TransPrune:
Token Transition Pruning for Efficient Large Vision-Language Model

Supplementary Materials

1. Theoretical Explanation

We provide a theoretical explanation showing that token transitions serve as indicators of semantic information. Due to the residual formulation, Transformer layers are not forced to modify every token representation. During optimization, tokens with little marginal contribution to the loss receive negligible gradient-driven updates, causing their residual update terms to approach zero. So, such tokens primarily propagate along the identity path and exhibit small token transition values.

Why intermediate layers? The prominence of TTV in intermediate layers can be explained from Information Bottleneck (IB). Deep representations balance input compression and task-relevant information preservation: lower layers retain high-fidelity inputs, higher layers align with the prediction objective, and intermediate layers form the critical transition where irrelevant information is discarded and predictive features are consolidated.

2. Configuration of TransPrune

As shown in Table 1, we present the specific configurations of TransPrune-High and TransPrune-Low.

Method	LLaVA-v1.5 TFLOPs	LLaVA-Next TFLOPs	Retained Ratio	Final Token
TransPrune-High	1.56	8.33	[0.875, 0.625, 0.125]	72
TransPrune-Low	1.19	6.41	[0.625, 0.1875, 0.0625]	36

Table 1. Configurations of TransPrune High and TransPrune Low. The retained ratio array indicates the proportion of tokens to be kept at each pruning layer, relative to the original number of tokens.

3. Comparison of different cosine-based transition metrics

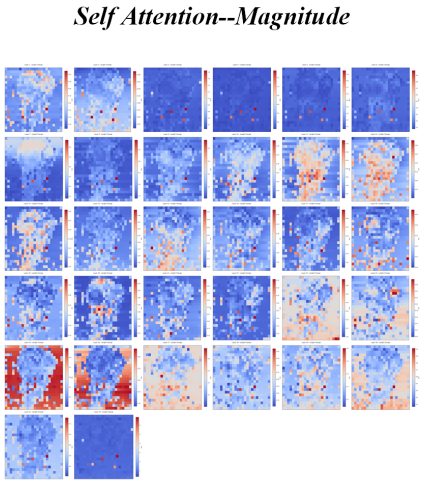
We experiment with several variants of cosine-based measures for directional change. As shown in Table 2, the performance of `cos` and `1-|cos|` is highly similar across benchmarks. We argue that both metrics capture comparable aspects of the angular deviation between token representations. While `cos` directly measures the signed angular similarity, `1-|cos|` transforms it into a measure of angular difference, emphasizing large directional shifts regardless of sign.

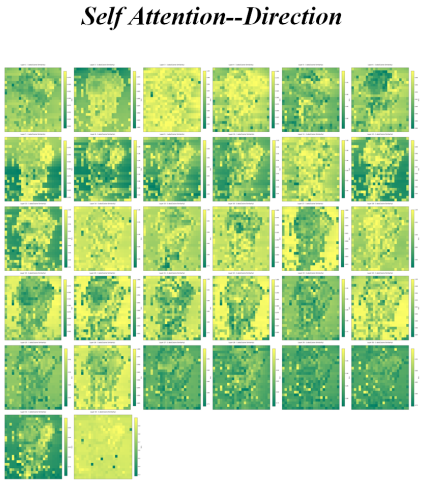
Methods	MME^P	SQA^I	GQA	MMB^{en}
cos	1539	69.5	61.4	66.1
$\|cos\|$	1519	69.6	61.4	65.5
1-cos	1513	69.6	61.4	65.6
1-$\|cos\|$	1540	69.5	61.4	66.0

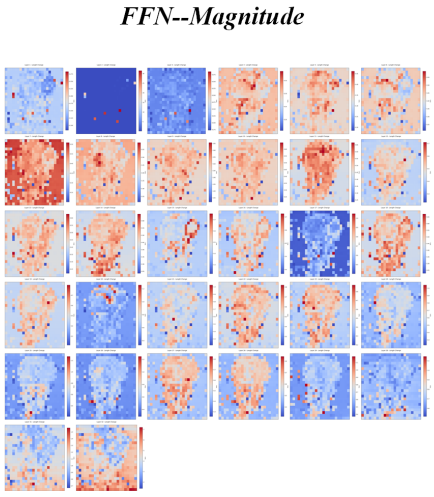
Table 2. Performance comparison among cosine-based variants for directional change measurement. `cos` and `1-|cos|` achieve almost identical results across benchmarks, indicating their similar sensitivity to directional deviation.

4. Visualization of Token Transition

As shown in Figure 1 and Figure 2, we provide additional visualization examples, showing that this phenomenon occurs not only in the 7B model but also in the 13B model.







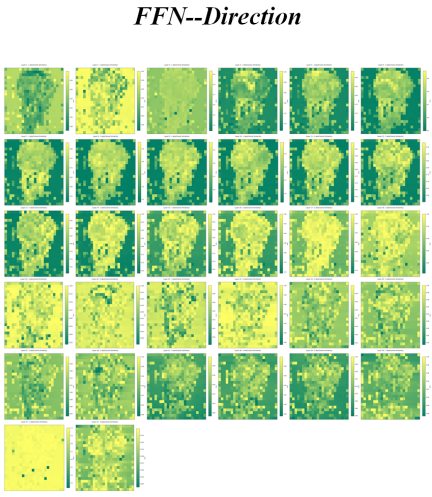
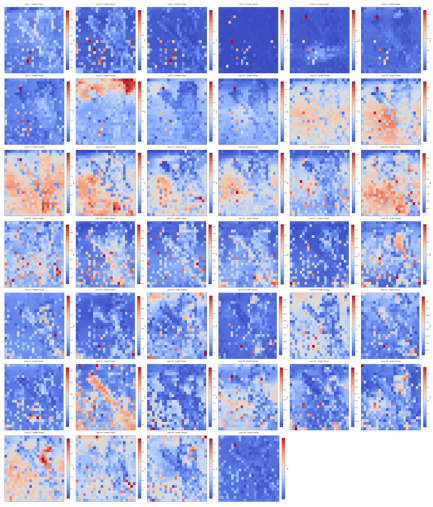
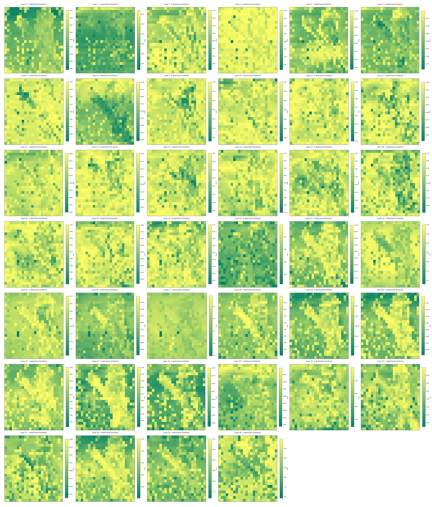


Figure 1. Visualization of token transition on LLaVA-v1.5-7B.

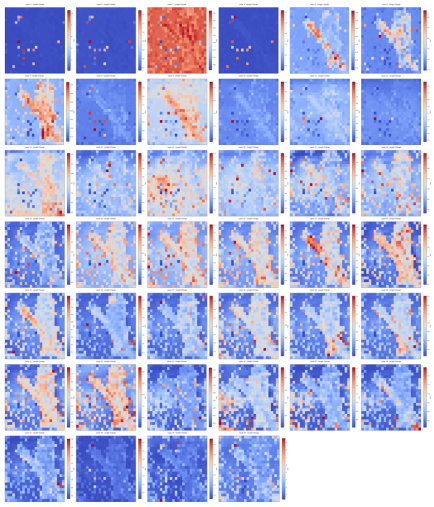
Self Attention--Magnitude



Self Attention--Magnitude



FFN--Magnitude



FFN--Direction

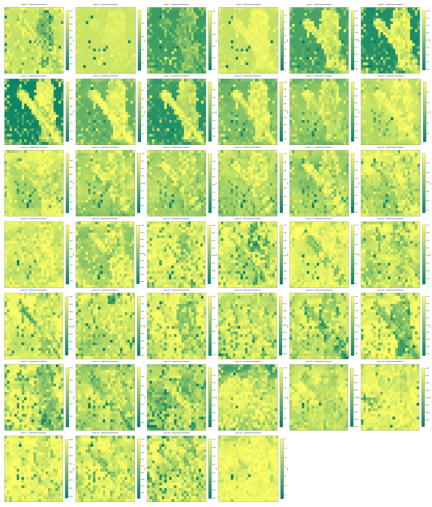


Figure 2. Visualization of token transition on LLaVA-v1.5-13B.