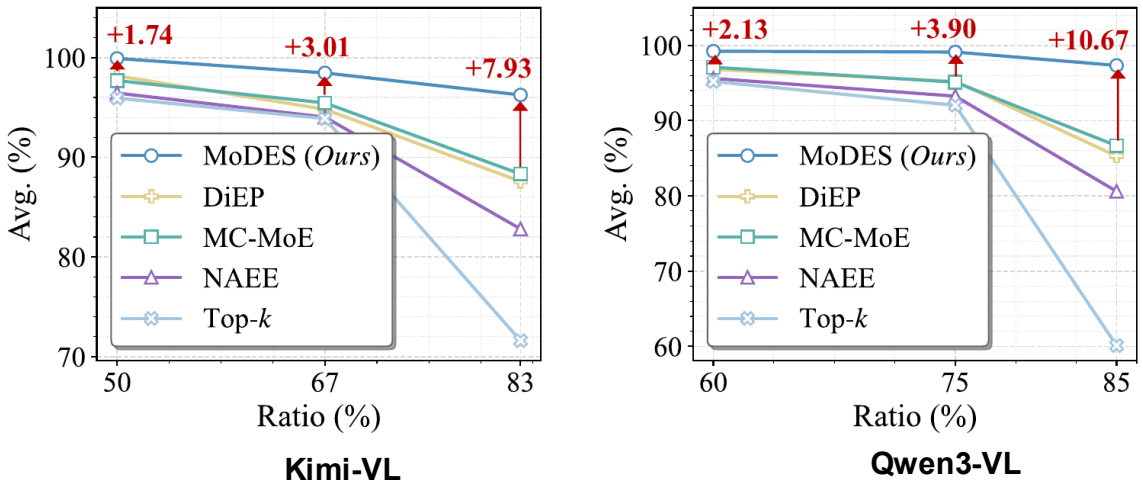
MoDES: Accelerating Mixture-of-Experts Multimodal Large Language Models via Dynamic Expert Skipping

Yushi Huang et al. • CVPR 2026

1. Problem setting and goal

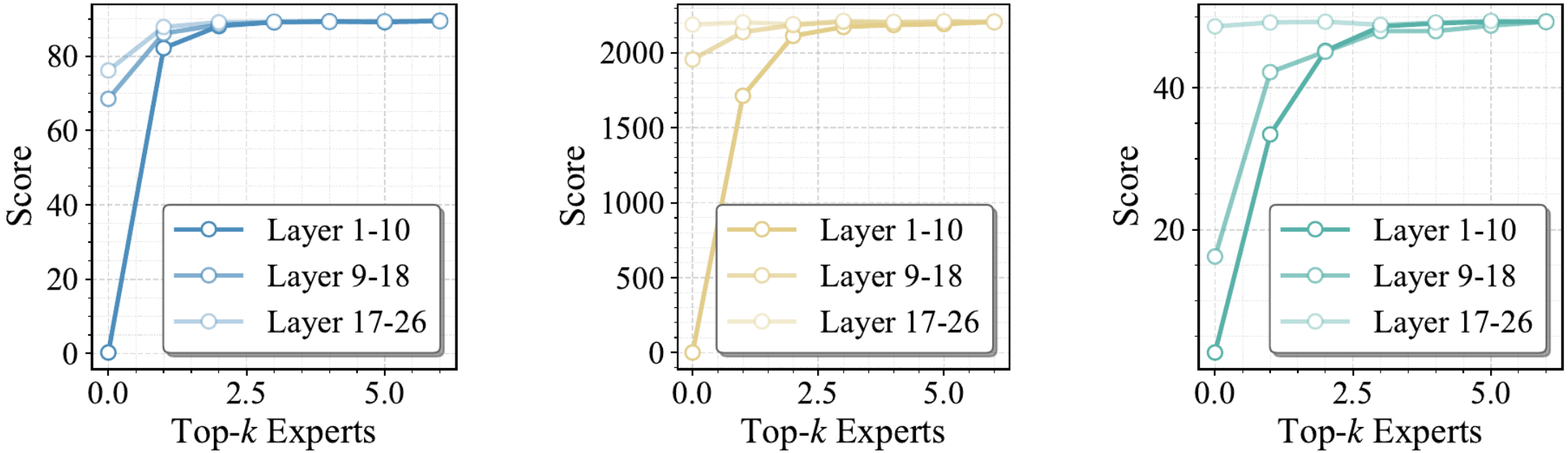
- Expert-skipping methods from text-only LLMs lose too much accuracy on multimodal MoE models.
- MoDES asks whether we can skip far more experts while keeping vision-language performance close to the original model.
- The method is training-free: calibrate once, then use a lightweight inference rule.

At 88% expert skipping on Qwen3-VL-30B, MoDES reaches 97.33 average score vs 86.66 for the best prior baseline.



2. Insight I: layer contributions are highly uneven

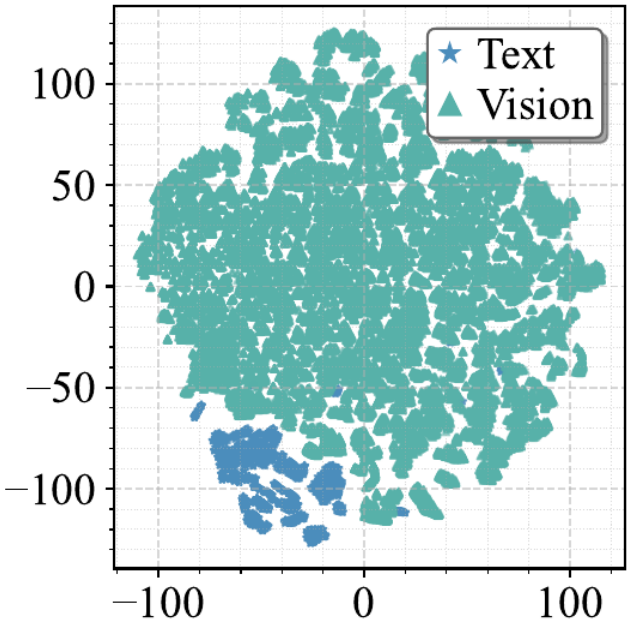
- Shallow MoE layers are much more sensitive to lowering top-k.
- So a single layer-agnostic skipping rule wastes accuracy where the model needs experts most.

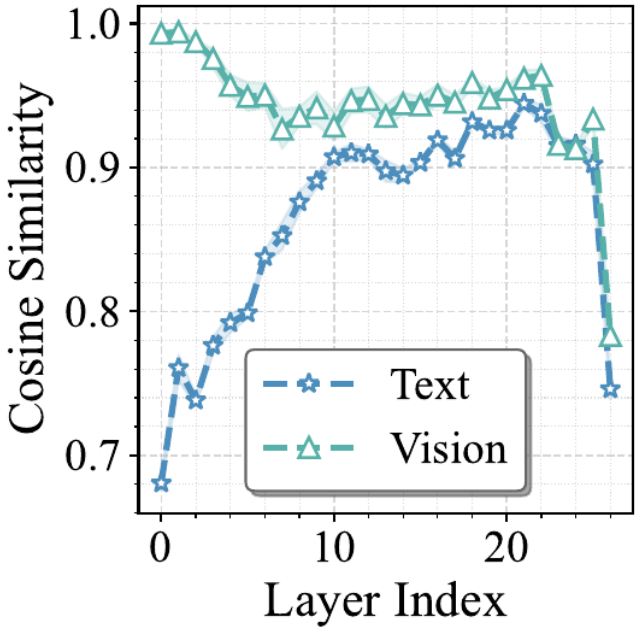


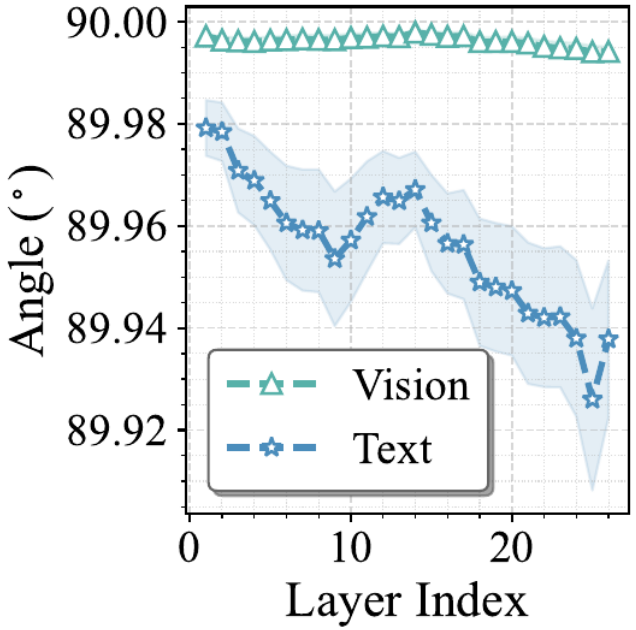
Takeaway: protect shallow layers; skip more in deeper layers.

3. Insight II: text and vision tokens behave differently

- Text and vision tokens enter FFN layers with different distributions and sensitivities.
- That means multimodal skipping should not use one shared threshold.

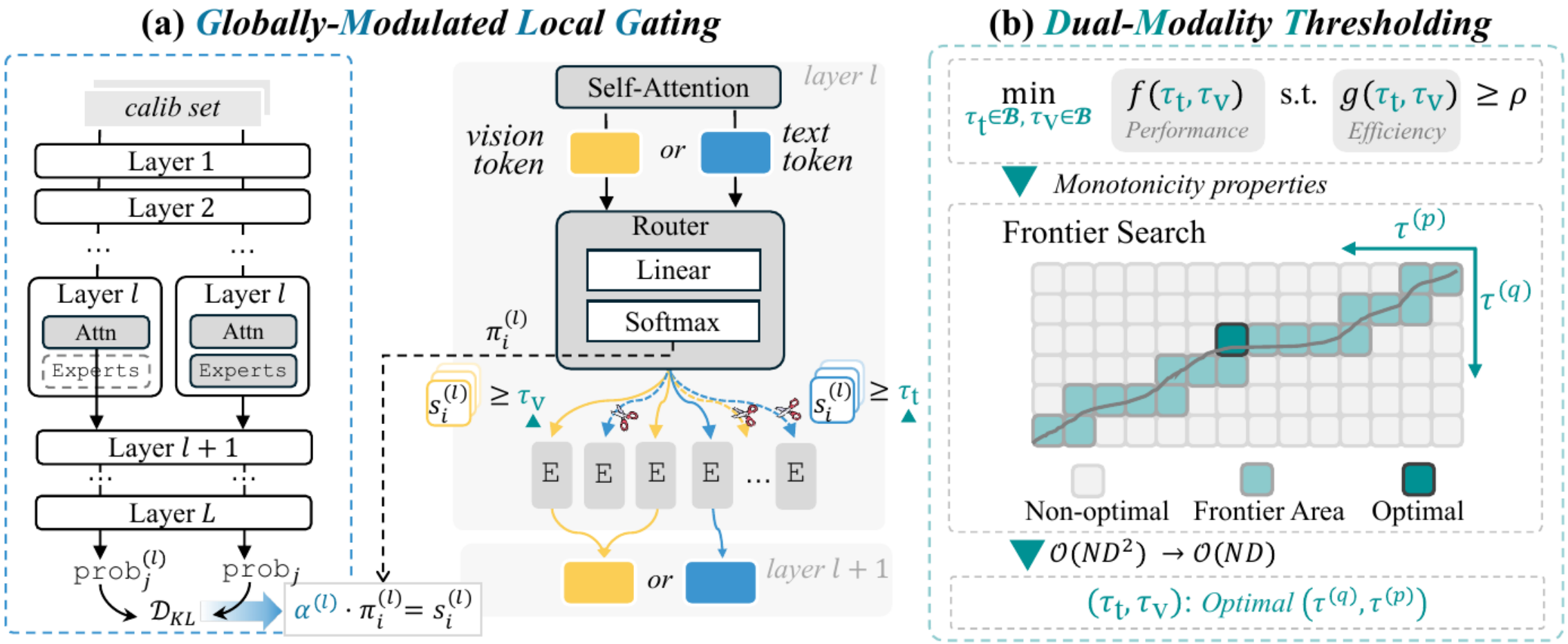






Takeaway: modality-aware thresholds are necessary.

4. MoDES overview



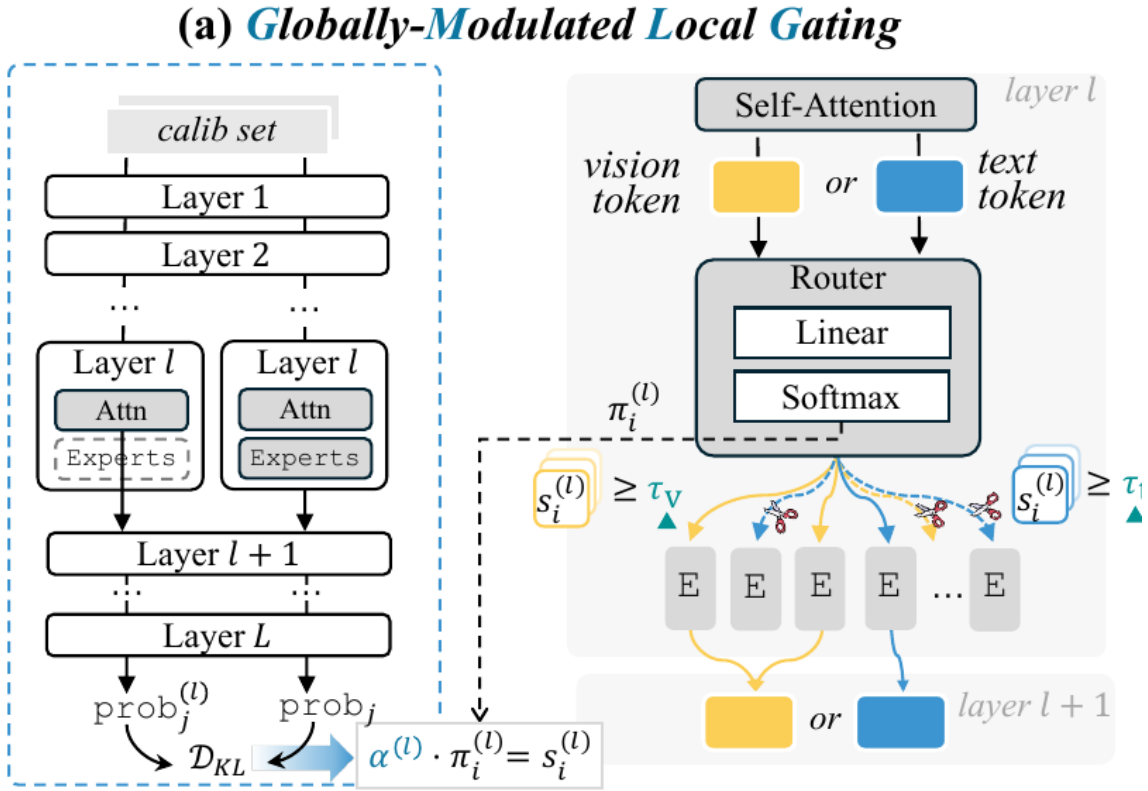
- GMLG scores experts with both layer importance and router probability.
- DMT applies different thresholds to text and vision tokens.
- Frontier search chooses the threshold pair efficiently.

5. Method I - GMLG

$$s_i^{(l)} = \alpha^{(l)} \times \pi_i^{(l)}$$

- $\pi_i^{(l)}$: router probability for expert i
- $\alpha^{(l)}$: offline layer-importance weight
- Low score = safe to skip

Same local router score should not be treated equally across all layers.

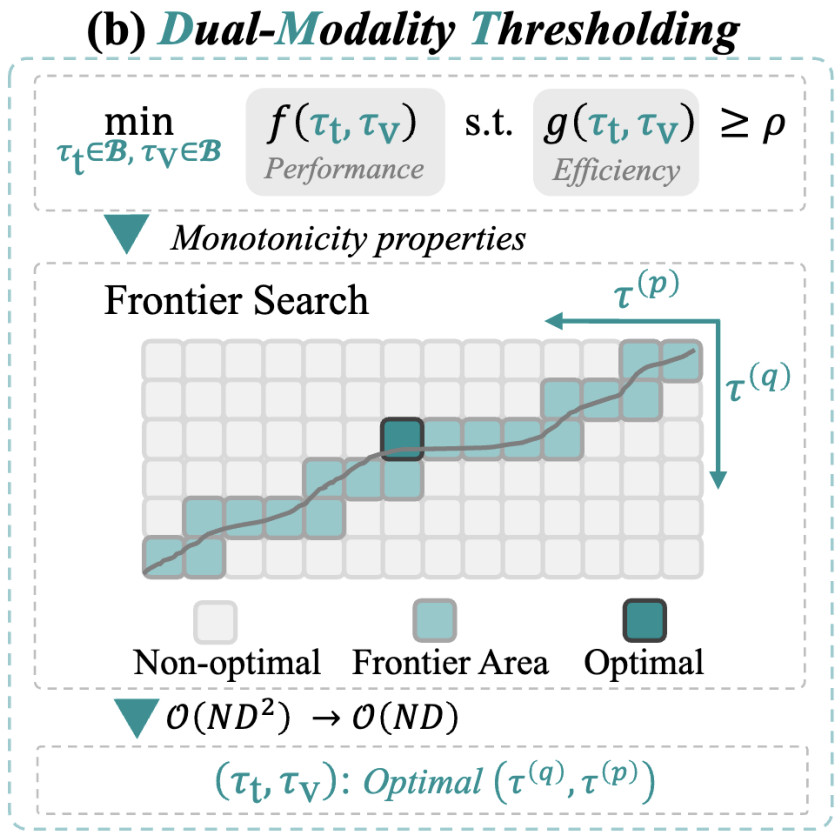


6. Method II - DMT

skip if score < τ_t (text) or τ_v (vision)

- Text tokens need a stricter threshold.
- Vision tokens can be skipped more aggressively.

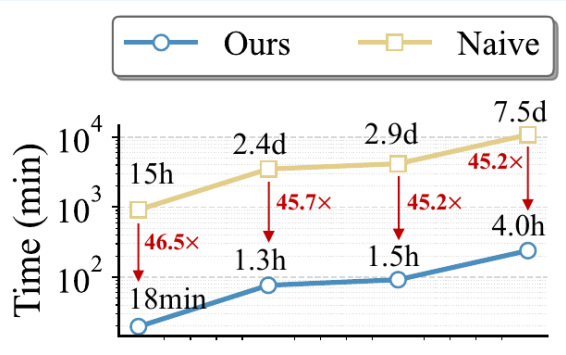
Inference cost stays tiny: only a score + threshold comparison.

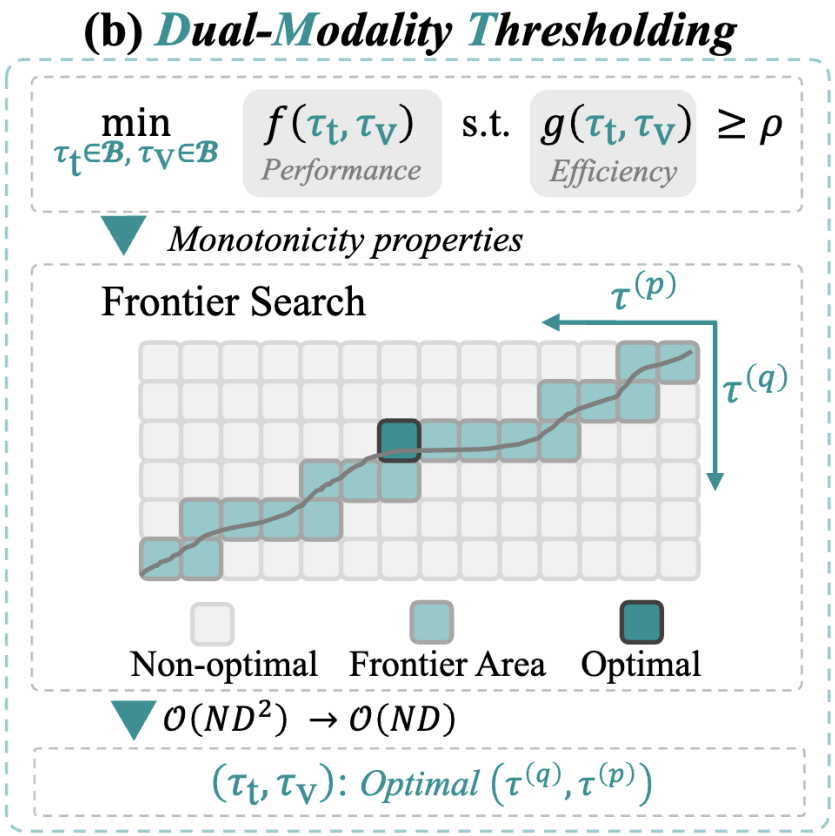


7. Frontier search makes threshold tuning practical

- Search for (τ_t, τ_v) under a target skipping ratio.
- Use monotonicity to avoid exhaustive search.
- Complexity drops from $\mathcal{O}(ND^2)$ to $\mathcal{O}(ND)$.

Paper claim: search time falls from days to hours.

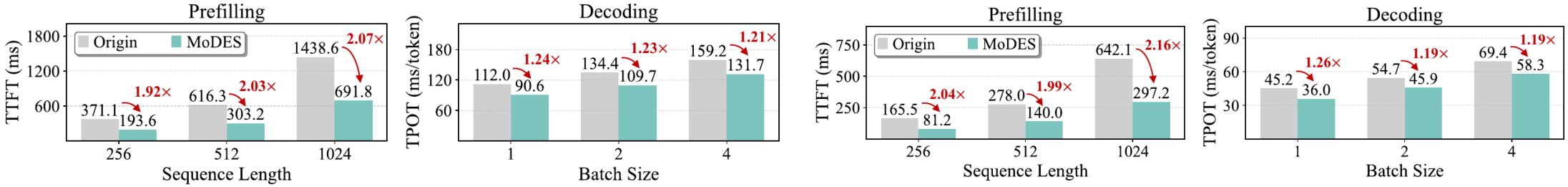




8. Main quantitative results

Model family	Skip	Best prior	MoDES	Gain
Kimi-VL-A3B	83%	88.32	**96.25**	**+7.93**
Qwen3-VL-30B	88%	86.66	**97.33**	**+10.67**
InternVL-30B	88%	86.20	**97.03**	**+10.83**
InternVL-GPT-OSS-20B	75%	90.07	**97.89**	**+7.82**

9. Efficiency and speedup



- 2.16× prefill speedup and 1.26× decoding speedup are reported in the abstract for Qwen3-VL.
- Calibration + search remain practical even for 20B-30B models.

Why it matters:
no retraining cost, better speed, much smaller quality loss.

10. Qualitative behavior

Question: Is the actress in this picture named Anne Hathaway, and which movie is this picture from?
Kimi-VL-A3B-Instruct: Yes, the actress in this picture is named Anne Hathaway, and this picture is from the movie "The Devil Wears Prada ."
MC-MoE: Yes, the actress in the picture is Anne Hathaway, and this is from the movie "Rachel Getting Married."
DiEP: The actress's name is Anne Hathaway, and the movie from is Gattino.
MoDES (*Ours*): Yes, the actress in the picture is Anne Hathaway, and this is from the movie "The Devil Wears Prada."

Question: Please carefully observe the image and come up with a caption for the image.
Kimi-VL-A3B-Instruct: The image depicts a bunch of ripe bananas resting on a wooden surface, with a computer keyboard and a cup visible in the background, suggesting a casual or home office setting.
MC-MoE: The image features a bunch of bananas prominently displayed on a wooden table. The background is blurred, with a cup of coffee placed on the table, and a newspaper also visible. The arrangement of the objects suggests a casual, yet organized setup.
DiEP: The image depicts a stack of bananas, which are yellow and slightly green at the bottom, and a bunch of green bananas at the top. There are also some bananas in the middle of the stack, but the middle bananas are green as well.
MoDES (*Ours*): The image depicts a stack of bananas placed on a wooden surface, with a computer keyboard and a cup visible in the background. The bananas are yellow with some brown spots, indicating ripeness. The setting appears to be a workspace or desk area.

Question: How many pens are shown in the video? **Kimi-VL-A3B-Instruct**: Three pans. **MC-MoE**: Two pans. **DiEP**: Two pans. **MoDES (*Ours*)**: Three pans.

Across image and video cases, MoDES stays much closer to the original model than previous skipping baselines.

11. Takeaways

- Training-free expert skipping for MoE MLLMs.
- Use both layer importance and token modality.
- Frontier search makes the rule practical to tune.

97.33
Avg. score retained
at 88% skipping
(Qwen3-VL-30B)

2.16× prefill

1.26× decode

One-line message:
skip experts, but do it with depth awareness and modality awareness.

Thank you!