DynImg: Key Frames with Visual Prompts are Good Representation for Multi-Modal Video Understanding

Supplementary Material

1. More Ablations

Different composition of temporal prompts. We also explore various composition choices for DynImg. In the temporal prompts of Sec.3, non-keyframes are uniformly arranged in a line and overlaid below the base keyframe. Here, we investigate alternative overlay methods, as presented in Tab.A. The first four methods involve arranging temporal prompts in a row or column and overlaying them in different positions around the base keyframe. It is observed that varying the overlay direction yields similar performance results, with optimal performance achieved when temporal prompts are placed below. The last two methods utilize a sandwich-like format, where temporal prompts are separated into two lines. Their relative temporal sequence to the keyframe decides which side of the base keyframe they are placed in horizontal or vertical directions. It is found that employing this sandwich-style format significantly reduces performance. We speculate that this occurs because non-keyframe temporal prompts occupy a disproportionately larger area than that in the DynImg, overshadowing the base keyframe and affecting its resolution and spatial fine-grained representation.

Sequence	Top	Down	Left	Right	Vert	Horiz
Acc	77.9	**78.6**	78.0	78.4	75.2	74.9
Score	3.9	**4.2**	4.0	4.1	3.4	3.3

Table A. Ablation studies of different compositions of temporal prompts. "Top", "Down", "Left", and "Right" refer to stacking non-keyframes in a row or column and overlaying them on the direction of the keyframes. "Vert" and "Horiz" refer to a sandwich-like structure. Specifically, temporal prompts before and after the keyframes are overlaid to the keyframes separately, in a horizontal or vertical direction, according to the time order.

Generality over Different MLLMs. To demonstrate the broad applicability of our approach, we evaluate DynImg across various LLMs and vision encoders. As shown in Tab. B, our method consistently improves performance regardless of the underlying model architecture, indicating strong generalization capabilities.

We choose SigLip as the default vision encoder based on recent studies (e.g., LLaVA-OV, VideoLLaMA3), which have shown its superiority in fine-grained visual understanding and multi-modal alignment. This is also reflected in our experiments, where models using SigLip achieve higher accuracy and user scores compared to those with CLIP or ViT-L encoders.

Even with a strong baseline accuracy of 74.9 without the format of DynImg, incorporating DynImg leads to further improvements (+3.7). Notably, the results using Vicuna-based LLMs in Tab. B suggest that the gains primarily stem from DynImg itself rather than being attributed to a more advanced language model. These findings confirm that DynImg is effective across different MLLM backbones.

LLM	Qwen	Vicuna	Qwen	Qwen	Qwen
Visual	SigLip	Clip	Clip	Vit-L	SigLip
DynImg		✓	✓	✓	✓
Acc	74.9	77.9	78.3	78.2	78.6
Score	3.8	4.0	4.2	4.2	4.2

Table B. Ablation studies of different model choices. We conduct experiments to quantify how DynImg enhances performance across various LLM and visual encoder configurations, benchmarked against baselines without DynImg.

2. Visualization

Visualizations in Fig. A. shows how DynImg preserves attention for fast-moving objects in the visual feature extraction, which are neglected by the baseline.

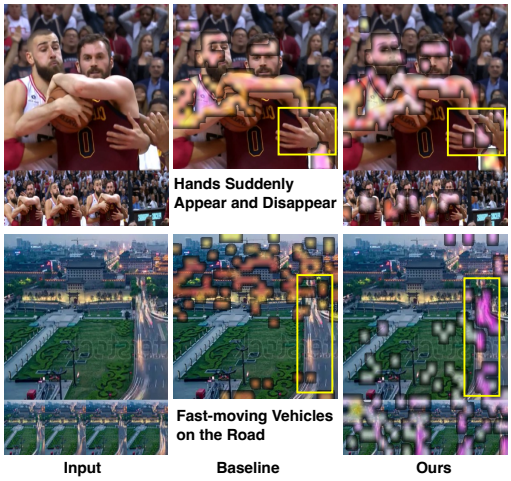

Figure A. PCA analysis on features of the visual encoder. The first column shows the generated DynImg images, the second column the baseline results (without DynImg), and the third column our DynImg results. In the first row, a hand appears in one frame. Previous methods ignore it during feature extraction, while our method retains attention to the region. In the second row, a fast-moving small object on the road is shown. The baseline fails to focus on it, whereas our method maintains proper attention.