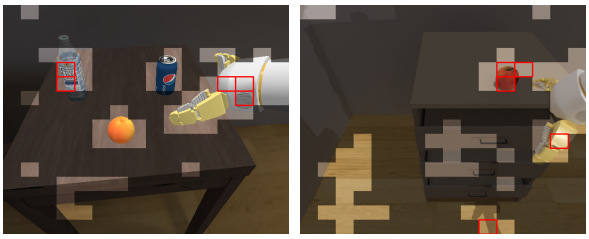
Beyond Attention Magnitude: Leveraging Inter-layer Rank Consistency for Efficient Vision-Language-Action Models

Anonymous ACL submission

Abstract

Vision-Language-Action (VLA) models excel in robotic manipulation but suffer from significant inference latency due to processing dense visual tokens. Existing token reduction methods typically treat attention magnitude as a static importance signal. In this work, we show that this signal is not uniformly reliable: high-attention tokens can help in some observations yet harm policy execution in others. Instead of abandoning attention altogether, we ask when it should be trusted. We introduce **TIES** (**T**au-guided **I**nter-layer **E**fficient **S**election), a dynamic framework that uses inter-layer token ranking consistency to predict the reliability of attention-based pruning and adapt token selection accordingly. TIES ensures robust token selection without requiring additional training. On the CogACT + SIMPLER benchmark, TIES improves average success rates by 6% while reducing token usage by 78%, and shows transferability across additional decoder families and benchmarks.



(a) *move bottle near orange* ✔ (b) *open middle drawer* ✘

Figure 1: **Informative and misleading high-attention tokens in two CogACT manipulation states.** Each panel shows a SIMPLER observation under the corresponding language instruction. Unshaded patches denote highly attended visual tokens, and the top 5 tokens are highlighted in red. The green check mark and red cross indicate whether retaining these high-attention tokens aligns with successful policy execution. In (a), high-attention tokens concentrate on task-relevant regions; in (b), they focus on distracting regions, illustrating that attention magnitude is not uniformly reliable across observations.

1 Introduction

Learning robust and generalizable manipulation policies remains a central challenge in robotics. Recent Vision-Language-Action (VLA) models address this challenge by combining visual observations and language instructions in a unified sequence model, enabling strong cross-task transfer and open-vocabulary control (Zitkovich et al., 2023; Kim et al., 2024; Li et al., 2024a; Black et al.). However, these benefits come with substantial inference cost: once visual patches are projected into the Transformer, self-attention scales quadratically with sequence length (Vaswani et al., 2017), making dense visual tokens a major obstacle to real-time deployment.

Existing efforts to accelerate VLA inference fall into two broad categories. One line focuses on model-level acceleration, such as lightweight architectures (Wen et al., 2025a), quantization (Park et al., 2024; Fang et al., 2025), and dynamic early-exit or adaptive decoding (Yue et al., 2024; Wang et al., 2025b; Tan et al., 2025). Another line targets visual token reduction, motivated by the observation that visual tokens dominate the sequence budget. In both VLMs and VLAs, many pruning methods rank tokens by attention magnitude or closely related saliency scores, sometimes supplemented with task-specific heuristics, caching rules, or action-aware modules (Chen et al., 2024; Xu et al., 2025; Yang et al., 2025; Li et al., 2025; Wang et al., 2025a; Zhang et al., 2024). This design is appealing because it is cheap and often training-free.

Yet attention magnitude is an uneasy foundation for pruning. Beyond the broader debate on whether attention weights are reliable explanations (Bibal et al., 2022), recent multimodal pruning work has shown that redundancy reduction or non-attention visual cues can rival or even outperform pure attention-based selection (Wen et al., 2025b; Alvar

et al., 2025; Zhang et al., 2025). These observations motivate moving beyond a single static pruning signal. However, most existing methods still make a global choice: they either trust attention throughout inference, or replace it with another fixed metric. This misses an important possibility: the usefulness of attention may itself depend on the current observation.

Our starting point is therefore not to ask whether attention magnitude is universally good or bad, but when it is reliable. We find that the answer is highly observation-dependent. Across tasks, and even across states within the same task, retaining top-attention tokens can either help policy execution or introduce spurious bias. This suggests that attention-based pruning should be conditional rather than uniform: the key problem is to estimate, from the current frame, whether attention rankings are trustworthy.

To this end, we study inter-layer token ranking dynamics and find that Kendall rank consistency (τ) provides a useful indicator of attention reliability. Intuitively, when token rankings remain overly fixed across layers, attention is more likely to be locked onto redundant or misleading patterns; when rankings are revised across layers, top-attention tokens tend to be more informative for control. Based on this observation, we propose **Tau-guided Inter-layer Efficient Selection (TIES)**, a lightweight training-free framework that predicts how much the current observation should trust attention-based pruning, and then adaptively mixes attention-driven selection with diversity-preserving retention. To further control overhead, TIES refreshes its pruning decision only once every fixed H control steps.

In summary, our contributions are as follows: **(1)** We provide a diagnostic study showing that attention magnitude is not a uniformly reliable pruning signal for VLA inference: top-attention tokens can help in some observations but hurt in others. **(2)** We show that this reliability is strongly observation-dependent, motivating adaptive token pruning rather than globally trusting or discarding attention. **(3)** We propose TIES, a training-free dynamic pruning framework that uses inter-layer Kendall τ to predict attention reliability and adapt the pruning strategy online. On CogACT with the SIMPLER benchmark, TIES reduces token usage by over 78% while improving average success rate by 6%, and it remains competitive in additional transfer settings.

2 Related Work

2.1 Vision-Language-Action (VLA) Models

Modern VLA models build on large-scale multimodal pretraining and robot data collected from both simulation and real-world environments (Liu et al., 2023; O'Neill et al., 2024). Following RT-2 (Zitkovich et al., 2023) and OpenVLA (Kim et al., 2024), recent systems such as CogACT (Li et al., 2024a) and π_0 (Black et al.) have expanded VLA modeling to autoregressive, diffusion-based, and flow-based decoders. While these models improve generalization and control quality, their large visual token budgets make inference increasingly expensive, especially in closed-loop control.

2.2 Efficient VLA Inference

VLA acceleration has been studied from several complementary angles, including lightweight model design (Wen et al., 2025a), quantization (Park et al., 2024; Fang et al., 2025), early-exit and dynamic execution (Yue et al., 2024), speculative decoding (Wang et al., 2025b), and token-aware action reuse (Tan et al., 2025). These approaches reduce inference cost at the architecture or decoding level and are largely orthogonal to our method. Our focus is instead on visual token selection, since visual tokens dominate the Transformer sequence length and directly determine the attention cost.

2.3 Visual Token Pruning Signals

Many visual token pruning methods are attention-centric. FastV (Chen et al., 2024) established a simple plug-and-play recipe for pruning VLM visual tokens according to attention scores. In the VLA setting, EfficientVLA (Yang et al., 2025) similarly relies on attention maps to identify task-relevant tokens, while SP-VLA (Li et al., 2025) and SpecPrune-VLA (Wang et al., 2025a) combine pruning with additional saliency, scheduling, or action-aware heuristics. VLA-Cache (Xu et al., 2025) further exploits temporal redundancy through token caching. These methods show that token reduction is practical, but most still implicitly assume that high-attention tokens are broadly trustworthy throughout inference.

At the same time, recent VLM work has begun to question attention magnitude as a universal pruning signal. DivPrune (Alvar et al., 2025) emphasizes diversity and redundancy reduction, Wen et al. (2025b) argue that duplicated tokens matter more

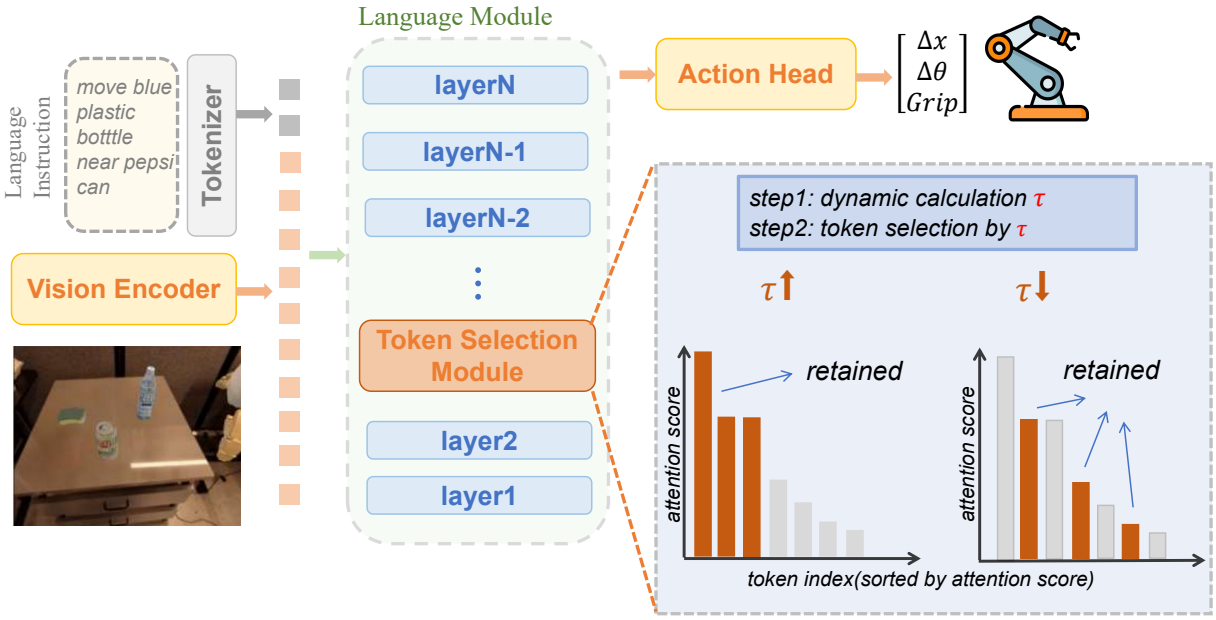


Figure 2: **Overview of TIES.** During offline calibration, TIES estimates reference quantiles of inter-layer Kendall τ values from sampled frames. During online inference, it recomputes τ once every fixed H control steps, then uses the resulting reliability estimate to determine how many tokens should be retained by attention-based selection versus diversity-preserving sampling.

than "important" ones under many settings, and Zhang et al. (2025) show that visual cues beyond text-visual attention can improve pruning. Several recent studies, e.g., FasterVLM(Zhang et al., 2024), also report that attention magnitude alone can be brittle under distributional or architectural changes. Our work is complementary to this line, but differs in a key way: rather than replacing attention with another globally applied metric, we model the *observation-dependent reliability* of attention itself, and use a cheap inter-layer consistency indicator to decide how much the current frame should trust attention-based pruning.

3 Key Insights

3.1 Preliminary

Vision-Language-Action (VLA) models map image observations and language instructions to robot actions through a multimodal Transformer. Let the language input contain N_l tokens and the visual input contain N_v tokens. The total sequence length is $N = N_l + N_v$, and self-attention therefore incurs $O(N^2)$ computation and memory cost. In practice, $N_v \gg N_l$, so inference latency is dominated by visual tokens. This makes visual token reduction a central efficiency problem for real-time control.

Most existing pruning methods rank visual tokens by attention magnitude. Given an attention matrix A, where $A_{i,j}$ is the attention weight from token i to token j, a common importance score for token j is $S_j = \sum_{i=1}^{N} A_{i,j}$. Tokens are then sorted by S_j, and only the top-k visual tokens are retained for subsequent layers. This attention-based ranking is simple and effective in many cases, but it also implicitly assumes that high-attention tokens are consistently reliable indicators of utility for downstream action generation.

3.2 Negative Effects of High-Attention Tokens

Our investigation is motivated by recent discussion regarding the interpretability of attention mechanisms. While conventional wisdom suggests that tokens with high attention weights are semantically critical, recent studies offer a more nuanced perspective. For instance, Xiao et al. (2023) suggests that high-attention tokens primarily serve as sinks for information flow during the Transformer's forward pass to maintain numerical stability, rather than being purely task-relevant. Similarly, Zhang et al. (2024) observed that attention maps in cross-modal Transformers are often noisier and less reliable proxies for token importance compared to those in standalone vision encoders. Furthermore, Darcet et al. (2023) argues that vision encoders disperse semantic information across multiple patches and often assign high weights to "register" tokens that lack semantic meaning. This renders individ-

ual visual tokens uninterpretable to human cognition and casts doubt on the validity of simple saliency heuristics.

We challenge the presumed synergy between high attention and token importance through a counter-intuitive experiment. Specifically, we evaluate the performance impact of retaining a reduced set of tokens under two extreme regimes: **Top-k**, where we only keep k tokens with the highest attention scores, and **Bottom-k**, where we retain a subset of tokens excluding the highest-ranked ones (specifically, we keep the last k tokens among the top K). We compare these against the original baseline with full 256 tokens reserved. As shown in Figure 3, the role of high-attention tokens varies drastically across different robotic tasks.

Performance Inversion. In the *Drawer* task, we observe a striking logic inversion. The Bottom-45 strategy achieves a success rate of 77.31%, which not only significantly outperforms the Top-45 strategy but also exceeds the full-token Baseline by a margin of 4.61%. This suggests that in certain manipulation tasks, the highest-ranked tokens can act as spurious signals or noise that mislead the model's policy, and removing them can effectively serve as a denoising mechanism.

Task Dependency. Conversely, in the *MoveNear* task, the conventional assumption holds true. Top-45 achieves 80.0%, clearly outperforms Bottom-45 with 73.75%. Furthermore, as k decreases from 45 to 25, the performance of the Bottom-k strategy in *MoveNear* collapses more rapidly compared to the *Top-k* strategy, indicating a higher reliance on primary attention signals for this specific task.

These results demonstrate that the credit or reliability of high-attention tokens is highly task-dependent and can even vary across different states within the same task. Blindly relying solely on magnitude as a selection criterion is insufficient and can be detrimental to policy performance.

3.3 Identifying Indicators for Attention Reliability

To develop an adaptive token selection mechanism, the primary challenge lies in identifying a reliable metric that distinguishes whether high-attention tokens are beneficial (Type 1) or detrimental (Type 2) to policy performance.

Limitations of Static Statistical Metrics. A natural starting point is to examine the attention distribution's sharpness using Shannon Entropy. One

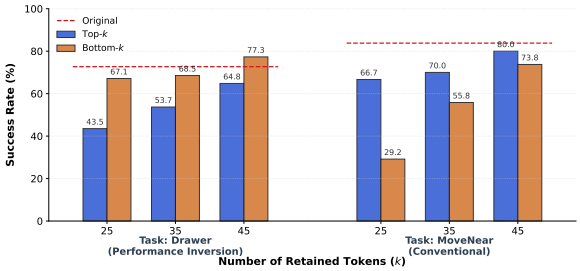


Figure 3: **Top-k versus Bottom-k pruning on two SIMPLER tasks.** Results are shown for CogACT with a full-token baseline of 256 visual tokens. In *MoveNear*, retaining the highest-attention tokens is more effective, whereas in *Drawer*, removing the highest-attention tokens can yield better success rates. This contrast motivates predicting when attention-based pruning should be trusted.

might hypothesize that low-entropy distributions represent more confident and thus more reliable token selection. However, our analysis on the SIMPLER benchmark refutes this. As shown in Figure 4, the entropy distributions for both Type 1 and Type 2 scenarios exhibit substantial overlap. The mean entropy for both types also remains close, indicating that magnitude-based or distribution-based static metrics do not reliably separate reliable frames from unreliable ones.

From Static Magnitude to Dynamic consistency. Inspired by the observation that token importance is not static but flows and evolves across Transformer layers(Zhang et al., 2024), we shift our focus from absolute weights to the inter-layer rank dynamics. For each visual Transformer layer l, we compute an attention score $S_j^{(l)} = \sum_i A_{i,j}^{(l)}$ for every visual token j, sort visual tokens by $S_j^{(l)}$, and keep the top K indices as a ranked list $\mathbf{r}^{(l)}$. For two adjacent layers, we take the intersection of their top-K sets and compute Kendall rank correlation on the relative order of the common tokens. Our final consistency metric τ is the average of these adjacent-layer Kendall scores across the selected layer range. A higher τ indicates that the model's focus remains on a relatively more stable set of tokens across L layers, while a lower τ reflects more active rank revision.

The Divergence of Rank Consistency. Our quantitative analysis reveals a clear divergence in rank consistency between the two types. As illustrated in Figure 4, the Kendall τ distributions for Type 1 and Type 2 are well separated: Type 1 exhibits lower consistency, indicating a more dy-

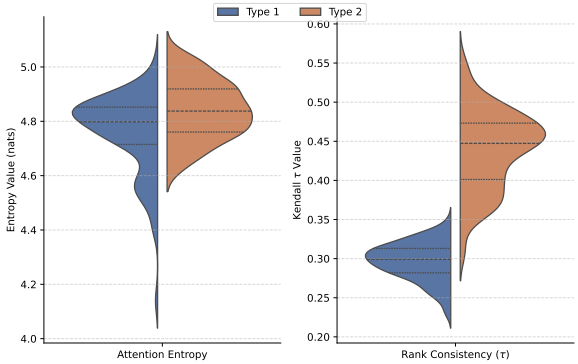


Figure 4: **Distribution of entropy and Kendall correlation** The figure shows the distribution of Shannon entropy and Kendall correlation values for frames where high-attention tokens are beneficial (Type 1) and detrimental (Type 2). The entropy can hardly discriminate the two types, while Kendall correlations separate them well.

namic reassessment of token importance as the representation evolves. Type 2 exhibits higher consistency, where the attention rankings remain relatively fixed across the depth of the model. Statistical evaluation shows a strong negative correlation between τ and task types, achieving an AUC of 0.91 in predicting attention reliability.

4 Method

The core of our approach is TIES, a training-free dynamic token pruning framework. The detailed algorithm is in Appendix1. In a word, TIES adaptively modulates the reliance on high-attention tokens based on the inter-layer rank consistency (τ) of the input and choose the pruning strategy accordingly.

4.1 τ-guided Token Selection

We utilize the Kendall rank correlation coefficient τ as a dynamic indicator to distinguish whether high-attention tokens are beneficial or detrimental to policy performance. For each frame, TIES constructs a top-K ranking at each visual Transformer layer and computes the frame-level score τ_t by averaging Kendall correlations between adjacent layers across L layers. We do ablation study in 6, and use the K=30 and L=16 across all experiments unless otherwise noted.

Let K_{ret} denote the retained-token budget and let $y_t \in [0, 1]$ denote the fraction of retained tokens allocated to the attention-based selector, i.e., FastV-style top-attention retention. The remaining fraction $1 - y_t$ is filled by a alternative stragegy, which is Uniform in our main setting. During offline calibration, we sample $M = 100$ frames and estimate the 25th, 50th, and 75th percentile values of τ, denoted by q_{25}, q_{50}, and q_{75}.

Hard-TIES uses a binary allocation rule,

$$y_t = \begin{cases} 1, & \tau_t \leq q_{50}, \\ 0, & \tau_t > q_{50}. \end{cases}$$

Thus, low-τ frames use purely attention-based retention, whereas high-τ frames switch entirely to the fallback strategy.

Soft-TIES replaces this hard switch with a piecewise-linear interpolation:

$$y_t = \begin{cases} 1, & \tau_t \leq q_{25}, \\ \frac{q_{75}-\tau_t}{q_{75}-q_{25}}, & q_{25} < \tau_t < q_{75}, \\ 0, & \tau_t \geq q_{75}. \end{cases}$$

Equivalently, on the interval $[q_{25}, q_{75}]$, y_t can be written as a linear function $y_t = k\tau_t + b$. TIES then retains $\lfloor y_t K_{ret} \rfloor$ tokens by attention ranking and fills the remaining $K_{ret} - \lfloor y_t K_{ret} \rfloor$ slots by Uniform sampling from the unselected visual tokens. Consequently, Soft-TIES degenerates to pure attention-based pruning when τ_t is low and to pure Uniform retention when τ_t is high.

4.2 Implementation Details

To minimize the computational overhead of calculating τ in high-frequency control loops, we use a fixed refresh horizon instead of a similarity trigger. Concretely, TIES recomputes τ on the current frame once every H control steps and reuses the resulting allocation rule for the intervening steps. This design keeps the runtime predictable and makes the extra cost of τ estimation easy to amortize. In the main results, we use $H = 20$, which provides a favorable latency-performance trade-off.

5 Experiment

5.1 Experimental Settings

Simulation Implementation Details. To evaluate our approach, we mainly conduct experiments on the SIMPLER(Li et al., 2024b) environment, a simulation-based benchmark specifically designed for table-top manipulation tasks. SIMPLER aims to faithfully reproduce real-world robotic dynamics for platforms such as the Google Robot and WidowX, and has been shown to exhibit strong

sim-to-real consistency. In this setup, the VLA model receives 224×224 RGB visual observations along with natural language task descriptions (e.g., "Pick coke can") as inputs, and generates a sequence of actions in a 7-DoF Cartesian control space. SIMPLER offers two evaluation protocols: Visual Matching, which emphasizes visual consistency with real-world scenes, and Variant Aggregations, which introduces a variety of environmental variations including changes in lighting, backgrounds, and surface textures. For the Google robot, both protocols are supported, and each includes the same set of four manipulation tasks: (1) Pick coke can; (2) Move near; (3) Open/close drawer; and (4) Open the top drawer and place the apple. We adopt task success rate as the primary evaluation metric.

Baselines. We conduct our primary experimental validation of TIES on the CogACT(Li et al., 2024a) framework, which features a sophisticated architecture incorporating DINOv2(Oquab et al., 2023) and SigLIP(Zhai et al., 2023)) as vision encoders, a Llama2-7B(Touvron et al., 2023) backbone for multimodal reasoning, and a Diffusion Transformer (Peebles and Xie, 2023) for action trajectory synthesis. To evaluate its performance, we compare TIES against several state-of-the-art baselines, including: FastV(Chen et al., 2024), which enhances inference speed via redundant visual token pruning; VLA-Cache(Xu et al., 2025), which employs temporal analysis to skip redundant computations by caching static tokens; and EfficientVLA(Yang et al., 2025), a comprehensive approach integrating token pruning, layer reduction, and decoder optimization. We additionally report two advanced VLM methods DivPrune(Alvar et al., 2025) and DART(Wen et al., 2025b), which replace attention with other metrics to guide the pruning.

Implementation Details. In all experiments, pruning was applied from the second Transformer layer onward. All our empirical evaluations were conducted on NVIDIA A800 GPUs.

5.2 Main Results on SIMPLER

We evaluate the performance of TIES using the SIMPLER benchmark across two distinct protocols: Visual Matching (VM) and Variant Aggregation (VA). As summarized in Table 1, the two TIES variants achieve the strongest average performance among the evaluated pruning methods on the two protocols and both improve over the full-token CogACT baseline in average success rate.

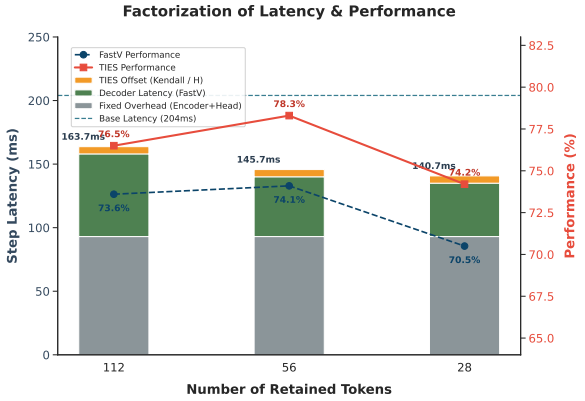


Figure 5: **Factorization of step latency and model performance under different vision token pruning ratios.** The bar decomposes the step latency into non-optimizable overhead (Vision Encoder + Action Head), Transformer Decoder latency, and our introduced Kendall sampling offset (5.7 ms). The line illustrates the average performance.

In the Visual Matching protocol, which emphasizes consistency with real-world scenes, Hard-TIES achieves the highest average success rate at 78.3% while retaining only 56 tokens, a 5.6% improvement over the unpruned CogACT baseline. In the Variant Aggregation protocol, where lighting and texture changes make control more challenging, Soft-TIES reaches the highest average success rate at 67.6%, compared with 59.9% for the unpruned baseline. These results suggest that adaptive trust in attention-based pruning can improve efficiency without sacrificing policy quality.

Task-specific analysis further illustrates the limitations of standard attention-driven selection. In the Variant Aggregation *Drawer* task, the baseline achieves a 28.8% success rate, while Soft-TIES improves this to 44.3%. This result indicates that fewer tokens can sometimes yield better control when highly attended tokens emphasize spurious signals. By using Kendall τ to identify such cases, TIES reduces reliance on high-magnitude tokens when they appear less trustworthy.

6 Ablation Study

6.1 Step Latency factorization under different pruning ratios

To investigate the efficacy of our proposed dynamic selection mechanism, we conduct a detailed ablation study by varying the number of retained vision tokens (112, 56, and 28). As shown in Figure 5, we factorize the step latency and trace the corresponding performance variations to analyze the

SIMPLER	Method	PickCan	MoveNear	Drawer	DrawerApple	Average
Visual Matching	CogACT	91.7	83.8	72.7	42.6	72.7
	FastV	92.6	81.4	69.8	52.4	74.1
	VLA-Cache	92.0	**83.3**	70.5	51.6	74.4
	EfficientVLA	94.7	82.4	69.8	55.4	75.5
	Divprune	82.3	71.3	79.6	**60.2**	73.4
	DART	**95.7**	78.3	76.4	39.8	72.6
	Soft-TIES	93.3	79.6	76.9	57.4	76.8
	Hard-TIES	91.0	80.0	**86.6**	55.6	**78.3**
Variant Aggregation	CogACT	88.6	76.5	28.8	45.6	59.9
	FastV	91.4	78.6	27.6	50.6	62.1
	VLA-Cache	91.7	**79.3**	32.5	45.8	62.3
	EfficientVLA	**94.4**	77.2	27.6	51.3	62.6
	Divprune	82.9	61.5	**48.6**	57.1	62.5
	DART	89.4	62.8	43.1	60.3	63.9
	Soft-TIES	88.7	76.2	44.3	**61.1**	**67.6**
	Hard-TIES	89.7	74.7	41.0	54.6	65.0

Table 1: **Main results on CogACT with the SIMPLER benchmark.** We report task success rates under the Visual Matching and Variant Aggregation protocols, with all pruning methods retaining 56 visual tokens. Bold indicates the best result within each column. TIES improves the average success rate over the full-token baseline while operating with substantially fewer visual tokens.

edge of our method over the baseline and FastV. From the latency perspective (bars), the bottom runtime for the Vision Encoder and Action Head remains stable, establishing a theoretical lower bound for structural optimization. Since our method executes the Kendall step periodically every $H = 20$ steps, it introduces a predictable, constant computation offset of $\frac{114\,\text{ms}}{20} = 5.7$ ms atop FastV's decoder curve. Crucially, this marginal latency investment yields 5% performance gain (red line). This demonstrates that our dynamic evaluation paradigm successfully offsets the information loss of token pruning by allocating computation to where it matters most, delivering a superior Pareto-optimal curve for resource-constrained deployment.

6.2 Sensitivity Analysis on Kendall τ Hyper-parameters

We conduct a joint sensitivity analysis on two core hyper-parameters: tracking layers (L) and Top-K in Figure 6. First, evaluating L reveals a quasi-linear increase in localized Kendall latency, expanding from 59 ms ($L = 4$) to 185 ms ($L = 32$) due to deeper cross-layer tensor slicing and sort-ranking. However, this resource consumption yields diminishing returns for downstream tasks. On our primary curve (Top-$K = 30$), increasing L from 4 to 16 boosts the success rate from 70.48% to a peak of 78.28%, proving that intermediate attention aggregation distills stable rank correlation. Conversely, extending L to 32 degrades performance to 78.16%, as overly deep layers introduce redundant features and background jitter. Meanwhile, Top-$K = 30$ consistently outperforms other settings. Consequently, we select $L = 16$ and Top-$K = 30$ as the default configuration for resource-constrained deployment.

6.3 Comparisons of Alternative Strategies

Figure 7 shows the average performance of different token selection strategies on the Visual Matching task with 56 tokens retained. Specifically, Middle retains tokens within the median range of attention scores, while Uniform uniformly samples tokens from the remaining pool after excluding the top-ranked ones. DivPrune(Alvar et al., 2025) serves as a state-of-the-art diversity-aware baseline for VLMs that incorporates deduplication.

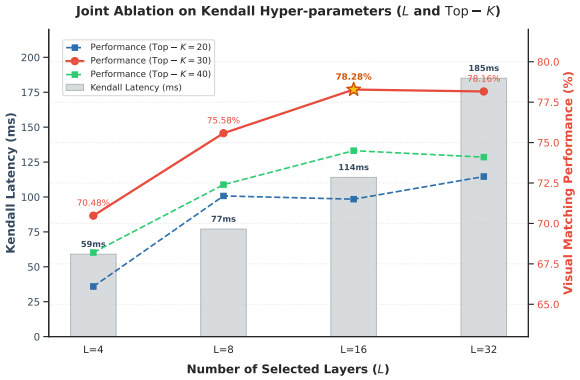


Figure 6: **Joint ablation of Kendall layer range(L) and Kendall** (Top- K)**.** The gray bars illustrate the computational latency of the Kendall computation. The gold star highlights our frozen hyper-parameter suite ($L = 16$, Top- $K = 30$), which yields the optimal balance along the accuracy-efficiency Pareto frontier.

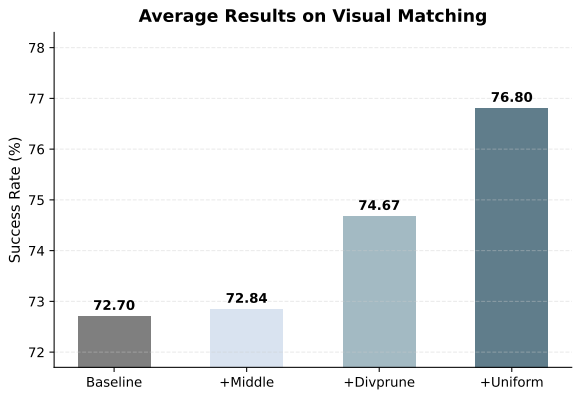


Figure 7: **Alternative fallback strategies within TIES.** Average success rates are reported on the SIMPLER Visual Matching protocol with 56 retained tokens. Once τ identifies cases where attention is less reliable, several simple fallback strategies improve over the full-token baseline, showing that the reliability indicator matters more than any single fallback heuristic.

The results demonstrate that when integrated with hardTIES, all of these strategies consistently outperform the full-token baseline under the same pruning budget, with Uniform achieving the best performance. This success across different strategies indicates that once τ accurately identifies the reliability boundaries of top attention tokens, even simple rules can produce decent results.

6.4 Generalization Across Diverse Architectures and Benchmarks

To assess transferability across heterogeneous architectures and environments, we evaluate our method using OpenVLA (Kim et al., 2024) on the SIMPLER benchmark. Unlike CogACT, OpenVLA employs an auto-regressive decoder requiring multiple forward passes per action, making inference efficiency particularly critical. As shown in Table 2, TIES consistently outperforms FastV across these transfer experiments under different decoding paradigms, with the largest gain achieved under Variant Aggregation.

Furthermore, we validate our approach on the widely adopted LIBERO benchmark (Liu et al., 2023) using the fine-tuned OpenVLA-OFT model (Kim et al., 2025). TIES again delivers superior success rates compared to vanilla FastV. This consistent improvement across distinct benchmarks and decoder structures demonstrates that our proposed reliability signal possesses strong cross-architecture generalization. However, TIES does not uniformly exceed the full-token baseline, a gap most noticeable on LIBERO. This is primarily due to a ceiling effect: since the baseline already achieves an exceptionally high score (96.90%) close to a perfect success rate, the margin for further optimization is inherently constrained for any token-reduction technique.

	SIMPLER VM	SIMPLER VA	LIBERO
Full tokens	34.30	39.30	96.90
FastV	27.99	31.47	92.70
TIES	28.31	**39.44**	94.30

Table 2: **Transfer results across additional architectures and benchmarks.** We compare TIES with FastV using 56 retained visual tokens: OpenVLA is evaluated on SIMPLER Visual Matching and Variant Aggregation, and OpenVLA-OFT is evaluated on LIBERO. Bold indicates the best pruning result in each column, and underlining indicates the second-best pruning result.

7 Conclusion

In this work, we show that relying solely on high attention magnitude is insufficient for efficient VLA inference and can occasionally hurt manipulation performance. We further show that the reliability of attention-based pruning depends on the current observation, and that inter-layer token rank consistency provides a useful signal for estimating this reliability. Building on this insight, we develop TIES, a training-free dynamic token-selection strategy that adjusts how much it trusts attention during inference. Experiments on CogACT and additional transfer settings show that TIES can improve the efficiency-performance trade-off of visual token pruning for robotic manipulation.

Limitations

While TIES effectively mitigates attention traps and enhances inference efficiency, it relies on correlation between inter-layer rank consistency and token importance, lacking deep theoretical grounding in the causal chain of model decision-making. Our method leaves room to explore more sophisticated cross-modal alignment that could further refine the interaction between visual features and complex linguistic instructions.

References

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. 2025. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9392–9401.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, and 1 others. π0: A vision-language-action flow model for general robot control. corr, abs/2410.24164, 2024. doi: 10.48550. *arXiv preprint ARXIV.2410.24164*.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. 2023. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*.

Hengyu Fang, Yijiang Liu, Yuan Du, Li Du, and Huanrui Yang. 2025. Sqap-vla: A synergistic quantization-aware pruning framework for high-performance vision-language-action models. *arXiv preprint arXiv:2509.09090*.

Moo Jin Kim, Chelsea Finn, and Percy Liang. 2025. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, and 1 others. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, and 1 others. 2024a. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, and 1 others. 2024b. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*.

Ye Li, Yuan Meng, Zewen Sun, Kangye Ji, Chen Tang, Jiajun Fan, Xinzhu Ma, Shutao Xia, Zhi Wang, and Wenwu Zhu. 2025. Sp-vla: A joint model scheduling and token pruning approach for vla model acceleration. *arXiv preprint arXiv:2506.12723*.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. 2023. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and 1 others. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, and 1 others. 2024. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE.

Seongmin Park, Hyungmin Kim, Wonseok Jeon, Juyoung Yang, Byeongwook Jeon, Yoonseon Oh, and Jungwook Choi. 2024. Quantization-aware imitation-learning for resource-efficient robotic control. *arXiv preprint arXiv:2412.01034*.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205.

Xudong Tan, Yaoxin Yang, Peng Ye, Jialin Zheng, Bizhe Bai, Xinyi Wang, Jia Hao, and Tao Chen. 2025. Think twice, act once: Token-aware compression and action reuse for efficient inference in vision-language-action models. *arXiv preprint arXiv:2505.21200*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Hanzhen Wang, Jiaming Xu, Jiayi Pan, Yongkang Zhou, and Guohao Dai. 2025a. Specprune-vla: Accelerating vision-language-action models via action-aware self-speculative pruning. *arXiv preprint arXiv:2509.05614*.

Songsheng Wang, Rucheng Yu, Zhihang Yuan, Chao Yu, Feng Gao, Yu Wang, and Derek F Wong. 2025b. Spec-vla: speculative decoding for vision-language-action models with relaxed acceptance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26916–26928.

Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Zhibin Tang, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, and 1 others. 2025a. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters*.

Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. 2025b. Stop looking for important tokens in multimodal language models: Duplication matters more. *arXiv preprint arXiv:2502.11494*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

Siyu Xu, Yunke Wang, Chenghao Xia, Dihao Zhu, Tao Huang, and Chang Xu. 2025. Vla-cache: Efficient vision-language-action manipulation via adaptive token caching. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yantai Yang, Yuhao Wang, Zichen Wen, Luo Zhongwei, Chang Zou, Zhipeng Zhang, Chuan Wen, and Linfeng Zhang. 2025. Efficientvla: Training-free acceleration and compression for vision-language-action models. *arXiv preprint arXiv:2506.10100*.

Yang Yue, Yulin Wang, Bingyi Kang, Yizeng Han, Shenzhi Wang, Shiji Song, Jiashi Feng, and Gao Huang. 2024. Deer-vla: Dynamic inference of multimodal large language models for efficient robot execution. *Advances in Neural Information Processing Systems*, 37:56619–56643.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. 2025. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20857–20867.

Qizhe Zhang, Aosong Cheng, Ming Lu, Zhiyong Zhuo, Minqi Wang, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. 2024. [cls] attention is all you need for training-free visual token pruning: Make vlm inference faster. *arXiv e-prints*, pages arXiv–2412.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, and 1 others. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR.

A TIES Algorithm

Algorithm 1 τ-guided Dynamic Token Pruning (TIES)

Require: Input sequence $I = \{i_1, i_2, \dots\}$, retained-token budget K_{ret}, VLA model $\mathcal{M}$, calibration size M, refresh horizon H, layer range L

Phase 1: Calibration (Offline)
Sample M calibration frames and compute their Kendall scores $\{\tau_i\}_{i=1}^{M}$
Estimate calibration quantiles q_{25}, q_{50}, and q_{75}

Phase 2: Dynamic Inference (Online)
Initialize $\tau_{curr} \leftarrow$ None
for each timestep $t = 1, 2, \dots$ **do**
  if τ_{curr} is None **or** $(t-1) \bmod H = 0$ **then**
    Compute τ_{curr} for frame i_t from the first L layers of $\mathcal{M}$
  end if
  if Hard-TIES **then**
    Set $y_t \leftarrow 1$ if $\tau_{curr} \leq q_{50}$, else $y_t \leftarrow 0$
  else
    Set $y_t \leftarrow \min\Big(1, \max\Big(0, \frac{q_{75}-\tau_{curr}}{q_{75}-q_{25}}\Big)\Big)$
  end if
  $K_{attn} \leftarrow \lfloor y_t K_{ret} \rfloor$, $K_{uni} \leftarrow K_{ret} - K_{attn}$
  Select $\mathcal{T}_{attn}$: the top K_{attn} visual tokens by attention score
  Select $\mathcal{T}_{uni}$: K_{uni} uniformly sampled tokens from the remaining visual tokens
  Retain $\mathcal{T}_{final} \leftarrow \mathcal{T}_{attn} \cup \mathcal{T}_{uni}$ and execute $\mathcal{M}$
end for