <u>underlined</u>.

Hardware Efficiency



For INT4 inference, we employ TinyChat, which provides optimized INT4 CUDA kernels . These kernels minimize memory bandwidth overhead and maximize arithmetic intensity.

Hardware Efficiency



Deployment efficiency comparison between LLaVA-1.5 7B FP16 and GRACE 7B INT4 on one RTX4090 GPU.

Qualitative Analysis

Please describe the scene in the picture in detail.

LLaVA-1.5 7B (Baseline):The image captures a nighttime cityscape with a tall building in the center, surrounded by a city skyline. The building is lit up with a dazzling display of fireworks, creating a spectacular view. The fireworks are scattered throughout the scene, with some closer to the ground and others higher up in the sky.

GRACE 7B: The image captures a breathtaking view of a city skyline at night, bathed in the glow of a spectacular fireworks display. The fireworks, in hues of **yellow, red, and purple**, are exploding in the sky, creating a mesmerizing spectacle. The tallest building in the scene is the **Taipei 101 in Taiwan**, standing majestically in the center of the frame. The cityscape is dotted with other buildings, their lights twinkling like stars against the night sky. The **perspective of the image is from a high vantage point**, looking down upon the city, giving a sense of scale and grandeur to the scene.

Conclusion & Future Work

- ✓ Identified the limitations of PTQ for VLMs and pioneered a **unified framework**.
- ✓ Successfully applied the Information Bottleneck principle to **maximize task-relevant knowledge extraction** under **strict quantization capacity constraints**.
- ✓ Demonstrated that 4-bit VLMs can not only match but often **exceed standard FP16** baselines on complex visual reasoning tasks.

Future Work

- ☐ Extend our framework to **activation** quantization.
- ☐ Explore its application to **other multimodal architectures**. (Video/Multi-image/MoE)

The paper is current under review. The code will be released on GitHub once the submission is officially accepted.

https://arxiv.org/abs/2601.22709 (Arxiv link to this paper)

PULP Platform

Open Source Hardware, the way it should be!

Yanlong Chen yanlchen@student-ethz.ch

Institut für Integrierte Systeme – ETH Zürich
Gloriastrasse 35
Zürich, Switzerland

DEI – Universitá di Bologna
Viale del Risorgimento 2
Bologna, Italy

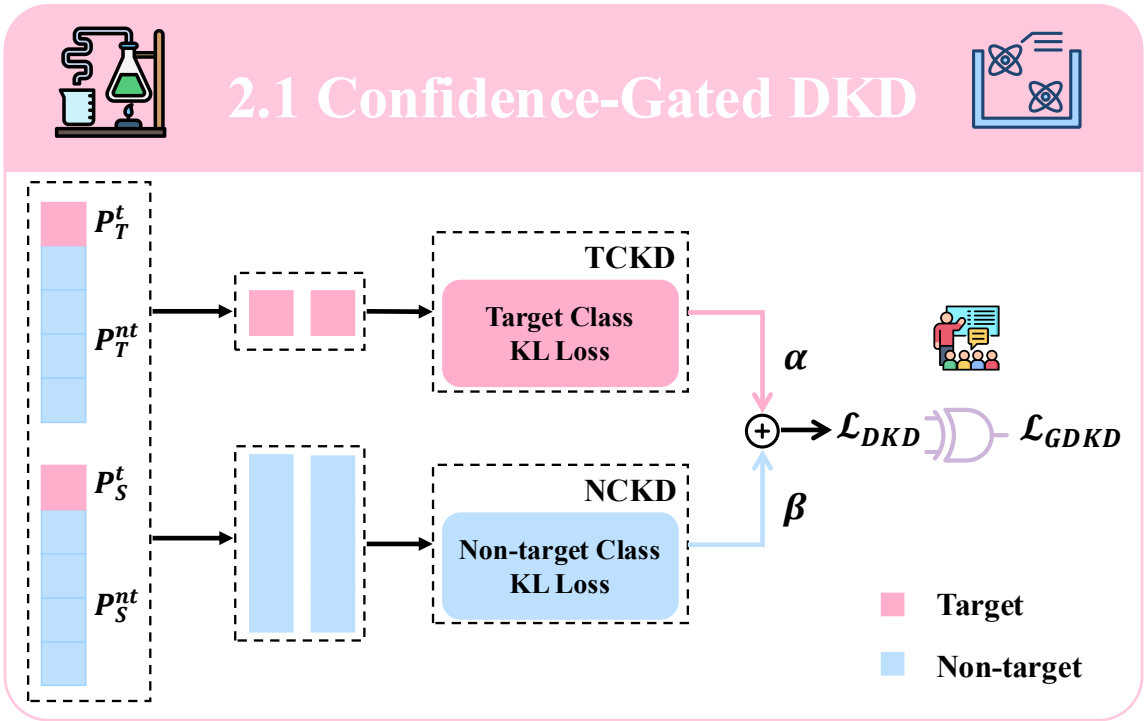
@pulp_platform

pulp-platform.org

youtube.com/pulp_platform

Confidence-Gated Decoupled Distillation

Teachers can produce **uncertain or noisy predictions**, forcing the low-capacity student to learn irrelevant noise.



1. Splits the distillation loss into Target Class Knowledge (TCKD) and Non-Target Class Knowledge (NCKD), allowing independent transfer.

2. Adaptively modulates the distillation loss based on teacher certainty, gating out signals with high entropy.

$$\mathcal{L}_{\text{GDKD}} = \frac{1}{N}\sum_{i=1}^{N} g_i \cdot \mathcal{L}_{\text{DKD}}^{(i)} \qquad g_i = \exp(-\tilde{h}_i)$$

INT4 Kernel

- The main bottleneck for VLM generation is not compute, but **memory bandwidth**. During decoding, the GPU spends most of its time **fetching weights** rather than doing math.

- TinyChat solves this by using W4A16 quantization. Because the weights are stored in INT4, we reduce the memory traffic by exactly 4 times. TinyChat's optimized CUDA kernels **fetch these compressed weights super fast**, do an **on-the-fly dequantization** back to FP16 in the registers, and then perform the matrix multiplication.