Quant Experts

Token-aware Adaptive Error Reconstruction with Mixture of Experts for Large Vision-Language Models Quantization

Chenwei Jia, Baoting Li, Xuchong Zhang, Mingzhuo Wei, Bochen Lin, Hongbin Sun

State Key Laboratory of Human-Machine Hybrid Augmented Intelligence

Institute of Artificial Intelligence and Robotics

Xi'an Jiaotong University

西安交通大學

XI'AN JIAOTONG UNIVERSITY

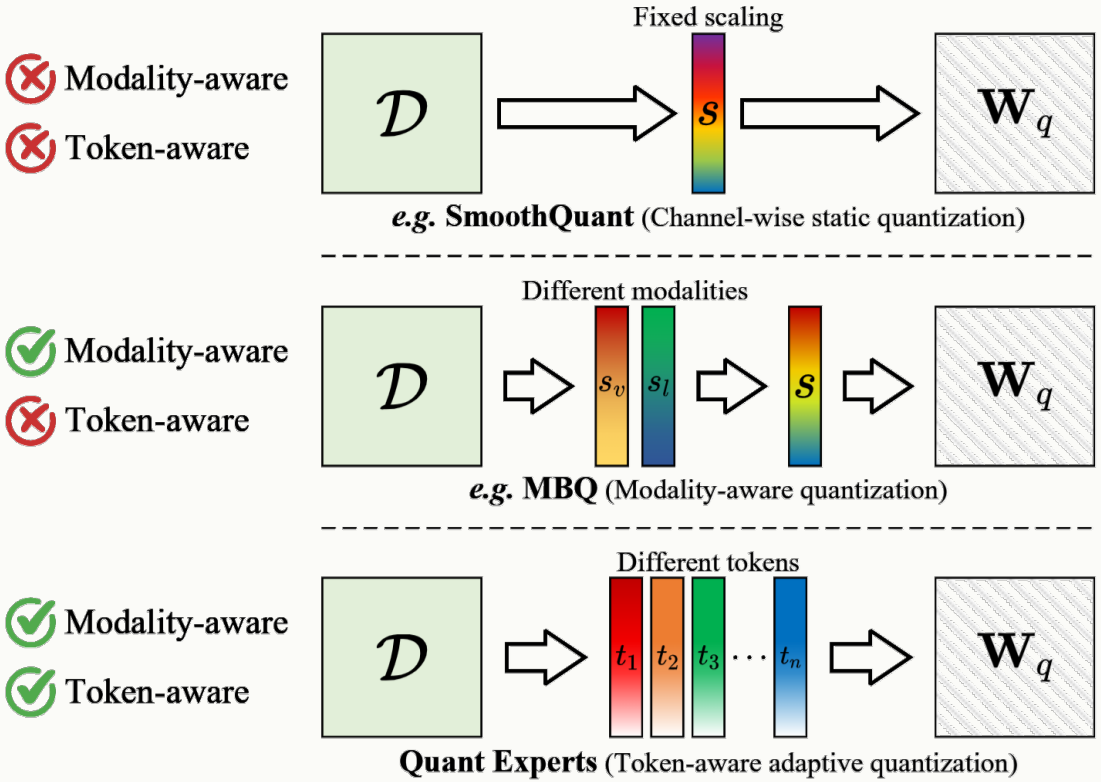
CVPR

JUNE 3-7, 2026

DENVER

COLORADO

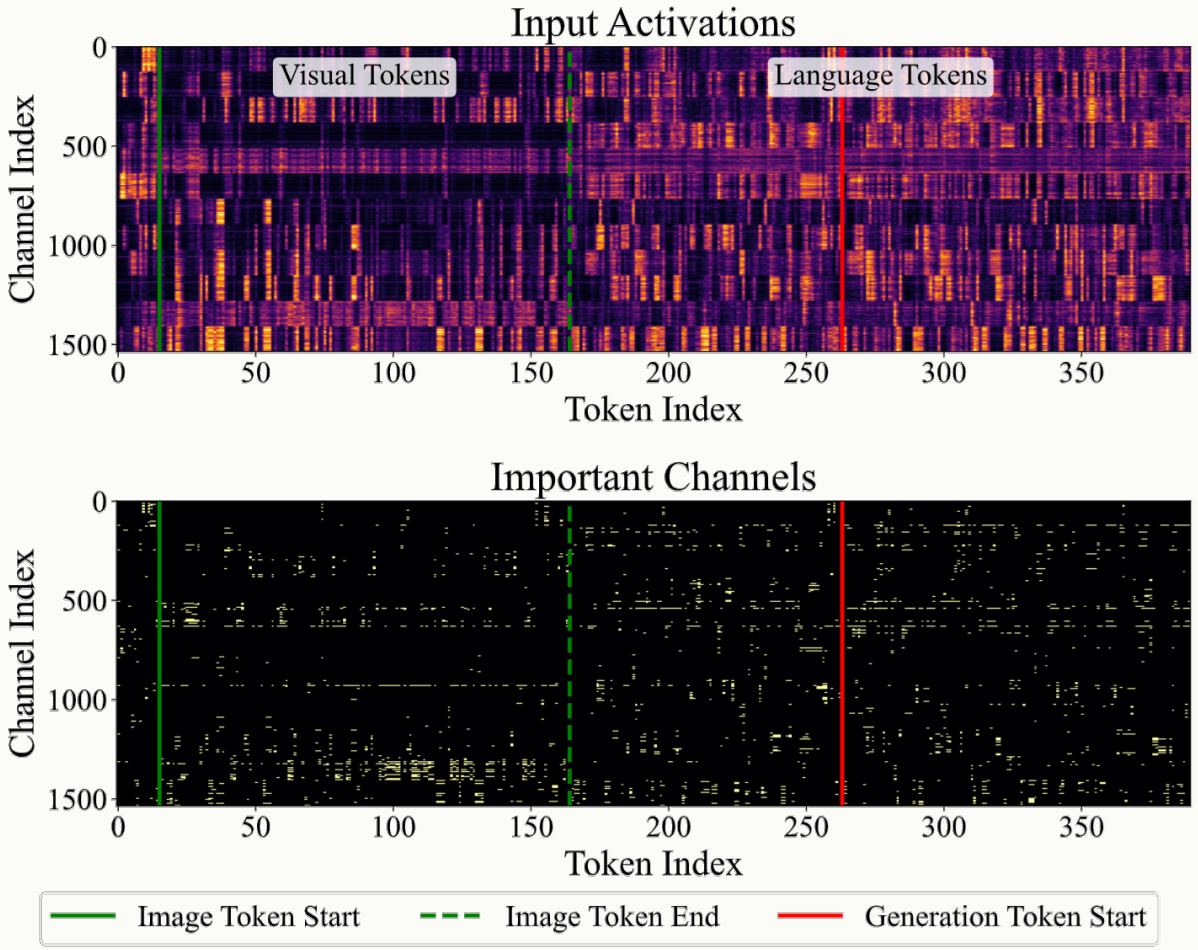
Problem: static PTQ misses token-level channel dynamics



- PTQ is a major approach for model compression, but quantization errors from low-precision conversion can severely degrade model performance.
- A common strategy to reduce quantization error is to reconstruct errors on important channels and preserve more information in these channels.
- Existing methods mainly focus on globally important channels or modality-level channel differences in VLMs.

Takeaway: the compensation granularity should be token-aware, not only channel-wise or modality-wise.

Observation 1: important-channel positions migrate across tokens



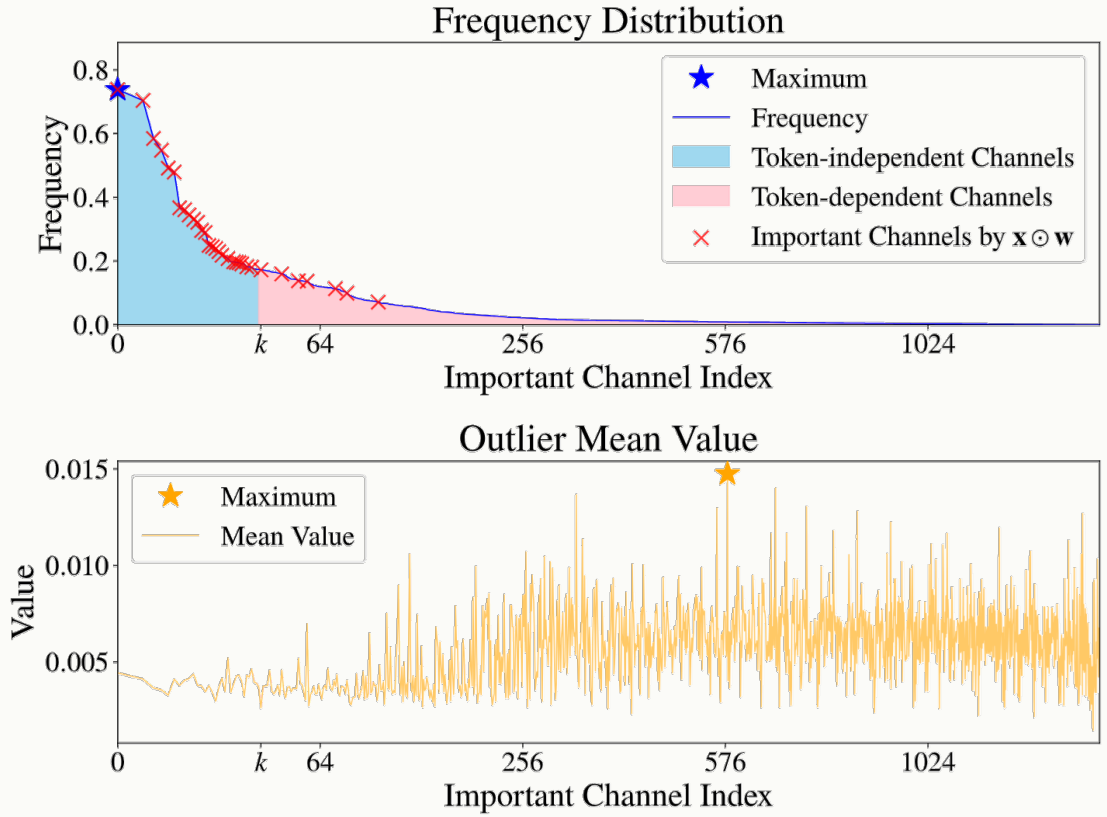
Per-token importance:

$$\mathbf{w} := \mathrm{Mean}_{\mathrm{row}}(|\mathbf{W}_f|),$$
$$\mathcal{C}_t = \mathrm{Top}\text{-}k(|x_t| \odot \mathbf{w}),$$

- We visualize the input activations and important-channel positions of a linear layer in Qwen2VL.
- Important-channel positions shift not only across modalities, but also across tokens within the same modality.
- Therefore, channel importance is not static; it changes with token semantics and context.

Takeaway: fixed global important channels cannot fully capture token-level quantization error.

Observation 2: global and local important channels coexist



- We count how frequently each important channel appears across calibration tokens.
- Only a small subset appears frequently across most tokens, forming token-independent global channels.
- Many low-frequency channels are token-dependent, but they can still have large outlier values and cause significant quantization error.

Takeaway: quantization error compensation should handle both global and token-specific important channels.

Quant Experts: token-aware adaptive error reconstruction

Token-aware error reconstruction objective

$$\arg\min_{\tilde{\mathbf{E}}^l} \left\| \left(\mathbf{E}^l - \left(\tilde{\mathbf{E}}^l_S + \tilde{\mathbf{E}}^l_R(\mathbf{x}^l) \right) \right) \mathbf{x}^l \right\|_F$$

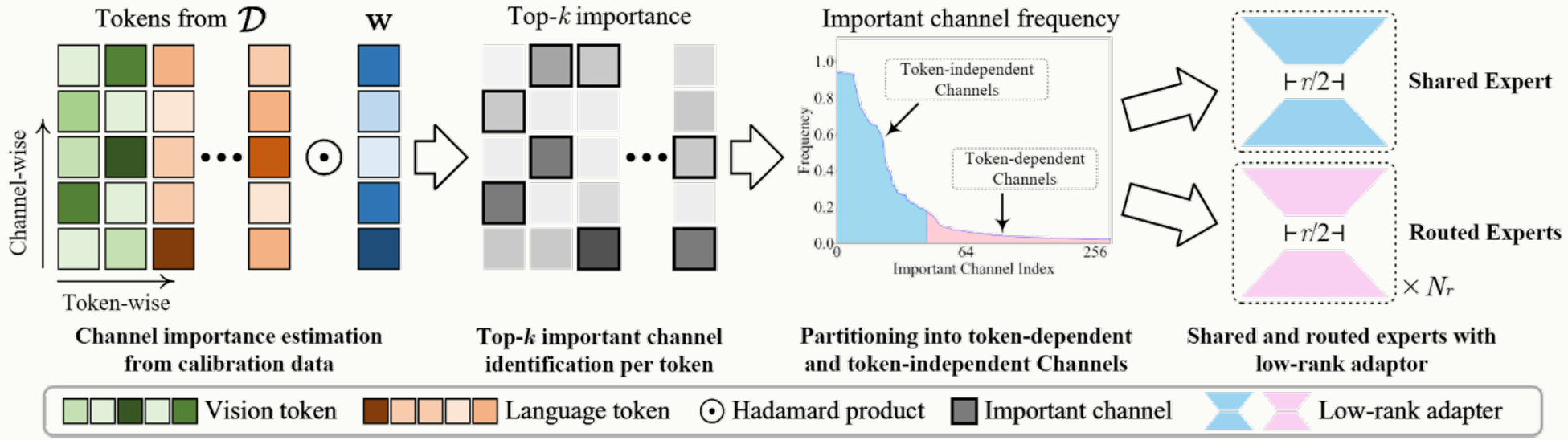
$$\tilde{\mathbf{E}}^l = \underbrace{\mathbf{L}^l_{SA}\mathbf{L}^l_{SB}}_{\text{Shared Expert}} + \underbrace{\mathbf{L}^{l,i^*}_{RA}\mathbf{L}^{l,i^*}_{RB}}_{\text{Routed Expert}}$$

$$i^* = \arg\min_i (\mathbf{R}^l |\mathbf{x}^l|)_i$$

1. **Decompose error** Split quantization error into global and token-dependent components.
2. **Reconstruct with experts** Use a shared low-rank expert for global error and routed low-rank experts for local error.
3. **Route by token** Select the routed expert that best matches the current input token.

Takeaway: QE reconstructs quantization error with one shared expert and one token-selected routed expert.

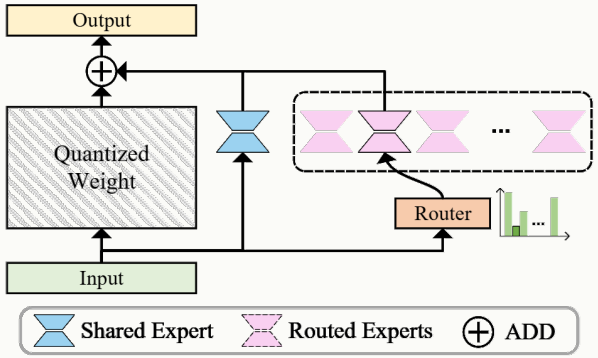
Quant Experts: token-aware adaptive error reconstruction



$$p(i) = \frac{1}{T}\sum_{t}^{T} \mathcal{O}_{t,i}^{l},$$

$$p(i,j) = \frac{1}{T}\sum_{t}^{T} (\mathcal{O}_{t,i}^{l}\mathcal{O}_{t,j}^{l}),$$

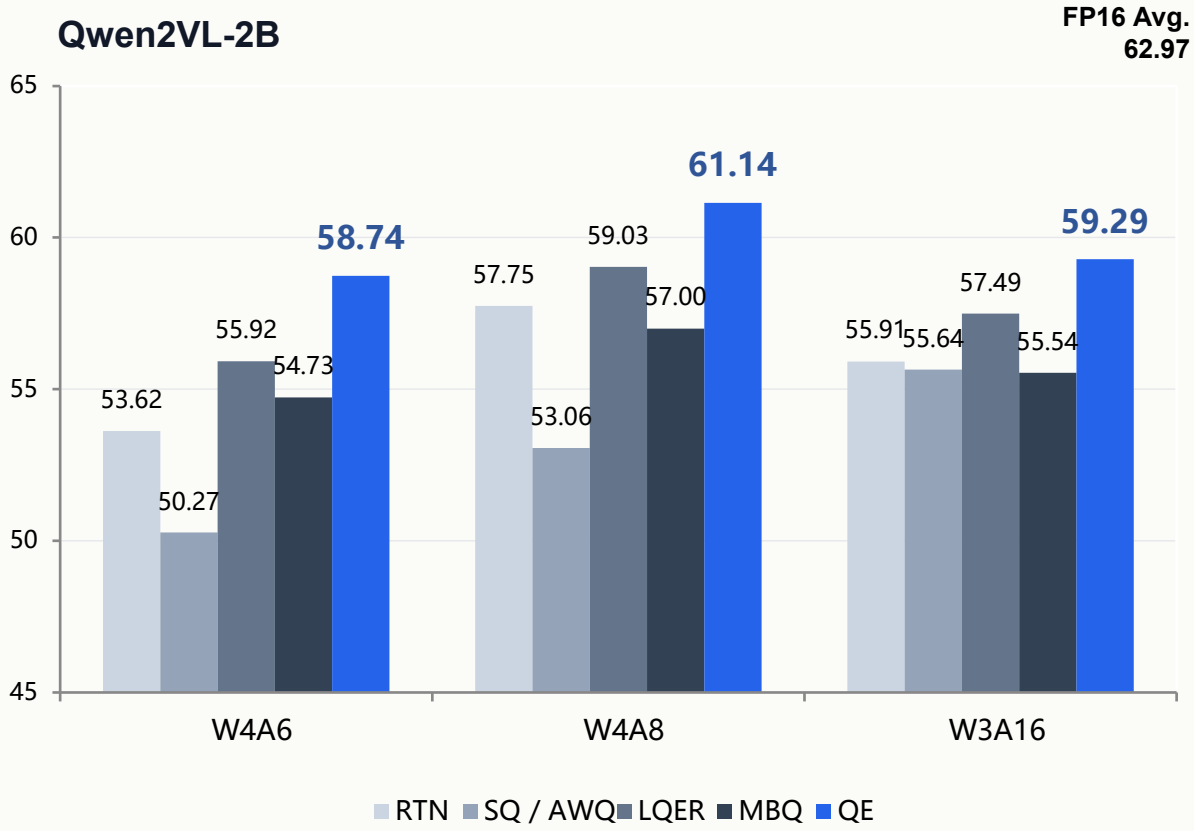
$$\mathbf{S}_{i,j} = (\log \tfrac{p(i,j)}{p(i)p(j)}) / -\log p(i,j),$$

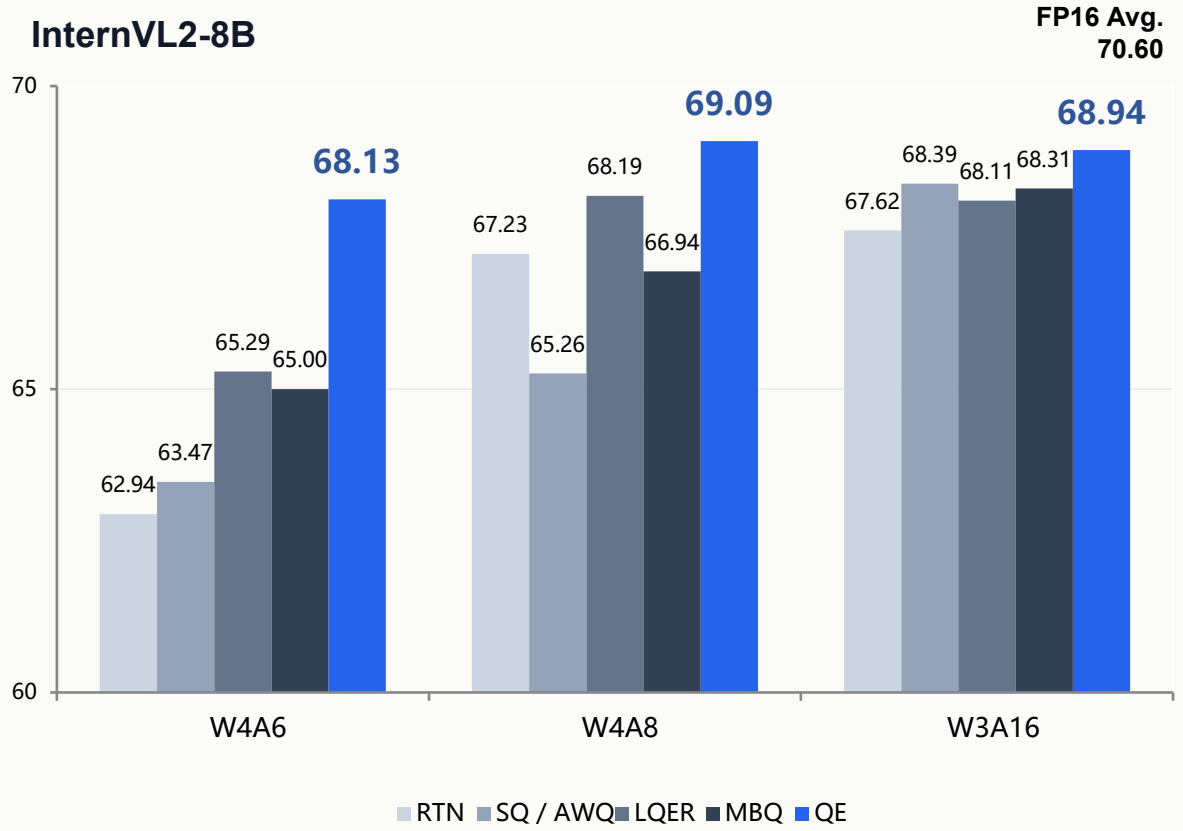


the normalized pointwise mutual information (NPMI)

Main results

Grouped bars report Avg. (↑) from Tables 1–2. QE is highlighted; the middle baseline is SQ for W4A* and AWQ for W3A16.





QE gain vs. best non-QE: Qwen2VL-2B **+2.82 / +2.11 / +1.80** | InternVL2-8B **+2.84 / +0.90 / +0.55** (W4A6 / W4A8 / W3A16)

Co-occurrence clustering for routed experts

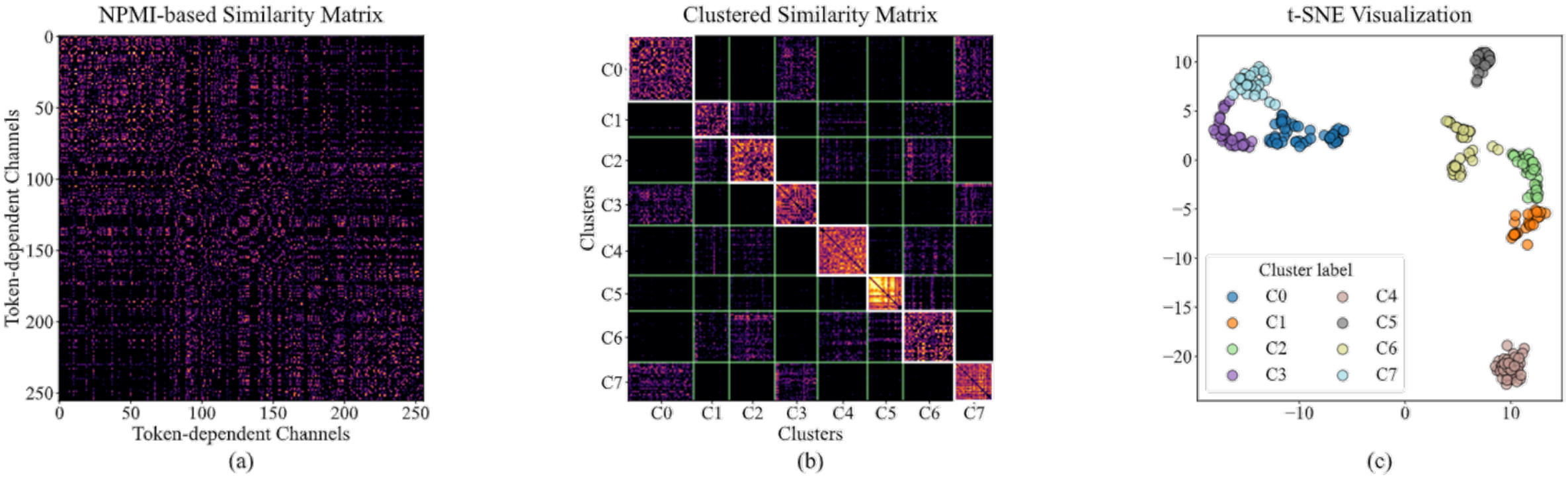


Figure 6. Illustration of the Co-Occurrence-Based Clustering in a Transformer Block of Qwen2VL-2B. (a) Similarity matrix $\mathbf{S}^l$ showing mutual co-occurrence among token-dependent channels, with brightness indicating similarity. (b) Channels with strong co-occurrence are grouped into the same cluster. (c) t-SNE [30] projection demonstrates that the clustering effectively captures their co-occurrence relations.

Ablation and efficiency

Setting	Component	MMMU (↑)	ScienceQA (↑)
FP16	-	39.89	76.95
W4A6	routed experts (REs)	34.56	68.72
	shared expert (SE)	35.22	69.61
	SE + random routing	35.89	70.00
	SE + random clustering	35.33	69.71
	QE (SE+REs)	**36.89**	**70.85**
W4A8	routed experts (REs)	36.00	71.94
	shared expert (SE)	36.78	73.13
	SE + random routing	37.89	73.67
	SE + random clustering	37.22	73.82
	QE (SE+REs)	**38.00**	**74.37**

Table 4. Ablation study results on Qwen2VL-2B model.

N_r	OCRBench	TextVQA	VizWiz	Avg. (↑)
2	68.40	73.14	59.70	67.08
4	68.50	73.13	60.41	67.35
8	69.60	73.30	60.58	67.83
16	69.90	73.52	60.75	68.06

Table 6. Impact of the number of routed experts on the performance of Qwen2VL-2B under the W4A6 quantization setting.

Complexity	Origin	QE
Computation	sd^2	$sd^2 + sd(2r + N_r)$
Memory	d^2	$d^2 + rd(1 + N_r)$

Table 7. Complexity analysis of the linear layer in QE method.

Shape (IC × OC)	W4A6	W4A8	W3A16
3584 × 3584	3.56×	3.50×	4.10×
3584 × 18944	3.60×	3.59×	4.50×
18944 × 3584	3.84×	3.77×	4.50×

Table 8. NPU speedup ratios of QE for Qwen2VL-7B linear layers compared with the fp16 model, measured during the prefill stage with a sequence length of s=128. "IC" and "OC" denote the input and output channel dimensions, respectively.

Thank you.

西安交通大學

XI'AN JIAOTONG UNIVERSITY

CVPR

JUNE 3-7, 2026

DENVER

COLORADO