VidPrism: Heterogeneous Mixture of Experts for Image-to-Video Transfer

Supplementary Material

6. Experiments and Further Analysis

6.1. Detailed Hyperparameters

In this section, we provide detailed training configurations on the Kinetics-400 dataset. The model is optimized over 40 epochs with a base learning rate of 1×10^{-4}. For the key hyperparameters of our proposed modules, the mixing score α in RgSTA is set to 0.7, and the gating threshold in DBI is configured to 0.6. Additionally, the loss weights for $\mathcal{L}_{ce}$, $\mathcal{L}_{rank}$, $\mathcal{L}_{div}$, and $\mathcal{L}_{gate}$ are set to 1, 0.1, 0.2, and 0.2.

6.2. Baseline MoE Experiment

Table 6. Performance comparison with the homogeneous MoE baseline (BMoE) on Kinetics-400.

Method	**Inputs**	**Top-1(%)**	**Top-5(%)**	**GFLOPs**
BMoE-B/16	8×4×3	80.3	95.0	159
BMoE-B/16	32×4×3	81.0	95.4	718
VidPrism-B/16	8×4×3	**84.0**	**96.8**	162
BMoE-L/14	8×4×3	84.2	96.6	629
BMoE-L/14	32×4×3	84.7	97.0	2590
VidPrism-L/14	8×4×3	**87.4**	**97.8**	632

To evaluate the necessity of expert heterogeneity, we compare VidPrism with a homogeneous MoE baseline (BMoE) under identical conditions. As detailed in Table 6, VidPrism-B/16 outperforms the equivalent BMoE model by a significant 3.7% in Top-1 accuracy. Even when BMoE is fed with much denser inputs (32×4×3, requiring ∼4.5× GFLOPs), it still lags behind our method by 3.0%. This superiority holds true on the ViT-L/14 backbone with a 3.2% margin. These results empirically validate that simply scaling computation or relying on homogeneous experts is highly suboptimal, and our heterogeneous design is key to capturing complex spatial-temporal dynamics.

6.3. Single Component Retention Experiment

To isolate the contributions of our modules, we tested activating RgSTA, DBI, or GlobalAttn individually. As shown in Table 7, solely enabling DBI or GlobalAttn only provides modest gains over the baseline ("None"), while applying RgSTA alone even causes a slight drop on HMDB-51. In stark contrast, integrating all modules within the full VidPrism architecture yields substantial Top-1 accuracy improvements of 1.7% on UCF-101 and 3.1% on HMDB-51. These results suggest that the effectiveness of our framework comes from their synergistic collaboration rather than a trivial accumulation of individual structures.

Table 7. Single-component retention analysis on UCF-101 and HMDB-51 datasets.

Type	**UCF-101**	**HMDB-51**
None	94.2	73.2
only RgSTA	94.3	73.0
only DBI	94.8	74.2
only GlobalAttn	94.7	73.5
VidPrism	**95.9**	**76.3**

Table 8. Cost analysis of our method. We report the number of parameters of the model, Top-1 accuracy on K400 and FPS during inference, and compare them with MoTE.

Method	Arch	Param(M)	Top-1(%)	FPS
MoTE [54]	ViT-B/16	98.8	83.0	338.93
	ViT-L/14	346.6	86.8	85.97
VidPrism	ViT-B/16	106.9	84.0	307.12
	ViT-L/14	345.9	87.4	86.62

6.4. Cost Analysis

We analyze the inference cost of our method in Table 8 and compare it with MoTE [54]. All measurements are conducted on a single RTX 3090 GPU with the number of input frames fixed to 8. Under the ViT-B/16 backbone, our method uses slightly more parameters than MoTE, which leads to a minor drop in FPS, but yields a 1.0% improvement in accuracy on K400, indicating a favorable accuracy–efficiency trade-off. When scaling up to ViT-L/14, our method achieves clear advantages: it uses fewer parameters, runs marginally faster, and still improves accuracy by 0.6%.

6.5. Few-shot Experiment with Larger Model

Building upon the configuration shown in Table 2, we further employed larger-scale models, including ViT-L/14 and VideoMAEv2-Large [37]. As detailed in Table 10, we benchmark our scaled variants against the baseline MoTE and the state-of-the-art method ETL[7]. The results indicate that increasing model capacity significantly boosts performance over both the baseline and the SOTA competitor. Specifically, VidPrism-L/14 demonstrates superior capability on HMDB51, where it surpasses ETL by a large margin in 4-shot, 8-shot, and 16-shot settings, achieving a peak accuracy of 77.5%. Furthermore, VidPrism-Large exhibits exceptional robustness on UCF101 and SSv2. On UCF101, it consistently leads across all shot counts and outperforms MoTE and ETL, reaching 98.0% accuracy. Notably, on the temporal-heavy SSv2 dataset, VidPrism-Large establishes a

new state-of-the-art by attaining 22.8% accuracy under the 16-shot protocol, substantially outperforming the previous best result achieved by ETL.

6.6. Variant of VidPrism

As shown in Figure 2, our VidPrism employs only the visual encoder of CLIP. The unified representation produced by the combination module is passed through a linear layer to obtain the final logits. In the variant, we incorporate CLIP's text encoder and compute the logits by measuring the similarity between the unified representation and the text embeddings. We adapt ViT-B/16 as the visual backbone and use 8 input frames. The model is trained on Kinetics-400, and we evaluate its few-shot performance on HMDB51 and UCF101. The results are presented in Table 11. The variant yields higher accuracy in few-shot regimes like HMDB51 K=2 by leveraging semantic priors. However, it lags behind VidPrism on the full Kinetics-400 dataset and 16-shot benchmarks. This indicates that relying on fixed text embeddings restricts the flexibility to optimize for specific temporal features as data availability increases.

6.7. Analysis of Impact on Number of Experts

Table 9. Expert isolation analysis on UCF-101 as a supplementary experiment to Table 3. We evaluate different combinations of expert scales by masking specific pathways during inference.

Retained experts	UCF-101	Δ
2,4	93.2	-1.4
4,8	94.1	-0.5
2,8	94.7	+0.1
2,4,8	**94.6**	-
4,8	93.8	-0.5
8,16	93.7	-0.6
4,16	94.4	+0.1
4,8,16	**94.3**	-
2,4	82.7	-11.4
4,16	93.9	-0.2
2,16	93.7	-0.4
2,4,16	**94.1**	-

Simply increasing the number of experts can introduce feature redundancy. To investigate this, we conducted isolation experiments by progressively masking specific expert outputs during inference, as shown in Table 9. When intermediate scales are dropped (2,8 or 4,16), accuracy slightly improves to 94.7% and 94.4%. This indicates that adjacent scales (like 4 and 8) may cause negative interference due to overlapping temporal granularities. Conversely, when omitting the slow-scale expert (16) from the {2,4,16} combination, performance reduces to 82.7%, exposing the model's critical reliance on distinct long-term context. These observations confirm that redundancy primarily arises from non-complementary temporal scales, while synergy across diverse resolutions is essential for model's capacity.

6.8. Further Explanation of Expert Heterogeneity

While content-aware sampling dictates input-level scale differences, the true heterogeneity of our experts is performed by the model's internal mechanisms. Specifically, RgSTA facilitates scale-based modeling to extract temporal representations, and DBI enables functional differentiation through selective bidirectional fusion. Furthermore, maintaining independent expert parameters allows the model to learn distinct representational spaces unconstrainted by resolution. Thus, the heterogeneity within VidPrism arises from the synergy of multi-rate inputs, interaction mechanisms, and experts. We also explored structural heterogeneity (e.g., using TimeSformer-style architectures for specific experts) but observed a $\sim$3% performance drop, likely due to optimization instability and temporal rate misalignment.

7. Additional Visualizations

7.1. Expert Usage Visualization on HMDB-51

We visualise expert usage on the HMDB51 dataset in Figure 5, separately aggregating training samples, correctly classified test samples, and misclassified samples. The activation patterns of the training set and correctly classified samples are highly consistent, indicating that the model learns a stable expert assignment strategy; moreover, different action categories exhibit distinct expert usage, evidencing a clear functional division of labour among heterogeneous experts. In contrast, the heatmaps of misclassified samples deviate from these patterns, for example, errors in category 40 are associated with over-activation of expert 0, suggesting that incorrect predictions often arise when the model fails to choose the most appropriate expert.

7.2. Visualization of Expert Specialization

To validate the effectiveness of our heterogeneous framework, we visualize the spatial saliency maps of experts operating at different temporal scales in Figure 6. These visualizations provide direct evidence that we have successfully addressed challenge 1 by preventing expert homogenization through specialized input provisioning. We observe a clear functional separation based on the input frame rate. Experts processing low-rate inputs such as the 2-Frames expert concentrate their attention heavily on the main actors and static semantic details, which is evident in the golfer's posture in Figure 6 (a). In contrast, experts handling high-rate inputs like the 16-Frames expert exhibit broader attention distributions that capture motion trajectories and environmental changes. This is exemplified by the dynamic waves in the Surfing sequence in Figure 6 (b). This confirms that our method effectively differentiates expert roles and ensures they learn complementary spatial and temporal features.

Table 10. Further Few-shot Experment.

Method	HMDB51				UCF101				SSv2			
	K=2	K=4	K=8	K=16	K=2	K=4	K=8	K=16	K=2	K=4	K=8	K=16
MoTE [54]	**61.3**	63.9	67.2	68.2	88.8	91.0	92.3	93.6	7.3	8.5	9.5	12.2
ETL [7]	61.2	62.3	67.1	70.4	86.1	90.2	92.7	94.8	**9.1**	10.4	13.4	17.5
VidPrism-B/16	55.0	64.1	67.6	74.1	88.1	91.6	93.8	96.0	4.9	7.3	9.3	14.1
VidPrism-Base	53.4	63.3	68.3	73.2	94.3	95.6	95.9	96.6	6.4	9.5	13.6	18.3
VidPrism-L/14	53.4	**65.6**	**72.6**	**77.5**	90.4	95.7	96.8	97.9	5.5	9.4	10.6	16.2
VidPrism-Large	55.6	64.8	70.2	74.7	**95.9**	**96.8**	**97.1**	**98.0**	7.7	**11.3**	**15.8**	**22.8**

Table 11. Research on variants of the clip text encoder.

Method	K400	HMDB51				UCF101			
		K=2	K=4	K=8	K=16	K=2	K=4	K=8	K=16
Varient	83.2	60.2	64.6	67.3	70.0	87.3	92.2	94.7	95.6
VidPrism	84.0	55.0	64.1	67.6	74.1	88.1	91.6	93.8	96.0

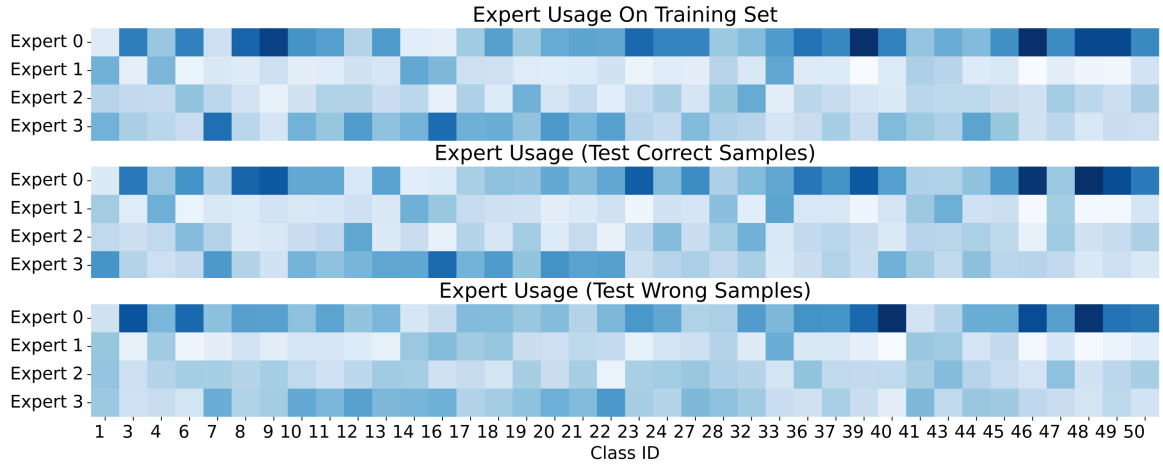
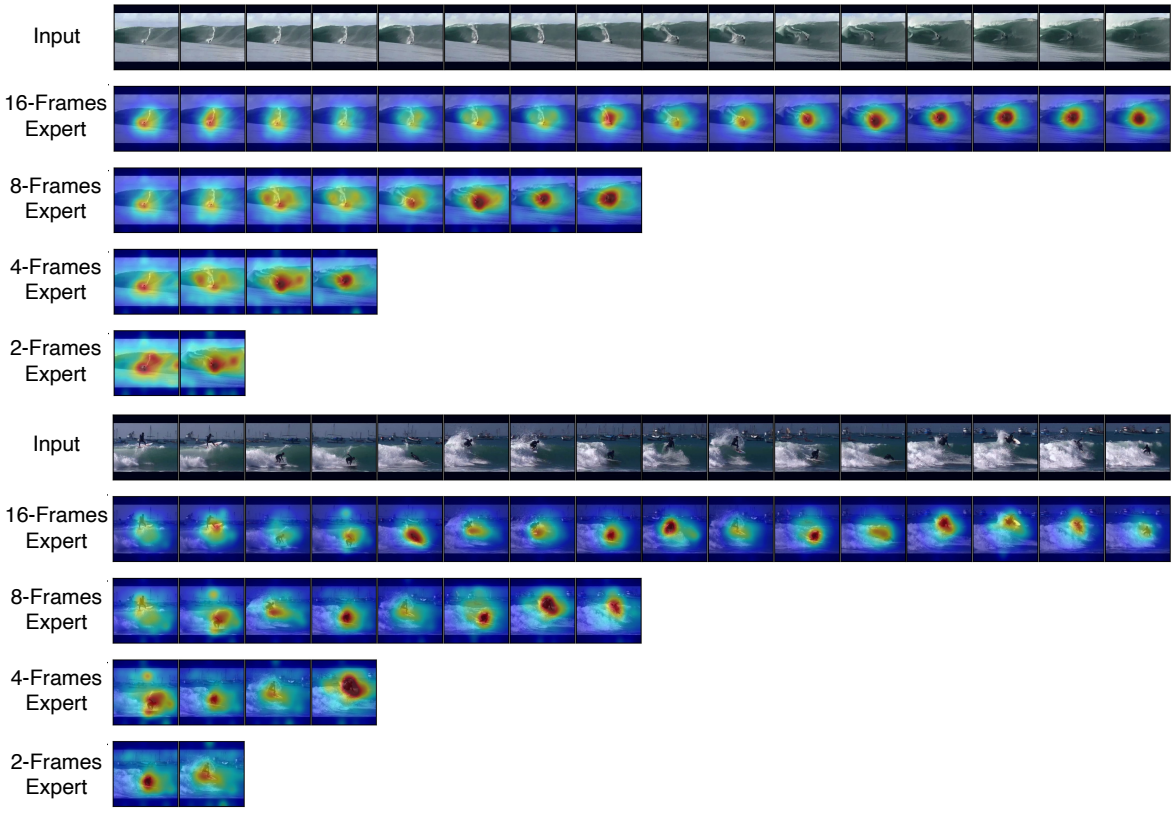


Figure 5. Visualization of Expert Usage on Training and Test Sets.

(a) Saliency Map for GolfSwing.



(b) Saliency Map for Surfing.

Figure 6. Visualization of Saliency Maps.