# Make Your Training Flexible: Towards Deployment-Efficient Video Models


上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

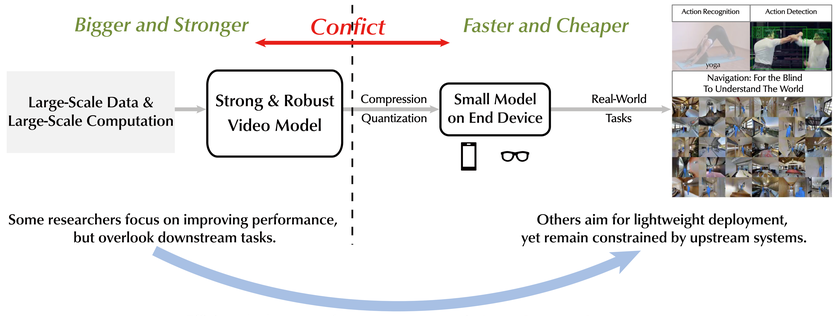
Cheneting Wang[1,2] Kunchang Li[2] Tianxiang Jiang[2,4] Xiangyu Zeng[2,5] Yi Wang[2,3] Limin Wang[1,2,5]

[1]Shanghai Jiao Tong University, [2]Shanghai AI Laboratory, [3]Shanghai Innovation Institute, [4]University of Science and Technology of China, [5]State Key Laboratory for Novel Software Technology, Nanjing University

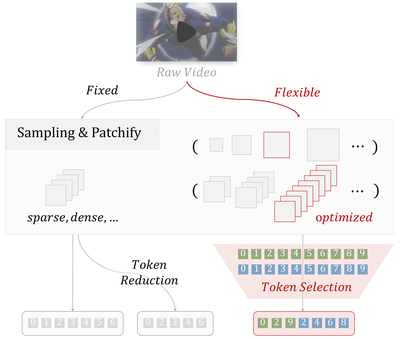


ICCV OCT 19-23, 2025 HONOLULU HAWAII


## Video Model's Deployment

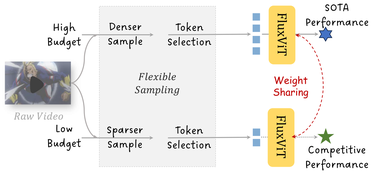



## FluxViT: Train Once, Deploy Anywhere

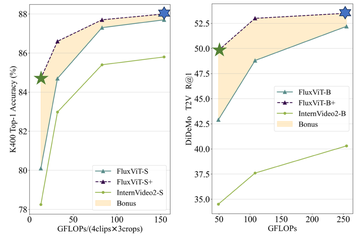



Common methods (left): Use rigid sampling and directly use token reduction for deployment.

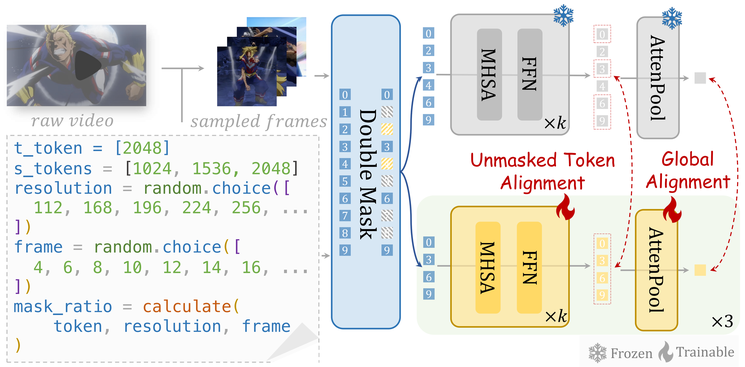
Flux (right): Employs flexible sampling and token selection for Token Optimization.




FluxViT deployed across varying computational budgets.




## Training Method




Flux integrated with the Unmasked Teacher framework.

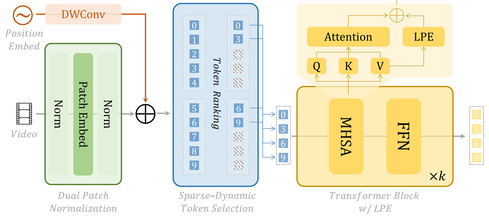



Illustration of the proposed Flux-ViT module.

| Method & Arch | K400 | *w/* TO | SSv2 | *w/* TO |
|---|---|---|---|---|
| Vanilla + ViT | 78.4 | 78.4 | 65.4 | 65.4 |
| Vanilla + FluxViT | 79.3 | 79.6 (↑1.2) | 66.0 | 66.4 (↑1.0) |
| Flux-Single + ViT | 79.2 | 80.3 (↑1.9) | 66.3 | 67.3 (↑1.9) |
| *With new positional embeddings* | | | | |
| *w/* RoPE | 79.5 | 80.7 (↑0.4) | 66.5 | 67.5 (↑0.2) |
| *w/* GPE | 79.4 | 80.5 (↑0.2) | 66.4 | 67.4 (↑0.1) |
| *w/* LPE | 79.7 | 81.0 (↑0.7) | 66.8 | 68.3 (↑1.0) |
| *w/* GLPE | 79.9 | 81.3 (↑1.0) | 67.0 | 68.6 (↑1.3) |
| *With DPN* | | | | |
| *w/* DPN | 79.8 | 81.2 (↑0.9) | 66.9 | 68.4 (↑1.1) |
| *Combining the two modules* | | | | |
| Flux-Single + FluxViT | 80.5 | 81.7 (↑3.3) | 67.6 | 69.3 (↑3.9) |
| Flux-Multi + FluxViT | 81.4 | 82.4 (↑4.0) | 68.4 | 70.0 (↑4.6) |

Ablation Study.

## Performance

| Model | Extra Data | #P(M) | GFLOPs | Top-1 |
|---|---|---|---|---|
| TimeSformer-L [5] | - | 121 | 2380×3 | 80.7 |
| VideoSwin-L [50] | IN-21K | 197 | 604×12 | 83.1 |
| VideoMAE-L [66] | - | 305 | 3958×21 | 86.1 |
| CoVeR-L [87] | JFT-3B+SMI | 431 | 5860×3 | 87.1 |
| UniFormerV2-L [39] | CLIP-400M+K710 | 354 | 12550×6 | 90.0 |
| UMT-L [42] | K710 | 431 | 5860×3 | 90.6 |
| VideoMAE2-H [72] | UnlabeledHybrid | 633 | 1192×15 | 88.6 |
| ViViT-H [2] | JFT-300M | 654 | 3981×12 | 84.9 |
| MTV-H [82] | IN-21K+WTS-60M | 1000+ | 6130×12 | 89.9 |
| CoCa-G [85] | JFT-3B+ALIGN-1.8B | 1000+ | N/A×12 | 88.9 |
| MViTv1-B [19] | - | 37 | 70×5 | 80.2 |
| MViTv2-B [44] | - | 37 | 255×5 | 81.2 |
| ST-MAE-B [21] | K600 | 87 | 180×21 | 81.3 |
| VideoMAE-B [66] | - | 87 | 180×15 | 81.5 |
| VideoSwin-B [50] | IN-21k | 88 | 282×12 | 82.7 |
| UniFormer-B [40] | IN-1k | 50 | 259×12 | 83.0 |
| UMT-B [42] | K710 | 87 | 180×12 | 87.4 |
| InternVideo2-B [77] | K710+MASH | 96 | 440×12 | 88.4 |
| **FluxViT-B**$_{c200}$ | - | 97 | 440×12<br>49×12 | 88.7 89.4<br>84.0 86.7 |
| **FluxViT-B**$_{c100}$ | K710+MASH | 97 | 440×12<br>255×12<br>108×12<br>49×12 | 89.6 90.0<br>89.3 89.7<br>87.3 88.9<br>84.7 87.4 |
| UniFormer-S [40] | IN-1k | 21 | 42×4 | 80.8 |
| MViTv2-S [44] | - | 35 | 64×5 | 81.0 |
| VideoMAE-S [66] | - | 22 | 57×15 | 79.0 |
| VideoMAE2-S [72] | - | 22 | 57×15 | 83.7 |
| InternVideo2-S [77] | K710+MASH | 23 | 154×12 | 85.8 |
| **FluxViT-S**$_{c200}$ | - | 24 | 154×12<br>13×12 | 86.4 87.3<br>79.7 84.0 |
| **FluxViT-S**$_{c100}$ | K710+MASH | 24 | 154×12<br>83×12<br>32×12<br>13×12 | 87.7 88.0<br>87.3 87.7<br>84.7 86.6<br>80.1 84.7 |

Comparison on scene-related task: Kinetics-400.

| Model | Extra Data | GFLOPs | Top-1 | Top-5 |
|---|---|---|---|---|
| TimeSformer-L [5] | IN-21k | 2380×3 | 62.3 | - |
| MViTv1-B [19] | K400 | 455×3 | 67.7 | 70.9 |
| MViTv2-B [44] | K400 | 255×3 | 70.5 | 92.7 |
| VideoMAE-B [66] | K400 | 180×6 | 69.7 | 92.3 |
| VideoMAE-L [66] | K400 | 596×6 | 74.0 | 94.6 |
| UniFormerV2-B [39] | CLIP-400M | 375×3 | 70.7 | 93.2 |
| UniFormerV2-L [39] | CLIP-400M | 1718×3 | 73.0 | 94.5 |
| UMT-B [42] | K710 | 180×6 | 70.8 | 92.4 |
| InternVideo2-B [77] | K710+MASH | 253×6 | 73.5 | 94.4 |
| **FluxViT-B** | K710+MASH | 440×6<br>255×6<br>108×6<br>49×6 | 75.3 75.6<br>75.1 75.5<br>72.0 75.1<br>56.8 73.9 | 95.1 95.1<br>94.9 95.1<br>93.3 94.8<br>84.8 94.4 |
| UniFormer-S [40] | IN-1K | 42×3 | 67.7 | 91.4 |
| VideoMAE-S [66] | K600 | 57×6 | 66.8 | 90.3 |
| InternVideo2-S [77] | K710+MASH | 83×6 | 71.5 | 93.4 |
| **FluxViT-S** | K710+MASH | 154×6<br>83×6<br>32×6<br>13×6 | 73.4 73.8<br>72.9 73.4<br>70.0 72.5<br>55.3 70.9 | 94.1 94.1<br>94.0 94.1<br>93.4 93.8<br>83.7 93.1 |

Motion-intensive task: SSv2.

| Model | *e2e* | BackBone | Top-1 |
|---|---|---|---|
| Distant Supervision [47] | ✗ | TimeSformer | 90.0 |
| ViS4mer [32] | ✗ | Swin-B | 88.4 |
| Turbo$_{f32}$ [25] | ✓ | VideoMAE-B | 87.5 |
| VideoMamba$_{f64}$ [43] | ✓ | VideoMamba-S | 88.7 |
| VideoMamba$_{f64}$ [43] | ✓ | VideoMamba-M | 90.4 |
| InternVideo2$_{f12}$ [77] | ✓ | InternVideo2-S | 90.0 |
| MA-LMM [26] | ✓ | MLLM | 93.2 |
| HERMES [33] | ✓ | MLLM | 93.5 |
| **FluxViT**$_{3072}$ | ✓ | FluxViT-S | 91.8 92.1 |
| **FluxViT**$_{2048}$ | ✓ | FluxViT-S | 91.5 91.9 |
| **FluxViT**$_{1024}$ | ✓ | FluxViT-S | 89.8 91.0 |
| **FluxViT**$_{3072}$ | ✓ | FluxViT-B | 93.9 94.1 |
| **FluxViT**$_{2048}$ | ✓ | FluxViT-B | 93.7 93.9 |
| **FluxViT**$_{1024}$ | ✓ | FluxViT-B | 92.5 93.2 |

Video classification task: COIN.

| Model | MSR | DDM | ANet | LSMDC | MSVD |
|---|---|---|---|---|---|
| InternVideo2-S$_{2048}$ [77] | 35.6 | 33.7 | 34.5 | 14.7 | 41.8 |
| Frozen-B [3] | 18.7 | 20.2 | - | - | - |
| VIOLET-B [22] | 25.9 | 23.5 | - | - | - |
| Singularity-B [37] | 34.0 | 37.1 | 30.6 | - | - |
| OmniVL-B [69] | 34.6 | 33.3 | - | - | - |
| CLIP4Clip-B [53] | 30.6 | - | - | 13.6 | 36.2 |
| UMT-B [42] | 35.2 | 41.2 | 35.5 | 19.1 | 42.3 |
| Internvideo2-B$_{2048}$ [77] | 40.3 | 40.3 | 41.5 | 18.7 | 49.1 |
| VINDLU-L [14] | 32.0 | 36.9 | 30.9 | - | - |
| InternVideo-L [75] | 40.7 | 31.5 | 30.7 | 17.6 | 43.4 |
| UMT-L [42] | 40.7 | 48.6 | 41.9 | 24.9 | 49.0 |
| ViClip-L [76] | 42.4 | 18.4 | 15.1 | 20.1 | 49.1 |
| InternVideo2-L [77] | 42.1 | 42.8 | 43.6 | 21.4 | - |
| LanguageBind-L [91] | 42.8 | 39.7 | 38.4 | - | 54.1 |
| LanguageBind-H [91] | 44.8 | 39.9 | 41.0 | - | 53.9 |
| VideoCoCa-G [83] | 34.3 | - | 34.5 | - | - |
| VideoPrism-G [83] | 39.7 | - | 52.7 | - | - |
| VAST-G [13] | 49.3 | 55.5 | - | - | - |
| **FluxViT-S**$_{2048}$ | 44.4<br>**45.0** | 48.3<br>**49.3** | 52.4<br>**52.4** | 20.8<br>**22.4** | 49.4<br>**49.7** |
| **FluxViT-S**$_{1024}$ | 42.2<br>**44.5** | 45.4<br>**49.0** | 47.2<br>**50.3** | 18.7<br>**20.5** | 48.1<br>**48.5** |
| **FluxViT-S**$_{512}$ | 36.8<br>**40.5** | 38.5<br>**45.8** | 38.2<br>**44.7** | 17.7<br>**19.0** | 45.5<br>**46.9** |
| **FluxViT-B**$_{2048}$ | 49.8<br>**49.9** | 52.2<br>**53.5** | 56.6<br>**56.7** | 23.7<br>**25.4** | 53.8<br>**54.2** |
| **FluxViT-B**$_{1024}$ | 48.0<br>**49.1** | 48.8<br>**53.0** | 51.8<br>**54.8** | 22.6<br>**24.1** | 52.8<br>**53.4** |
| **FluxViT-B**$_{512}$ | 42.6<br>**47.2** | 42.9<br>**49.8** | 42.8<br>**50.3** | 20.1<br>**22.8** | 50.7<br>**52.1** |

Zero-shot text-to-video retrieval.

| Encoder | #Tokens | *w/* TO | MVbench | Dream1k-FT |
|---|---|---|---|---|
| Clip-L [60] | 8×256 | ✗ | 45.6 | 28.4 |
| SigLIP$_{336}$-L [86] | 8×576 | ✗ | 46.7 | 29.2 |
| InternVideo2-L [77] | 8×224 | ✗ | 47.0 | 28.7 |
| SigLIP$_{336}$-L [86] | 4×576 | ✗ | 44.5 | 25.4 |
| UMT-L [42] | 4×256 | ✗ | 45.0 | 24.6 |
| **FluxViT-L** | 8×256<br>4×256<br>2×256 | ✗ | 48.3<br>46.9<br>46.0 | 29.0<br>27.9<br>25.6 |
| **FluxViT-L** | 2048<br>1024<br>512 | ✓ | 49.0 (↑0.7)<br>47.7 (↑0.8)<br>47.6 (↑1.6) | 29.5 (↑0.5)<br>28.5 (↑0.6)<br>27.5 (↑1.9) |

MLLM benchmarks.

## Key Takeaways

- **Train Once, Deploy Anywhere**: Optimize your input first at deployment! Flexible training sampling empowers a single video model to adapt seamlessly across diverse computational budgets and downstream scenarios via efficient Token Optimization.
- **New Augmentation for Teacher–Student Pretraining:** Flexible teacher sampling enriches feature diversity and training efficiency at no extra cost, boosting downstream generalization.
- **Open-Source and Efficient:** FluxViT achieves SOTA performance across a wide range of video understanding tasks with dramatically reduced computational demand.


wangcheting@pjlab.org.cn