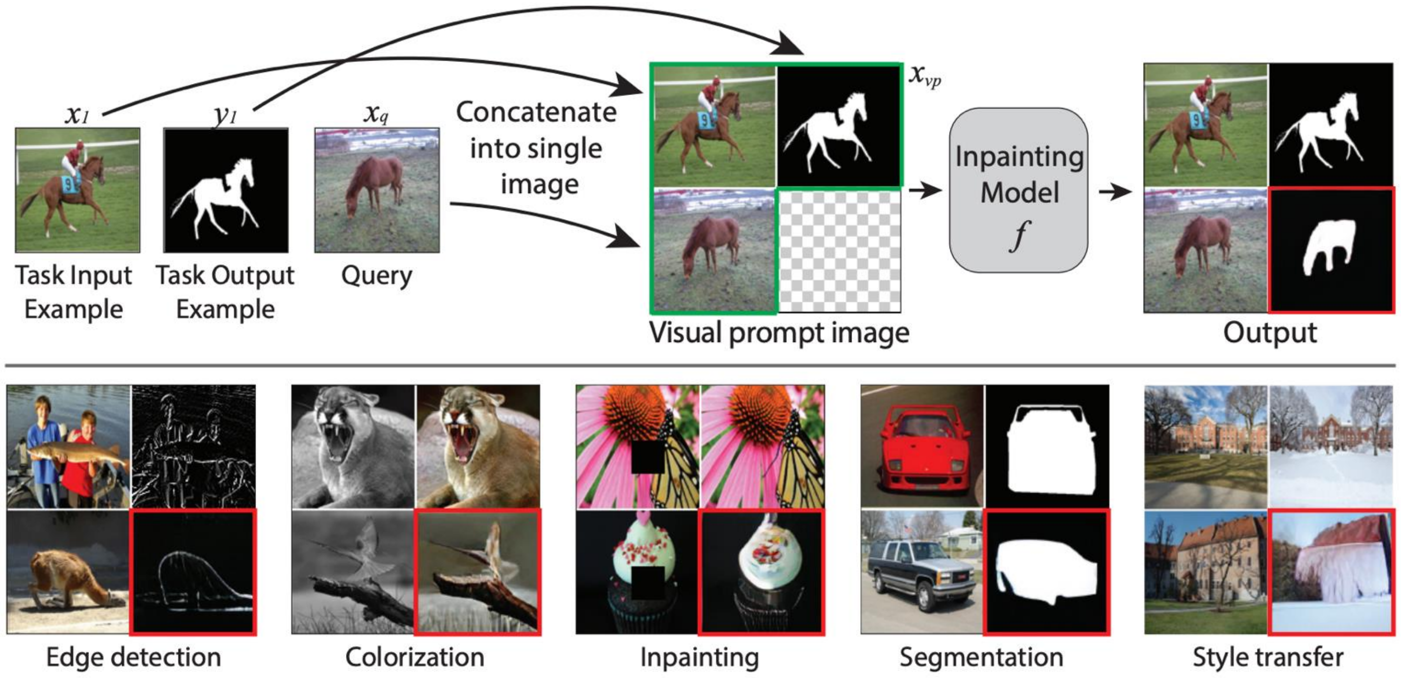
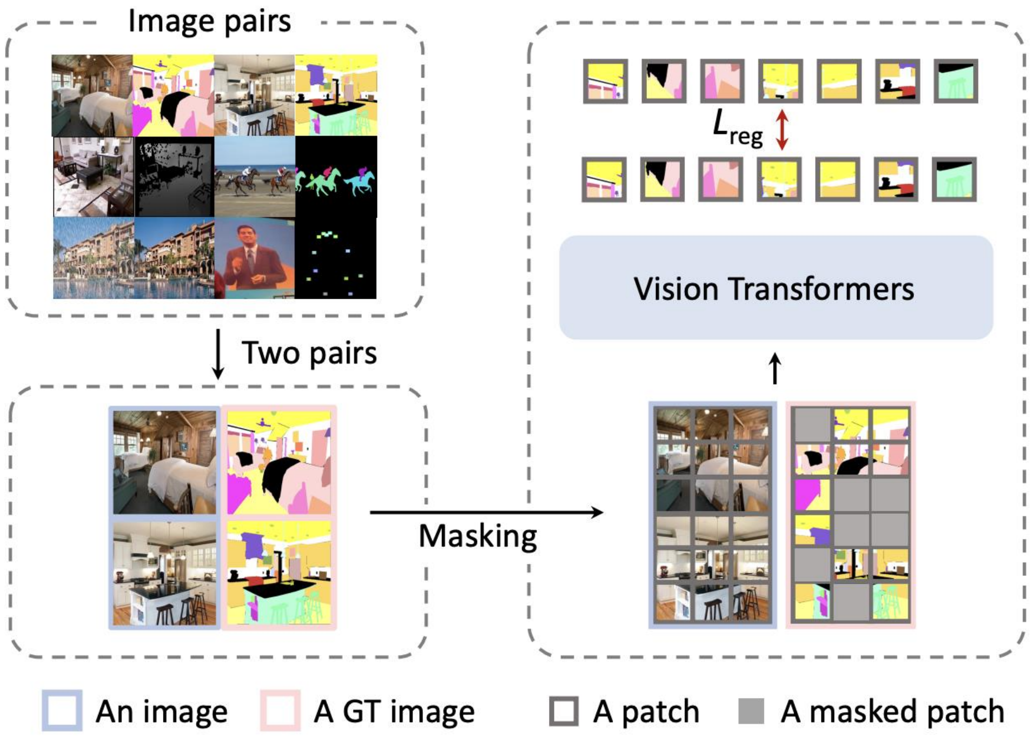
Towards Global Optimal Visual In-Context Learning Prompt Selection

Chengming Xu[*], Chen Liu[*], Yikai Wang[†], Yuan Yao, Yanwei Fu

Visual In-context Learning (VICL)

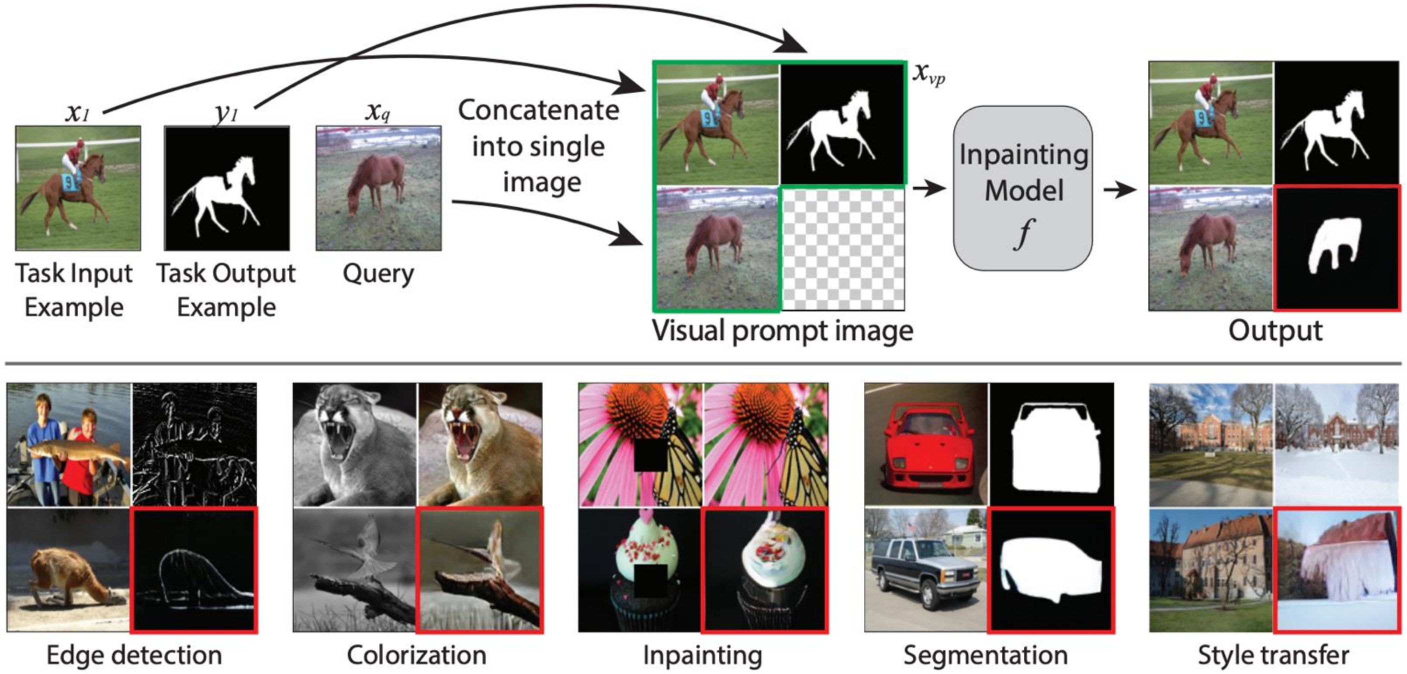




- inference using task&data domain provided by in-context prompts
- built on masked modelling

Bar A, Gandelsman Y, Darrell T, et al. Visual prompting via image inpainting[J]. NeurIPS2022.
Wang X, Wang W, Cao Y, et al. Images speak in images: A generalist painter for in-context visual learning[C]//CVPR2023.

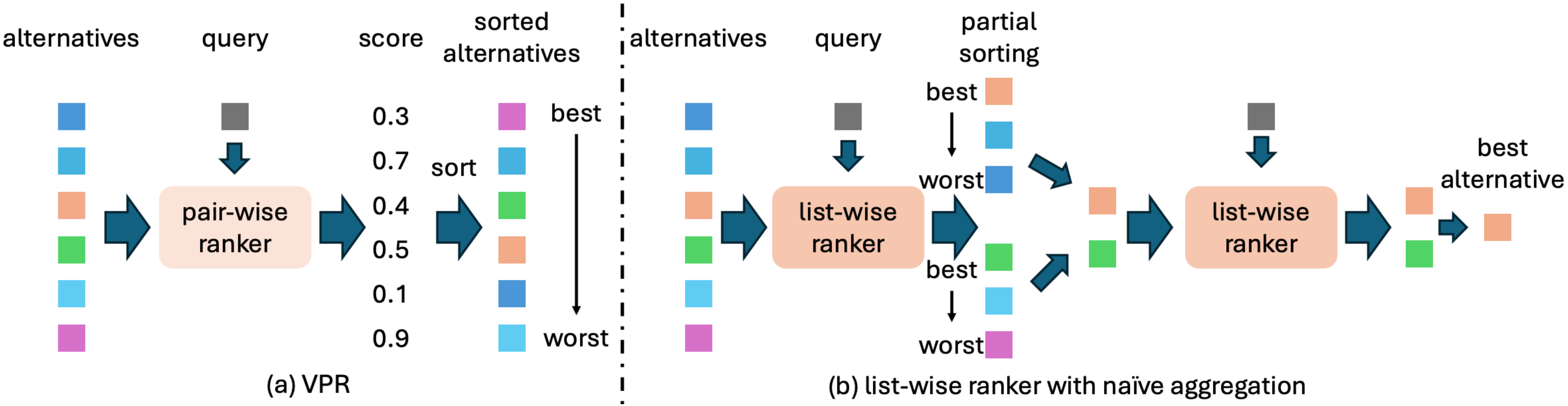
How to choose in-context examples?



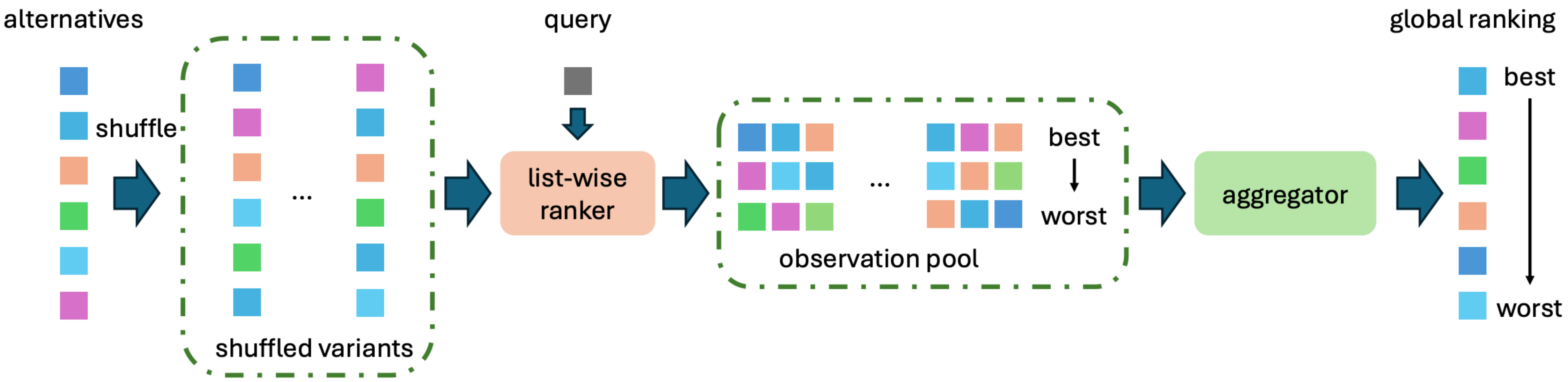
- random sample
- **ranking the candidates**
  - **right metric**
  - **proper comparison set**

Bar A, Gandelsman Y, Darrell T, et al. Visual prompting via image inpainting[J]. NeurIPS2022.
Wang X, Wang W, Cao Y, et al. Images speak in images: A generalist painter for in-context visual learning[C]//CVPR2023.

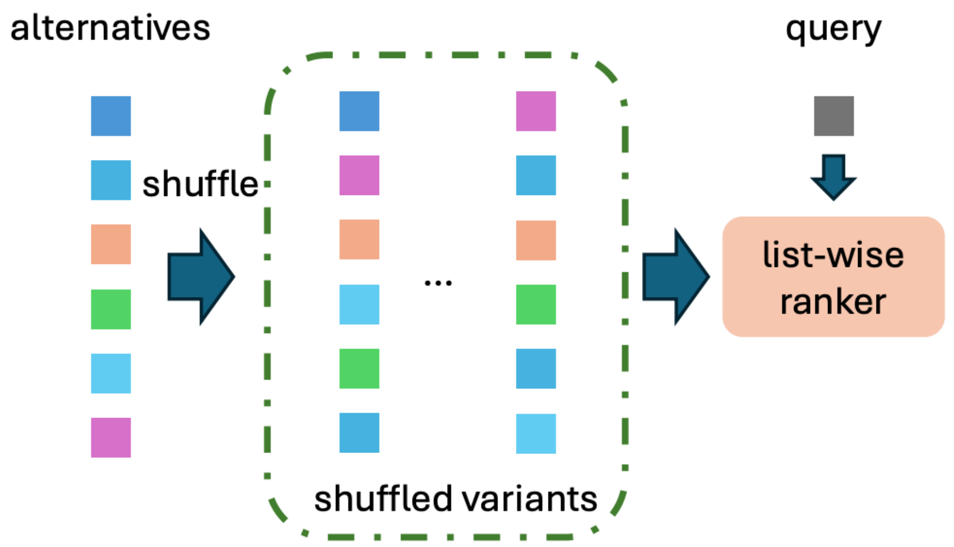
Rank in-context examples



List-wise ranker with consistency-aware aggregator



List-wise ranker with consistency-aware aggregator



transformer-based ranker

- sample a subset from alternative set
- concatenate their features from DINOv2
- process with extra transformer layers
- predict rankings with all class tokens

optimization: margin loss + NeuralNDCG + MSE

List-wise ranker with consistency-aware aggregator

Algorithm 1 Consistency-aware ranking aggregator

Input: Train set $\mathcal{X}_{train}$, query sample x_q, trained ranking models $\{\phi_k\}$, alternative set size K.

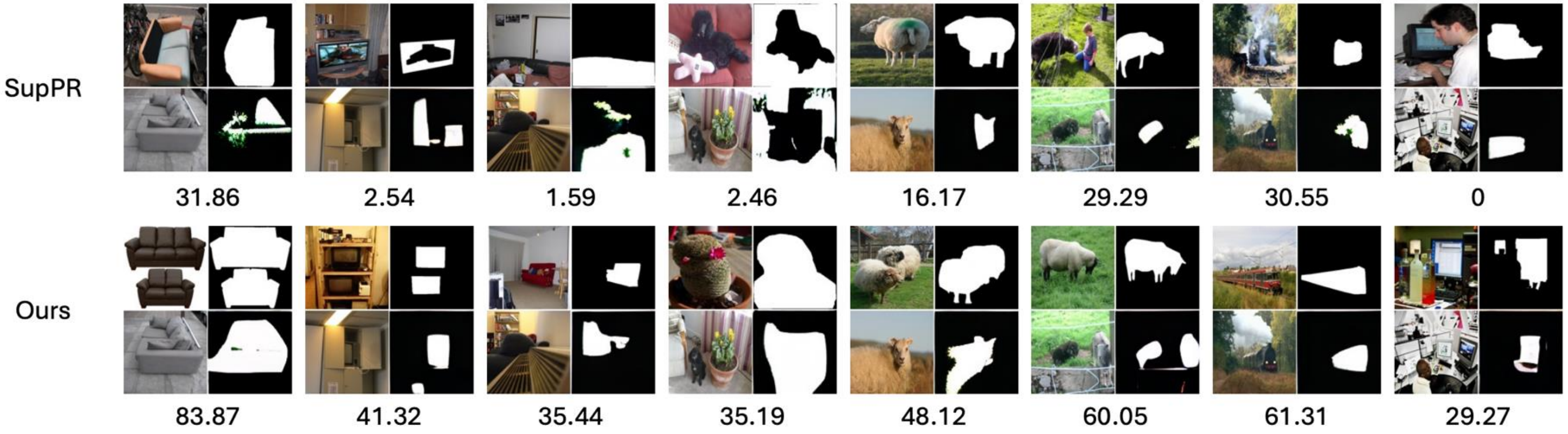
Alternative set $\mathcal{X}_R = topK_{\hat{x} \in \mathcal{X}_{train}}(sim(\hat{x}, x_q))$
Initial preference matrix set $\mathcal{S} := \emptyset$
for rank-k model ϕ_k **do**
    Build observation pool $\mathcal{X}_k$ from $\mathcal{X}_R$
    for randomly shuffled $\mathcal{X}_k^i$ from $\mathcal{X}_k$ **do**
        $\mathcal{R}_k^i = \bigcup_{x \in \mathcal{X}_k^i} \phi_k(x, x_q)$
        Aggregate $\mathcal{S}^i$ from $\mathcal{R}_k^i$
        $\mathcal{S} = \mathcal{S} \bigcup \mathcal{S}^i$
    end for
end for
Aggregate global ranking r as Eq. 3
return Top ranked sample.

aggregate all piecewise ranking prediction for a consistent global ranking

Experiment results

Table 1: Comparison of our method with previous in-context learning methods.

Model	Seg. (mIoU) ↑ Fold-0	Fold-1	Fold-2	Fold-3	Avg.	Det. (mIoU) ↑	Color. (MSE) ↓
MAE-VQGAN	28.66	30.21	27.81	23.55	27.56	25.45	0.67
UnsupPR	34.75	35.92	32.41	31.16	33.56	26.84	0.63
SupPR	37.08	38.43	34.40	32.32	35.56	28.22	0.63
Ours	**38.81**	**41.54**	**37.25**	**36.01**	**38.40**	**30.66**	**0.58**
prompt-SelF	42.48	43.34	39.76	38.50	41.02	29.83	—
Ours+voting	**43.23**	**45.50**	**41.79**	**40.22**	**42.69**	**32.52**	—



Experiment results

Backbone	**Strategy**	**Seg. (mIoU) ↑** Fold-0	Fold-1	Fold-2	Fold-3	Avg.	**Det. (mIoU) ↑**
CLIP	Naive	37.37	40.11	36.84	33.88	37.05	29.69
	Aggr.	38.58	41.34	37.66	35.91	38.37	30.79
DINOv1	Naive	38.78	40.02	36.92	35.12	37.71	28.03
	Aggr.	39.25	42.27	38.45	36.77	39.19	29.19
DINOv2	Naive	37.51	39.69	36.62	34.58	37.10	29.58
	Aggr.	38.81	41.54	37.25	36.01	38.40	30.66

Strategy	**Seg. (mIoU) ↑** Fold-0	Fold-1	Fold-2	Fold-3	Avg.	**Det. (mIoU) ↑**
#1 rank	38.81	41.54	37.25	36.01	38.40	30.66
#2 rank	38.13	41.66	37.62	35.35	38.19	30.76
#3 rank	38.66	41.08	37.36	35.91	38.25	30.61
top2 fusion	39.08	42.61	38.17	36.67	39.13	30.16
top3 fusion	40.07	42.48	38.77	37.61	39.73	31.85
top5 fusion	40.12	42.59	39.09	37.28	39.77	32.08

Thank you!