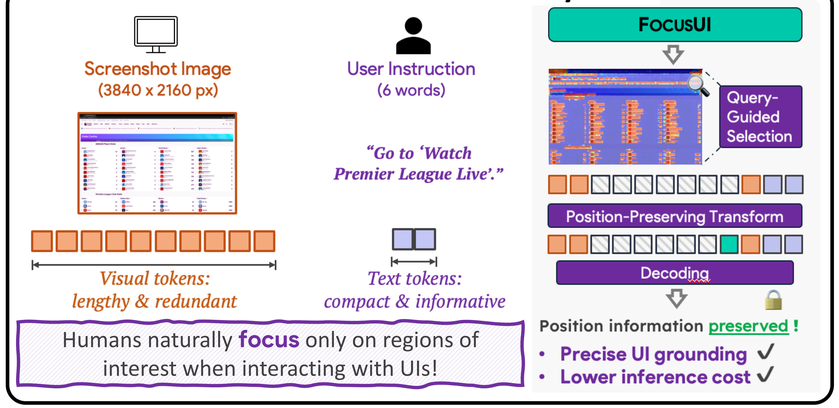
FocusUI: Efficient UI Grounding via Position-Preserving Visual Token Selection

Mingyu Ouyang[1], Kevin Qinghong Lin[2], Mike Zheng Shou[1], Hwee Tou Ng[1]

NUS National University of Singapore · University of Oxford

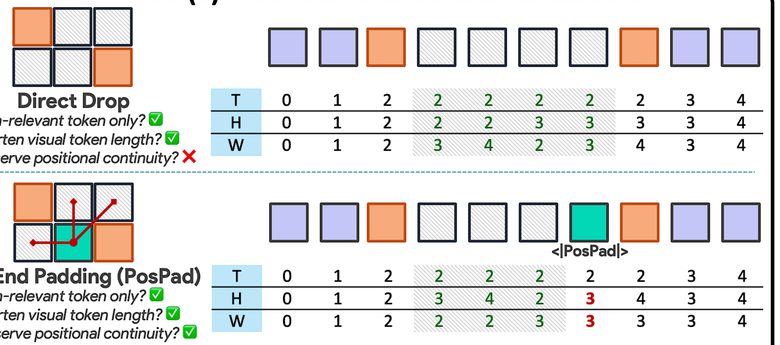
CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivation: Visual Redundancy in UI



Humans naturally **focus** only on regions of interest when interacting with UIs!

Method (c): PosPad Token Transformation



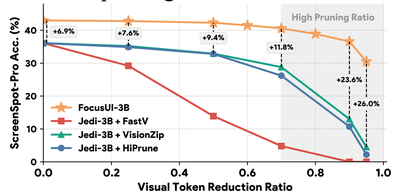
- A learnable token is placed at the last index of each contiguous sequence of dropped visual tokens, marking the removed spatial gaps.

Quantitative Results

- UI grounding benchmarks: Near-dense accuracy with about 30% of visual tokens

Model	ScreenSpot-V2							ScreenSpot-Pro							
	Mob.-T	Mob.-I	Des.-T	Des.-I	Web-T	Web-I	Avg	Dev	Cre.	CAD	Sci.	Office	OS	Avg-T	Avg-I	Avg
UI-TARS-1.5-7B [23]	-	-	-	-	-	-	90.0	31.8	40.2	31.8	47.2	65.6	33.2	-	-	42.6
Qwen2.5-VL-3B [3]	93.4	73.5	88.1	58.6	88.0	71.4	80.9	21.4	25.8	18.4	29.5	40.9	20.4	37.8	6.6	25.9
Qwen2.5-VL-7B [3]	97.6	87.2	90.2	74.2	93.2	81.3	88.8	29.1	24.9	13.8	31.1	45.7	22.4	39.9	7.6	27.6
Qwen2.5-VL-32B [3]	97.9	88.2	98.5	79.3	91.2	86.2	91.3	48.5	41.1	32.6	57.1	67.4	42.3	63.2	22.5	47.6
GUI-Actor-3B [30]	97.6	83.4	96.9	83.6	94.0	85.7	91.0	39.8	36.7	34.1	49.6	61.3	35.2	-	-	42.2
GUI-Actor-7B [30]	97.6	88.2	96.9	85.7	93.2	86.7	92.1	38.1	41.4	38.3	50.8	63.0	38.8	-	-	44.6
Jedi-3B [33]	96.6	81.5	96.9	78.6	88.5	83.7	88.6	38.1	34.6	23	38.6	57.0	25.0	49.8	13.7	36.1
Jedi-7B [33]	96.9	87.2	95.9	87.9	94.4	84.2	91.7	27.4	34	32.2	52.4	68.7	26.0	52.6	18.2	39.5
FocusUI-3B ($r = 100\%$)	99.2	85.9	96.1	87.3	95.4	81.9	91.5	43.1	37.0	37.6	48.4	61.7	38.3	59.3	18.9	43.8
FocusUI-3B ($r = 50\%$)	98.8	86.9	95.0	87.3	95.4	81.9	91.4	42.1	37.0	36.4	46.9	58.3	35.2	56.7	19.0	42.3
FocusUI-3B ($r = 30\%$)	98.5	85.3	96.1	87.3	94.3	81.9	91.0	38.1	35.8	33.3	44.5	57.8	37.2	55.0	17.4	40.6
FocusUI-7B ($r = 100\%$)	98.8	91.6	95.6	92.1	95.0	84.4	93.1	44.5	41.1	42.9	52.0	69.6	44.4	64.7	21.9	48.3
FocusUI-7B ($r = 50\%$)	98.8	92.2	93.9	87.3	95.0	85.2	92.6	42.8	40.5	40.2	51.6	67.0	40.3	61.7	21.9	46.5
FocusUI-7B ($r = 30\%$)	98.8	90.1	93.3	85.7	93.9	85.2	91.8	38.8	39.9	42.9	49.2	64.4	38.8	60.4	20.4	45.1

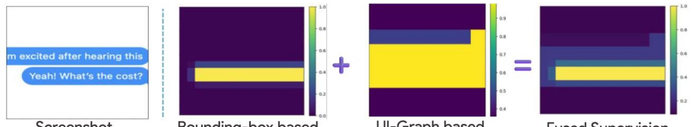
- Comparison with general visual token pruning methods + UI models



- Efficiency gain: 1.44x speedup with 30% retained visual tokens

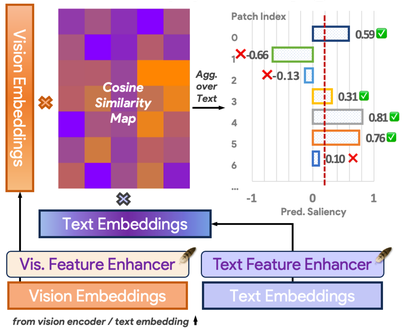
Model	%Ret. Ratio	#Vis. Token	per Sample Time (sec)	Max GPU Mem. (MB)	SS-Pro Acc
*Base Model: Qwen2.5-VL, max_pixel = 6400 * 28 * 28 = 4816000*					
FocusUI-7B	100%	5319	1.75 (1.00×)	20994 (1.00×)	48.3
FocusUI-7B	70%	3989	1.67 (1.05×)	18334 (0.87×)	47.7
FocusUI-7B	50%	2659	1.49 (1.18×)	17944 (0.85×)	46.5
FocusUI-7B	30%	1329	1.22 (1.44×)	17392 (0.83×)	45.1
*Base Model: Qwen3-VL, max_pixel = 6000 * 32 * 32 = 6144000*					
FocusUI-Qwen3-VL-2B	100%	4627	0.97 (1.00×)	6278 (1.00×)	39.8
FocusUI-Qwen3-VL-2B	70%	3470	0.90 (1.08×)	6142 (0.98×)	40.1
FocusUI-Qwen3-VL-2B	50%	2313	0.85 (1.14×)	5680 (0.91×)	40.4
FocusUI-Qwen3-VL-2B	30%	1156	0.71 (1.37×)	5170 (0.82×)	38.5

Method (a): Build per-patch Saliency Labels



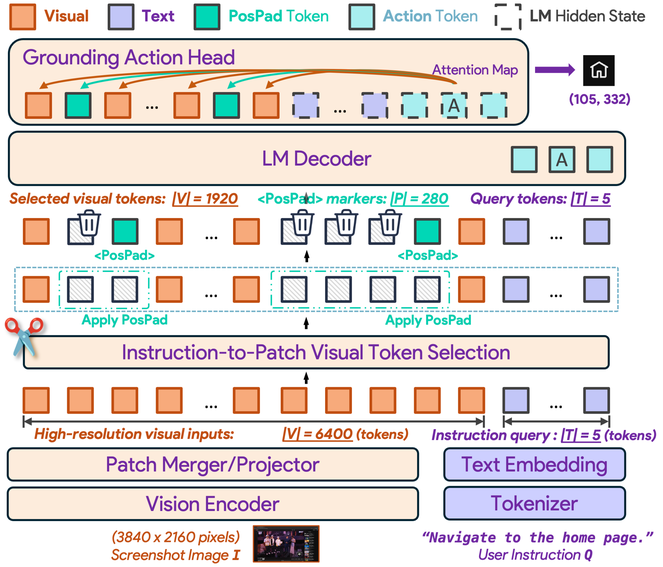
- Bounding-box based score identifies the region of target element from labelled UI grounding data (ground-truth bounding-boxes).
- UI-Graph based score suppresses redundant background by assigning higher importance to distinct patches.

Method (b): Query-guided Saliency Scorer



- A lightweight module is trained to predict the saliency for each visual patch given a query.
- It uses similarities between patch feature and query text embeddings in the VLM backbone.

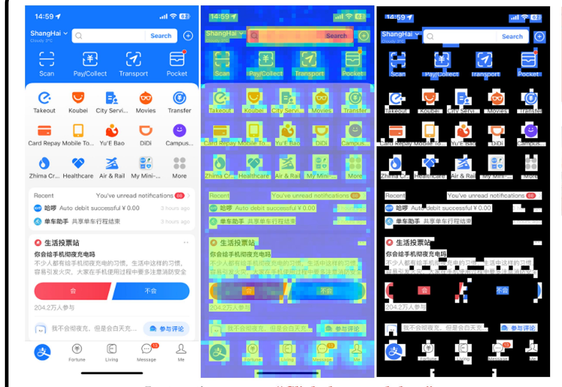
Method (d): Plug into VLM Inference

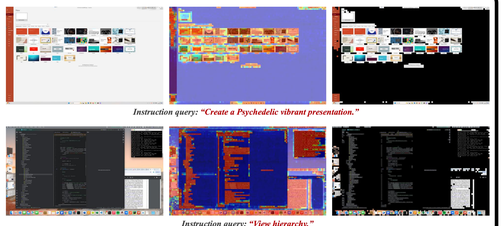


- FocusUI reduces the visual sequence before it is passed to the language model, which receives fewer visual tokens plus PosPad tokens for dropped regions.

The project page, code, and model are available!

Qualitative Results





- Original screenshot, predicted saliency, and selected visual tokens at retention rate = 30%.
- The performance gain comes from both selecting instruction-relevant UI patches and preserving their positional structure.