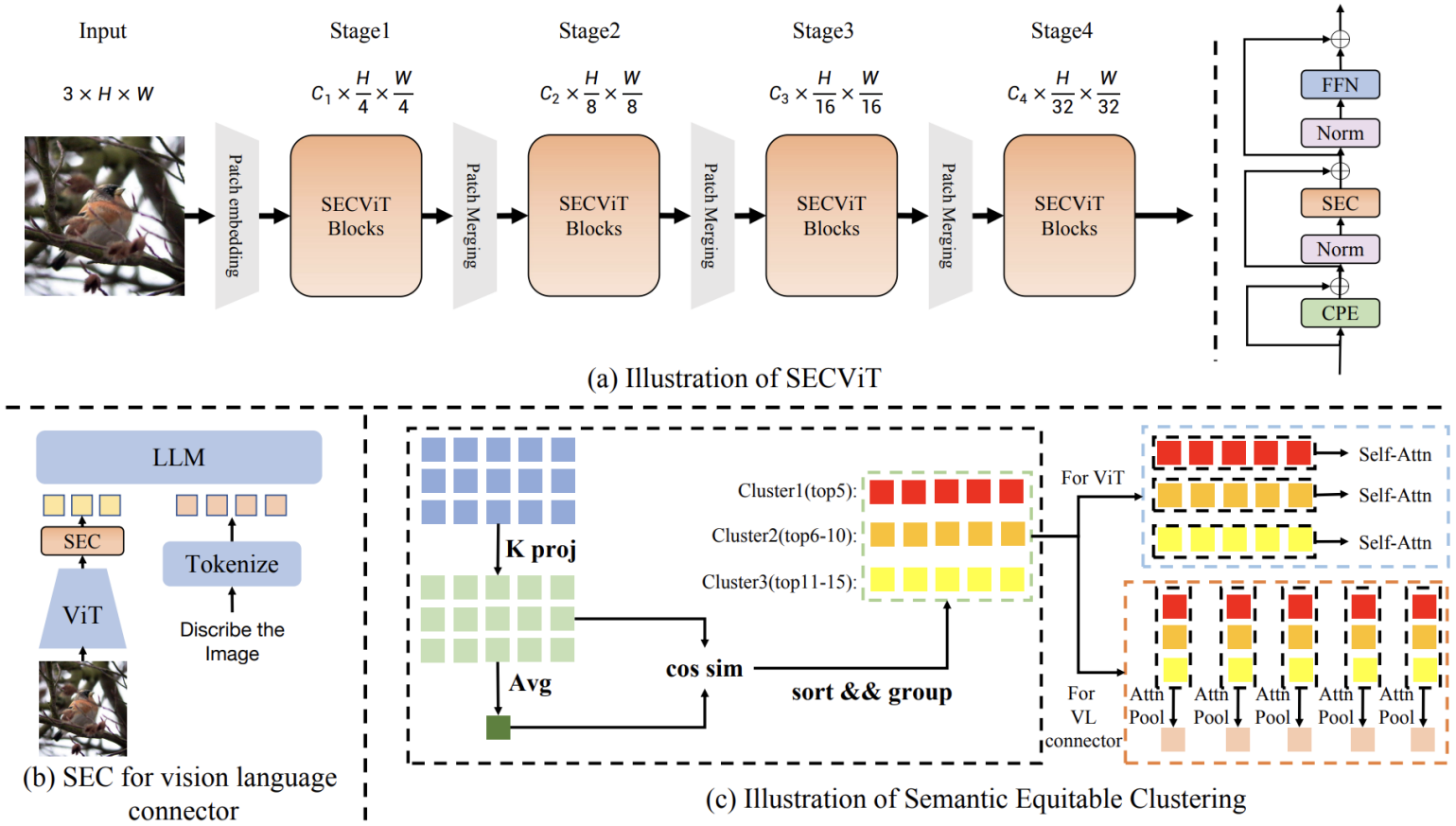
Semantic Equitable Clustering: A Simple and Effective Strategy for Clustering Vision Tokens

Qihang Fan

Institute of Automation, Chinese Academy of Sciences

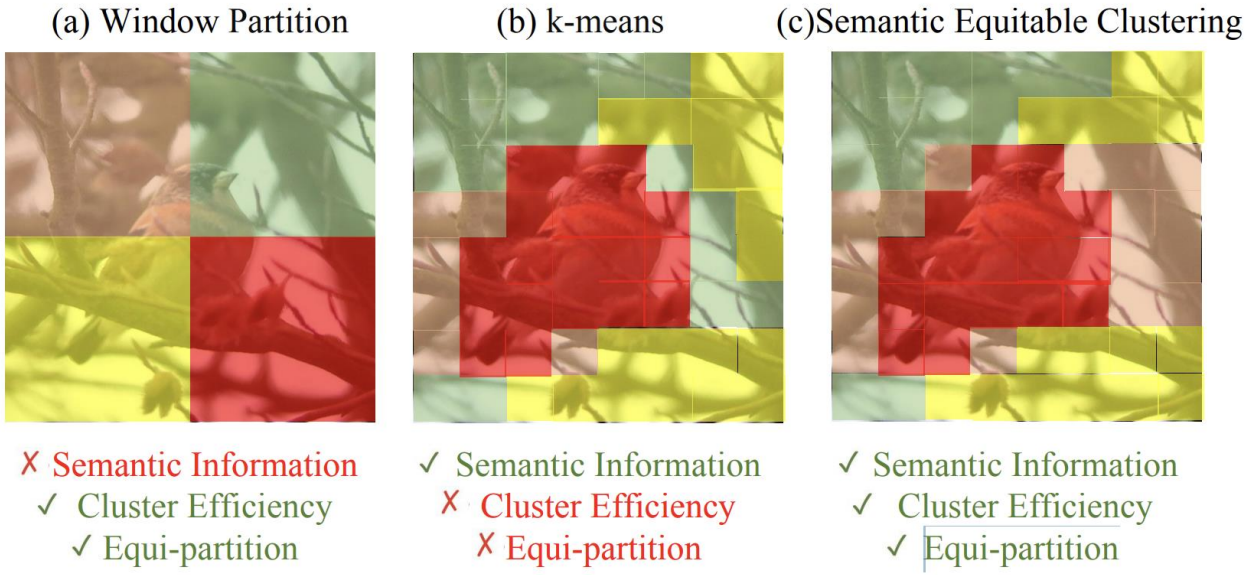
[ICCV'25] *Semantic Equitable Clustering: A Simple and Effective Strategy for Clustering Vision Tokens*

- **TL;DR:** An One-Step/Equi-partition Clustering Strategy for Vision Tokens for Vision and MLLM.



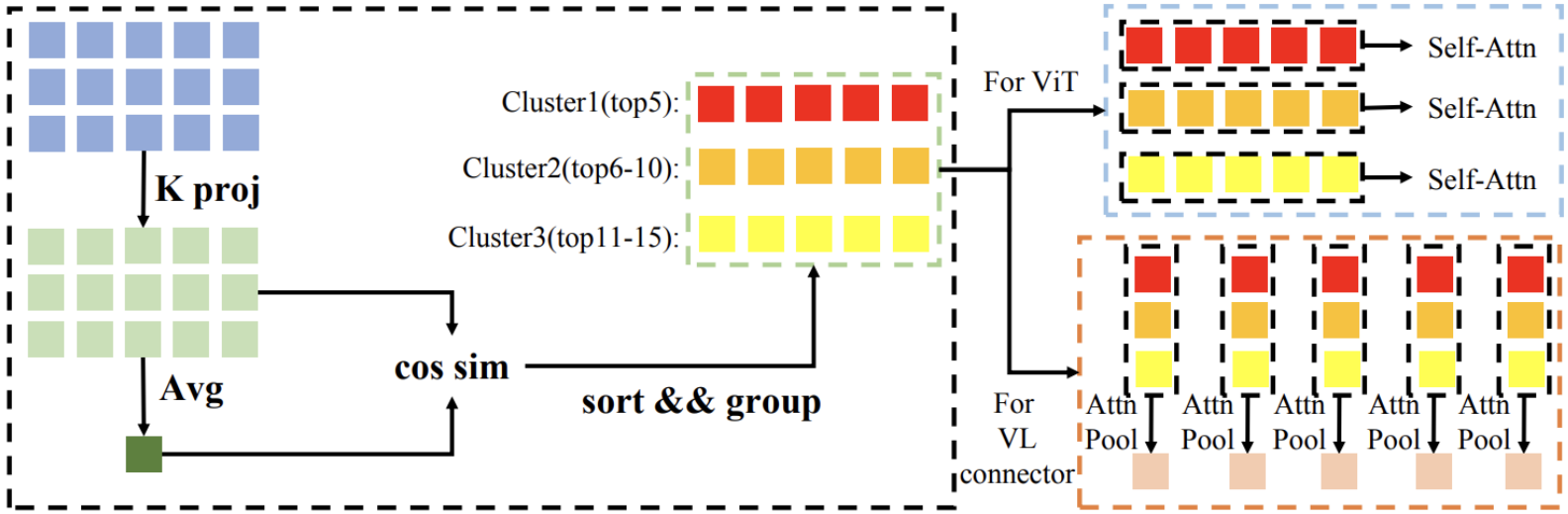
[ICCV'25] *Semantic Equitable Clustering: A Simple and Effective Strategy for Clustering Vision Tokens*

- **Motivation:**
  - Slide window partition ignores the semantic information.
  - Iterative methods such as k-means are too slow.
  - **We need a fatser and better one.**



[ICCV'25] *Semantic Equitable Clustering: A Simple and Effective Strategy for Clustering Vision Tokens*

- **Method: Utilize the Transformer's characatic.**



$$sim = \frac{K \cdot k_c}{||K|| \cdot ||k_c||},$$
$$idx = \text{argsort}(sim),$$
$$Q^* = Q[idx], K^* = K[idx], V^* = V[idx].$$

$$Q_m = Q^*[m \times N : (m+1)\,N],$$
$$K_m = K^*[m \times N : (m+1)\,N],$$
$$V_m = V^*[m \times N : (m+1)\,N].$$
$$Y_m = \text{Attn}\left(Q_m, K_m, V_m\right).$$

[ICCV'25] *Semantic Equitable Clustering: A Simple and Effective Strategy for Clustering Vision Tokens*

- **MLLM**

Model	Connector	V-T Num	Time	Speed	TextVQA	GQA	VQAv2
LLaVA-1.5	MLP	576+1	194s	1.0×	58.2	62.0	78.5
LLaVA-1.5+Resampler	Resampler	288+1	126s	1.5×	52.1	56.8	76.0
LLaVA-1.5+EViT	MLP+EViT	288+1	126s	1.5×	54.6	60.0	77.9
LLaVA-1.5+SEC	MLP+SEC	288+1	126s	1.5×	**60.1**	**63.5**	**78.9**
LLaVA-1.5+Resampler	Resampler	256+1	116s	1.7×	51.6	56.0	75.2
LLaVA-1.5+Pool	MLP+Pool	256+1	116s	1.7×	52.4	57.6	76.4
LLaVA-1.5+EViT	MLP+EViT	256+1	116s	1.7×	52.8	59.6	77.1
LLaVA-1.5+SEC	MLP+SEC	256+1	116s	1.7×	**59.6**	**63.2**	**78.6**
LLaVA-1.5+Resampler	Resampler	192+1	102s	1.9×	50.1	55.2	74.3
LLaVA-1.5+EViT	MLP+EViT	192+1	102s	1.9×	51.6	58.6	76.3
LLaVA-1.5+SEC	MLP+SEC	192+1	102s	1.9×	**57.7**	**62.7**	**78.4**
LLaVA-1.5+Resampler	Resampler	144+1	94s	2.1×	47.6	54.6	72.0
LLaVA-1.5+Pool	MLP+Pool	144+1	94s	2.1×	50.0	56.2	73.6
LLaVA-1.5+EViT	MLP+EViT	144+1	94s	2.1×	51.2	58.0	76.0
LLaVA-1.5+SEC	MLP+SEC	144+1	94s	2.1×	**56.8**	**62.0**	**78.0**

[ICCV'25] *Semantic Equitable Clustering: A Simple and Effective Strategy for Clustering Vision Tokens*

- **MLLM**

Model	LLM	Connector	V-T Num	Res	TextVQA	GQA	VQAv2	VisWiz	SQA_{img}	Speed (↑)
7B LLM										
Shikra [5]	Vicuna-7B	MLP	257	224	-	-	77.4	-	-	-
IDEFICS-9B [28]	LLaMA-7B	Cross Attn	257	224	-	38.4	50.9	35.5	-	-
Qwen-VL [1]	Qwen-7B	Resampler	256	448	-	59.3	78.8	35.2	67.1	-
Qwen-VL-Chat [1]	Qwen-7B	Resampler	256	448	-	57.5	78.2	38.9	68.2	-
LLaVA-1.5 [36]	Vicuna-7B	MLP	577	336	58.2	62.0	78.5	50.0	66.8	1.0×
LLaVA-1.5+SEC (ours)	Vicuna-7B	MLP+SEC	257	336	**59.6**	**63.2**	**78.9**	**52.8**	**69.6**	**1.7×**
13B LLM										
InstructBLIP [8]	Vicuna-13B	Q-Former	32	224	-	49.5	-	33.4	63.1	-
BLIP-2 [30]	Vicuna-13B	Q-Former	32	224	-	41.0	41.0	19.5	61.0	-
LLaVA-1.5 [36]	Vicuna-13B	MLP	577	336	61.2	63.3	**80.0**	53.6	71.6	1.0×
LLaVA1.5+SEC (ours)	Vicuna-13B	MLP+SEC	257	336	**62.3**	**64.3**	**80.0**	**54.7**	**72.0**	**1.8×**

Model	LLM	Connector	V-T Num	Res	POPE	MMB	MM-Vet	Speed (↑)
7B LLM								
MiniGPT-4 [60]	Vicuna-7B	Resampler	32	224	72.2	24.3	22.1	-
mPLUG-Owl2 [57]	LLaMA2-7B	Resampler	32	224	-	49.4	-	-
LLaMA-AdapterV2 [14]	LLaMA2-7B	LLaMA-Adapter	257	224	-	41.0	31.4	-
Shikra [5]	Vicuna-7B	MLP	257	224	-	58.8	-	-
Qwen-VL [1]	Qwen-7B	Resampler	256	448	-	38.2	-	-
Qwen-VL-Chat [1]	Qwen-7B	Resampler	256	448	-	60.6	-	-
LLaVA-1.5 [35]	Vicuna-7B	MLP	577	336	86.1	64.3	31.1	1.0×
LLaVA1.5+SEC (ours)	Vicuna-7B	MLP+SEC	145	336	**86.1**	**68.4**	**31.7**	**2.1×**
13B LLM								
MiniGPT-4 [60]	Vicuna-13B	Resampler	32	224	-	-	24.4	-
BLIP-2 [29]	Vicuna-13B	Q-Former	32	224	85.3	-	22.4	-
LLaVA-1.5 [35]	Vicuna-13B	MLP	577	336	86.2	67.7	36.1	1.0×
LLaVA-1.5+SEC (ours)	Vicuna-13B	MLP+SEC	145	336	**86.4**	**69.2**	**37.3**	**2.2×**

Thanks!