Supplementary Material
IFCap: Image-like Retrieval and Frequency-based Entity Filtering for Zero-shot Captioning

A Image-like Retrieval

We observe that Image-like Retrieval is also applicable to other models that employ text-to-text retrieval [2]. We perform **ILR** with $\epsilon_r = 0.02$ in the training time of Knight. In the COCO test set, every metric except METEOR is improved compared to vanilla Knight, verifying the effectiveness of our **ILR**.

B Design choice

We find the best threshold setting in heuristic way and adaptive way. In the former case Table 3, we set τ ranging from 1 to 8, which is the minimum and maximum value of the given setting. Above 8, performance freeze due to none of the entities being retrieved. In COCO test, we use $l = 9$ and $l = 7$ in Flickr30K test split. We can notice that each domain has different optimal τ, COCO at 5 and Flickr30K at 3 for the CIDEr score. In contrast to the heuristic way, we can assume such distribution exists from frequencies F. We try Gaussian distribution and Log-normal distribution with μ, $\mu + \sigma$, and $\mu + 2\sigma$, capturing upper 50%, 15.8%, and 2.2% based on the frequency of entity. In Table 4, we observe $\tau_{\text{adap}} = \mu + \sigma$ almost reproduce the performance of global optimal in the heuristic threshold. If ground truth does not exist or computing resource is limited, the adaptive threshold becomes attractive.

C Comparison with Baselines

We compare baselines [1, 2] with IFCap and IFCap* in every domain, including in-domain captioning, cross-domain captioning, and video captioning. Result can be found in Table 5

Method	COCO			
	B@4	M	C	S
Knight	27.8	**26.4**	98.9	19.6
Knight + ILR	**29.8**	25.6	**102.7**	**19.7**

Table 1: Effect of **Image-like Retrieval** on Knight.

HyperParameters	**COCO**	**Flickr30k**	**NoCaps**	**MSVD**	**MSR-VTT**
Epochs	5	30	-	10	10
$\boldsymbol{l}$	9	7	7	7	7
$\boldsymbol{\tau}$	5	3	3	5	6

Table 2: Hyperparameter table.

D Hyperparameter

We include details about our experiments in each dataset in Table 2.

Workshop on Video-Language Models at 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

τ	COCO				Flickr30k			
	B@4	M	C	S	B@4	M	C	S
1	6.5	18.7	6.4	17.0	6.8	18.9	3.9	15.4
2	21.4	26.5	80.3	21.0	18.9	**23.4**	52.2	**17.9**
3	28.1	26.8	103.6	**21.1**	23.5	23.0	**64.4**	17.0
4	30.2	**26.7**	107.7	20.7	**23.8**	22.3	61.1	15.9
5	**30.8**	**26.7**	**108.0**	20.3	**23.8**	21.9	59.1	15.3
6	30.4	26.4	106.2	19.9	23.6	21.7	57.3	15.0
7	30.0	26.1	104.6	19.6	23.6	21.6	56.5	14.8
8	29.8	26.0	103.4	19.4	23.7	21.6	55.9	14.7

Table 3: Ablation studies of heuristic threshold τ of **Entity Filtering**.

τ_{adap}	COCO				Flickr30k			
	B@4	M	C	S	B@4	M	C	S
Lognormal(μ, σ^2)								
μ	22.0	26.6	83.8	**21.1**	19.0	**23.4**	52.7	**17.9**
$\mu + \sigma$	29.1	**26.7**	**106.6**	20.7	22.0	22.9	**63.0**	17.2
$\mu + 2\sigma$	**29.6**	26.1	103.5	19.6	**23.3**	21.8	58.1	15.3
$N(\mu, \sigma^2)$								
μ	24.9	**26.7**	95.9	**21.1**	19.2	**23.2**	55.6	**17.7**
$\mu + \sigma$	**30.1**	26.6	**107.5**	20.4	22.3	22.5	**62.3**	16.4
$\mu + 2\sigma$	29.8	26.2	104.7	19.7	**23.4**	21.9	58.5	15.5
Best (H)	30.8	26.7	108.0	20.3	23.5	23.0	64.4	17.0

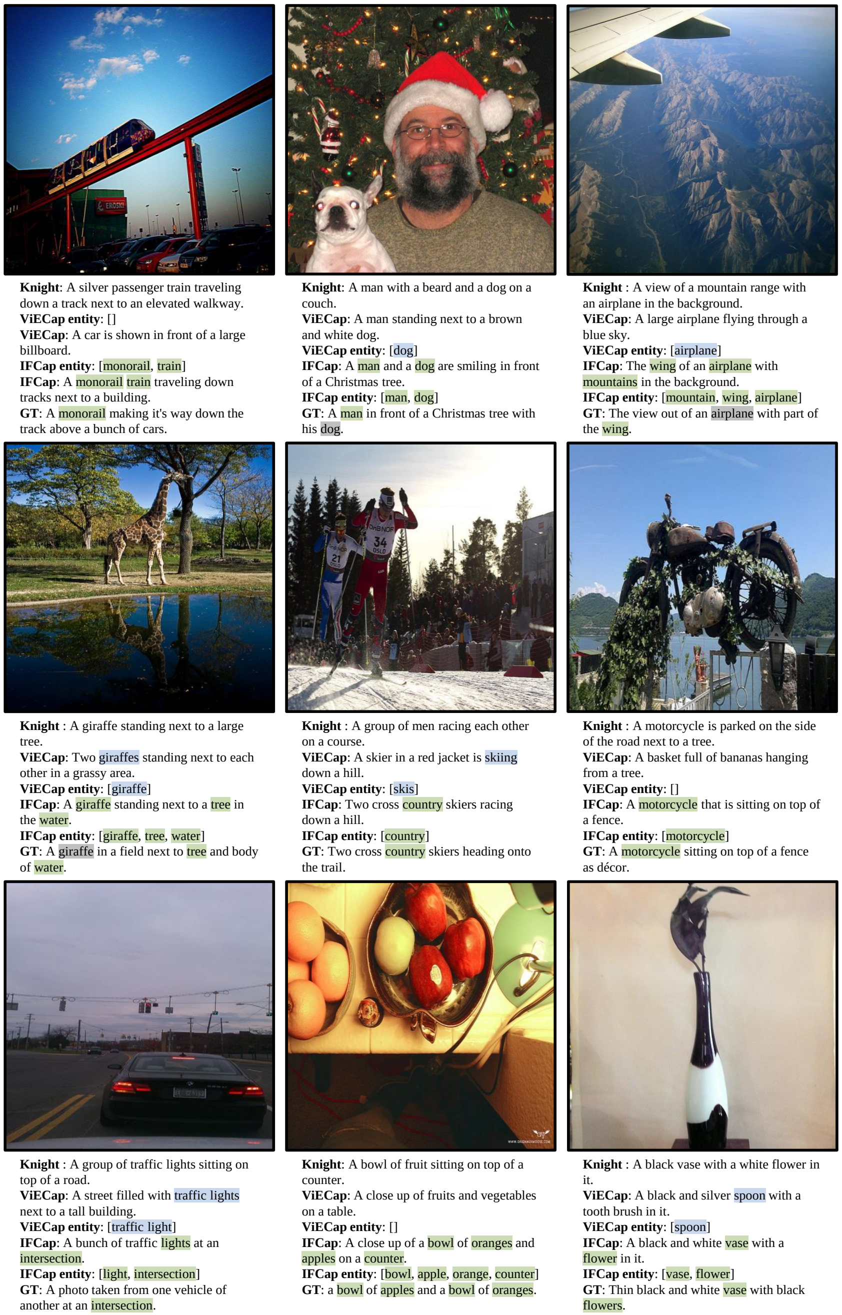
Table 4: Ablation studies of adaptive threshold τ_{adap} of **Entity Filtering.**

	In−domain				Cross−domain												Video Captioning			
Method	**COCO**		**Flickr**		**COCO ⟹ NoCaps Val**								**COCO ⟹ Flickr**		**Flickr ⟹ COCO**		**MSR-VTT**		**MSVD**	
					In		Near		Out		Entire									
	C	S	C	S	C	S	C	S	C	S	C	S	C	S	C	S	C	S	C	S
ViECap	92.9	18.2	47.9	13.6	61.1	10.4	64.3	9.9	65.0	8.6	66.2	9.5	38.4	11.2	54.2	12.5	-	-	-	-
Knight	98.9	19.6	56.3	16.3	-	-	-	-	-	-	-	-	48.9	14.2	64.4	15.1	31.9	**8.5**	63.8	5.0
IFCap*	102.0	20.0	59.8	15.8	70.1	11.2	**72.5**	10.9	**72.1**	**9.6**	**74.0**	10.5	47.5	12.7	60.7	13.6	20.8	4.1	40.2	3.4
IFCap	**108.0**	**20.3**	**64.4**	**17.0**	**75.8**	**12.4**	72.3	**11.6**	60.2	8.9	70.5	**10.8**	**59.2**	**15.6**	**76.3**	**17.3**	**38.9**	6.7	**83.9**	**6.3**

Table 5: Overall comparison among baselines and IFCap. ⋆: without **Entity Filtering** module in the inference time.

E Qualitative Results

We show additional qualitative results in Fig. 1.



Knight: A silver passenger train traveling down a track next to an elevated walkway.
ViECap entity: []
ViECap: A car is shown in front of a large billboard.
IFCap entity: [monorail, train]
IFCap: A monorail train traveling down tracks next to a building.
GT: A monorail making it's way down the track above a bunch of cars.

Knight: A man with a beard and a dog on a couch.
ViECap: A man standing next to a brown and white dog.
ViECap entity: [dog]
IFCap: A man and a dog are smiling in front of a Christmas tree.
IFCap entity: [man, dog]
GT: A man in front of a Christmas tree with his dog.

Knight : A view of a mountain range with an airplane in the background.
ViECap: A large airplane flying through a blue sky.
ViECap entity: [airplane]
IFCap: The wing of an airplane with mountains in the background.
IFCap entity: [mountain, wing, airplane]
GT: The view out of an airplane with part of the wing.

Knight : A giraffe standing next to a large tree.
ViECap: Two giraffes standing next to each other in a grassy area.
ViECap entity: [giraffe]
IFCap: A giraffe standing next to a tree in the water.
IFCap entity: [giraffe, tree, water]
GT: A giraffe in a field next to tree and body of water.

Knight : A group of men racing each other on a course.
ViECap: A skier in a red jacket is skiing down a hill.
ViECap entity: [skis]
IFCap: Two cross country skiers racing down a hill.
IFCap entity: [country]
GT: Two cross country skiers heading onto the trail.

Knight : A motorcycle is parked on the side of the road next to a tree.
ViECap: A basket full of bananas hanging from a tree.
ViECap entity: []
IFCap: A motorcycle that is sitting on top of a fence.
IFCap entity: [motorcycle]
GT: A motorcycle sitting on top of a fence as décor.

Knight : A group of traffic lights sitting on top of a road.
ViECap: A street filled with traffic lights next to a tall building.
ViECap entity: [traffic light]
IFCap: A bunch of traffic lights at an intersection.
IFCap entity: [light, intersection]
GT: A photo taken from one vehicle of another at an intersection.

Knight: A bowl of fruit sitting on top of a counter.
ViECap: A close up of fruits and vegetables on a table.
ViECap entity: []
IFCap: A close up of a bowl of oranges and apples on a counter.
IFCap entity: [bowl, apple, orange, counter]
GT: a bowl of apples and a bowl of oranges.

Knight : A black vase with a white flower in it.
ViECap: A black and silver spoon with a tooth brush in it.
ViECap entity: [spoon]
IFCap: A black and white vase with a flower in it.
IFCap entity: [vase, flower]
GT: Thin black and white vase with black flowers.

Figure 1: Qualitative result on COCO test set. We highlight the retrieved entities and their appearance in the generated captions with IFCap , ViECap and Intersection .

References

[1] Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. Transferable decoding with visual entities for zero-shot image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3136–3146, 2023.

[2] Junyang Wang, Ming Yan, Yi Zhang, and Jitao Sang. From association to generation: Text-only captioning by unsupervised cross-modal mapping. *arXiv preprint arXiv:2304.13273*, 2023.