Enhancing Relational Understanding in CLIP Leveraging HNC

Wen Wen

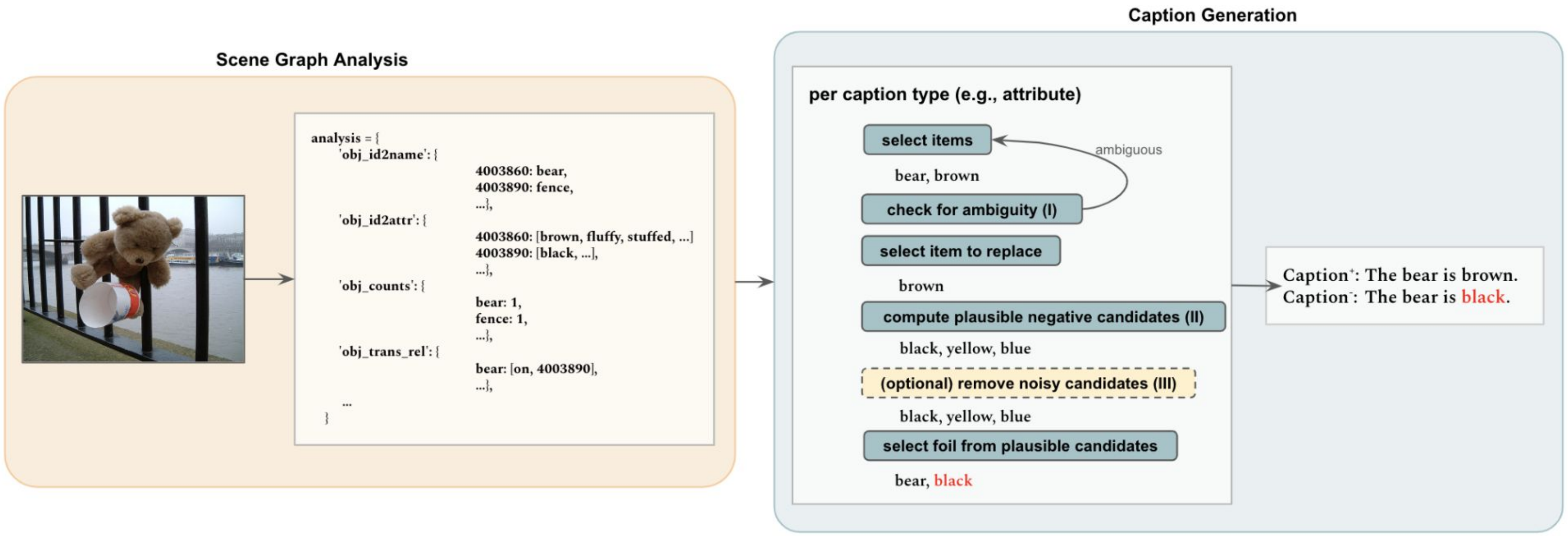
Research Question

Cos similarity score of image and caption:

Pos: "The trees are behind the fence."
→ViT-B/32: **0.1833**
Neg: "The trees are in front of the fence."
→ViT-B/32: **0.1907**
→ Margin(pos-neg): **-0.0073**

CILP cannot handle relations in images well.

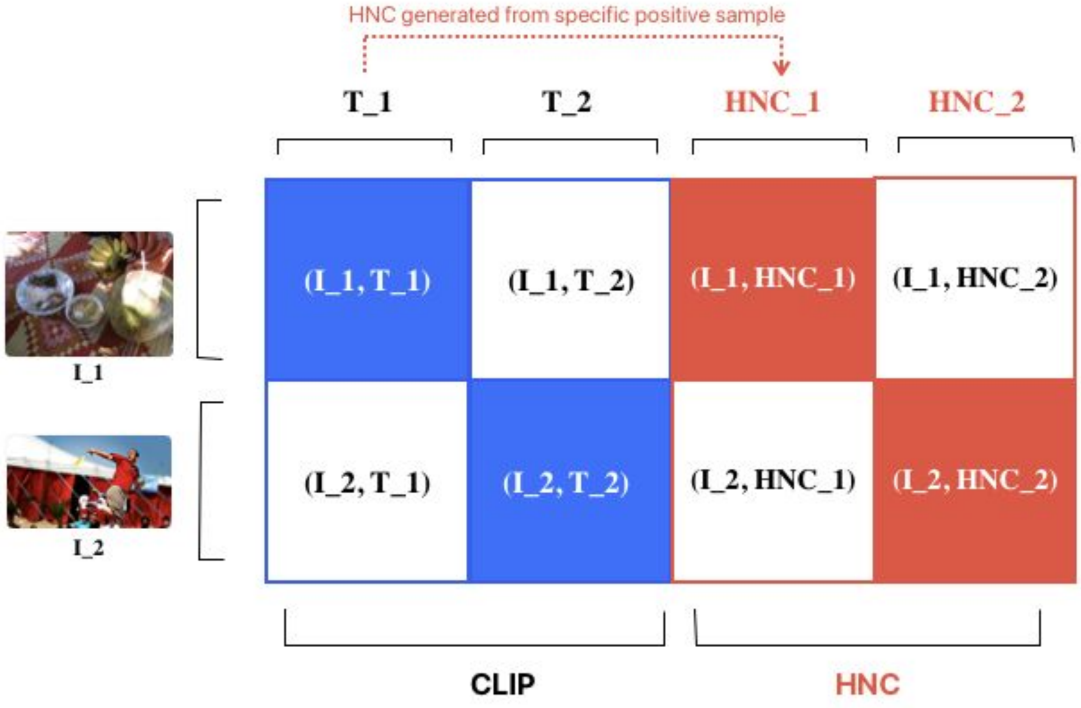
Hard Negative Caption(HNC) Dataset



For each scene graph, this pipeline is run through to generate hard negative captions.

https://aclanthology.org/2023.conll-1.24.pdf

Combine Contrastive Learning with HNC



Blue part:
Contrastive learning for CLIP

Red Part:
Additional HNC part

Loss Function

$$\mathcal{L}_i = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\frac{\mathrm{sim}(I_i, T_i)}{\tau}\right)}{\exp\left(\frac{\mathrm{sim}(I_i, T_i)}{\tau}\right) + \sum_{j \neq i} \exp\left(\frac{\mathrm{sim}(I_i, T_j)}{\tau}\right) + \alpha \exp\left(\frac{\mathrm{sim}(I_i, T_i^{\mathrm{HNC}})}{\tau}\right) + \sum_{j \neq i} \exp\left(\frac{\mathrm{sim}(I_i, T_j^{\mathrm{HNC}})}{\tau}\right)}$$

$$\mathcal{L}_t = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\frac{\mathrm{sim}(T_i, I_i)}{\tau}\right)}{\exp\left(\frac{\mathrm{sim}(T_i, I_i)}{\tau}\right) + \sum_{j \neq i} \exp\left(\frac{\mathrm{sim}(T_i, I_j)}{\tau}\right)}$$

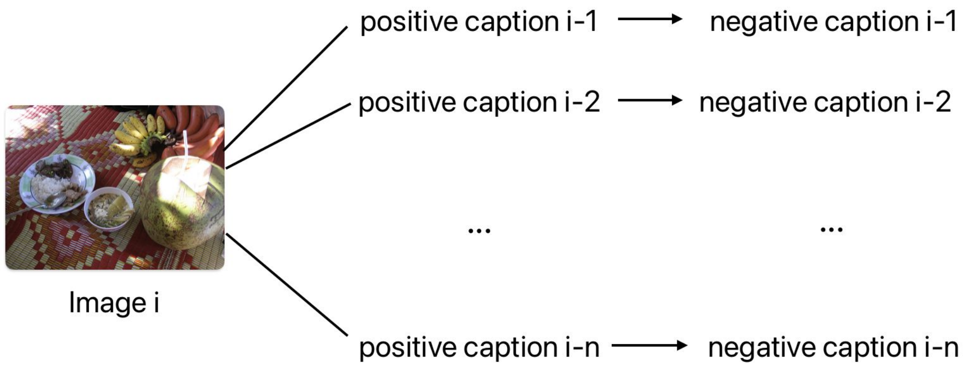
Evaluation method

- Positive capture and image cosine similarity score
- Hard negative caption and image cosine similarity score
- Margin: pos-hard neg
- Random negative and image cosine similarity score
- Measure the margin -> pos/hnc >= Threshold
- Threshold =[1, 1.1, 1.2, 1.5, 2, 3]

Avg_Pos	**Avg_Neg**	**Margin**	**Avg_Rand_Neg**	**threshold_1**	**threshold_1.1**	**threshold_1.2**	**threshold_1.5**	**threshold_2**	**threshold_3**
0.2471	0.2449	0.0022	0.1817	0.5393	0.1161	0.0459	0.0000	0.0000	0.0000

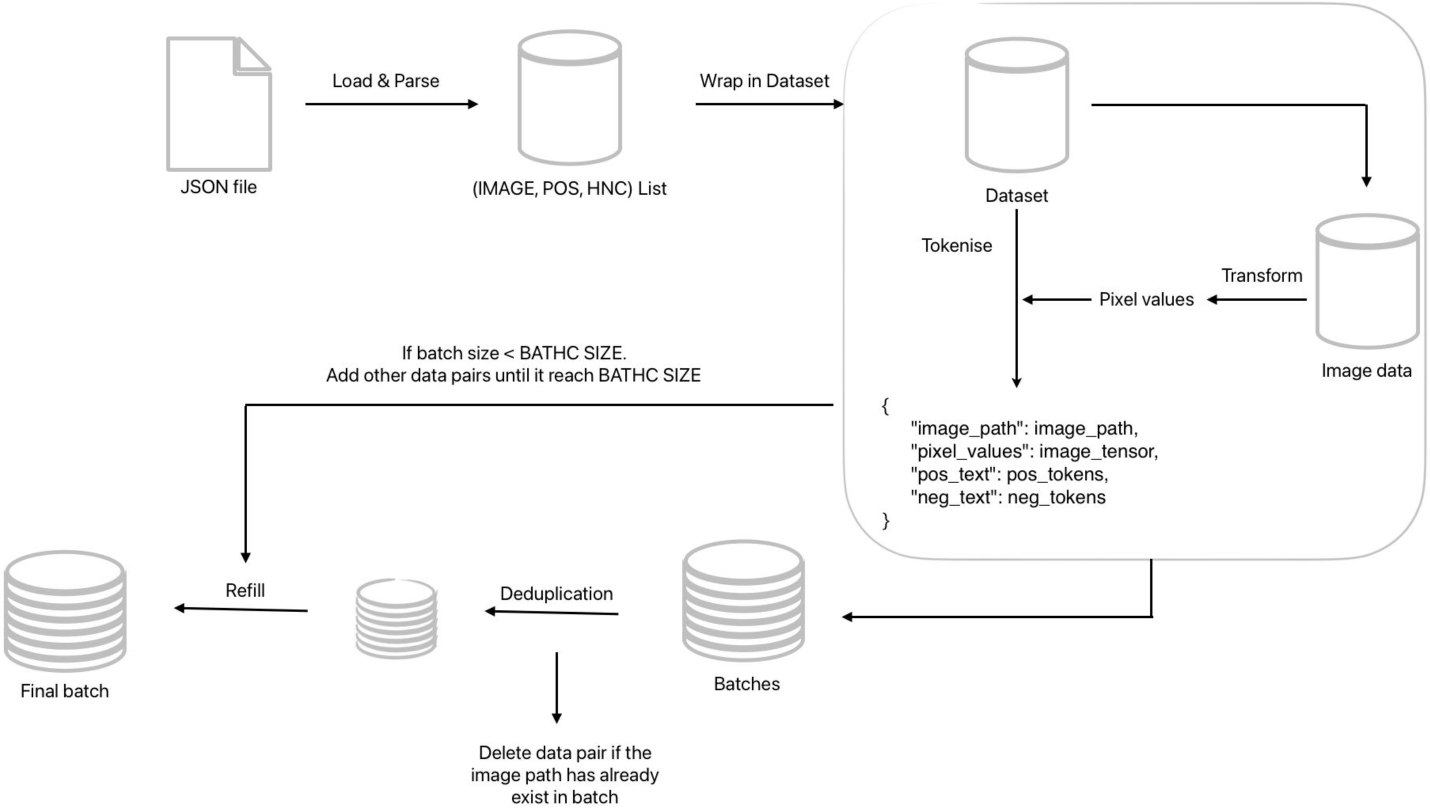
Scores on HNC test dataset for Vit-B/32

Data Preprocessing

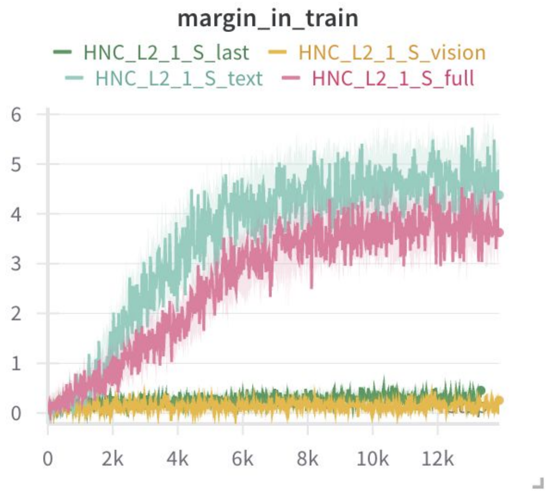


- One image has multiple positive captions and relative hard negative captions.
- For contrastive learning, it's important to remove the repeated image path in each batch.
- Otherwise, repeated positive captions can be treated as random negative captions for that image to influence training

Data Preprocessing



Training parameters



Mode	Token Embedding	Text Encoder	Text Projection	Visual Encoder	Vision Projection
text_encoder	True	True	True	–	-
vision_encoder	–	–	–	True	True
full_encoder	True	True	True	True	True
last_encoder	–	True (last block)	–	True (last block)	–

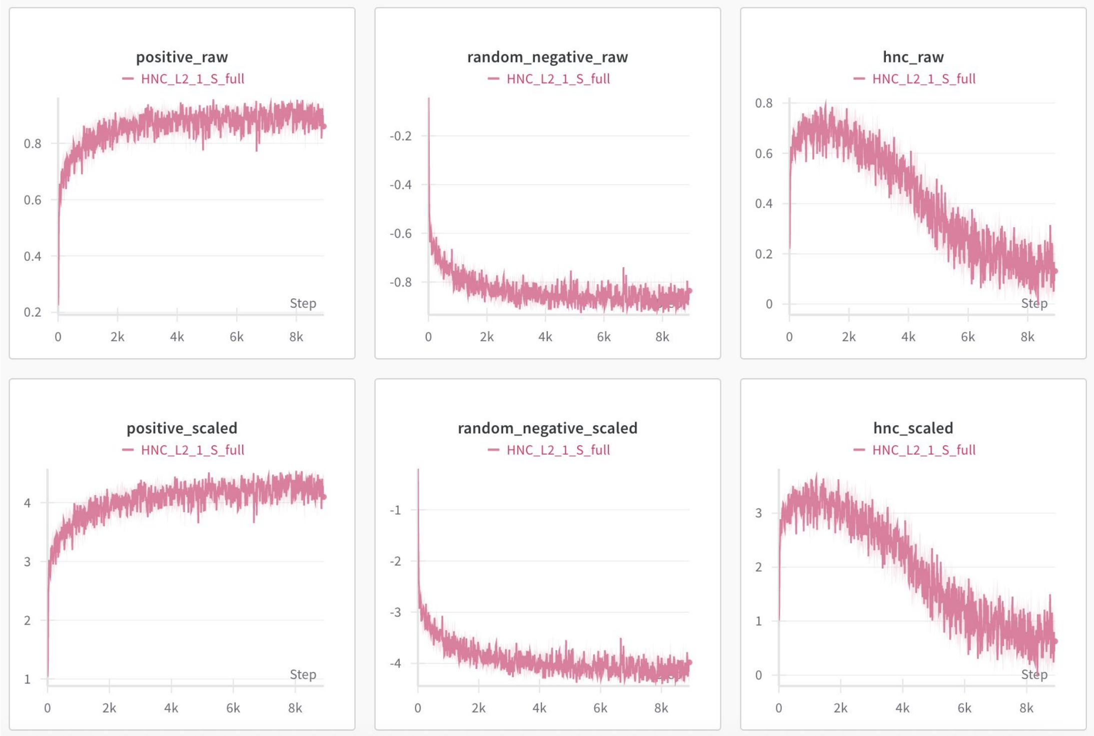
- Margin doesn't get improved during training when only train vision encoder or last encoder

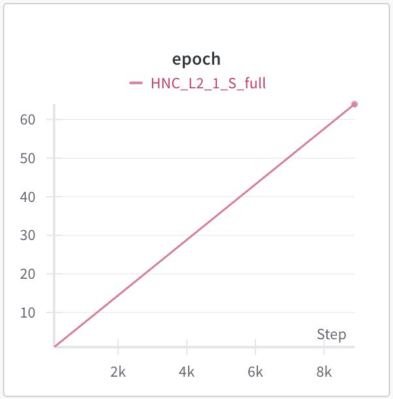
During training:

Positive score-> larger

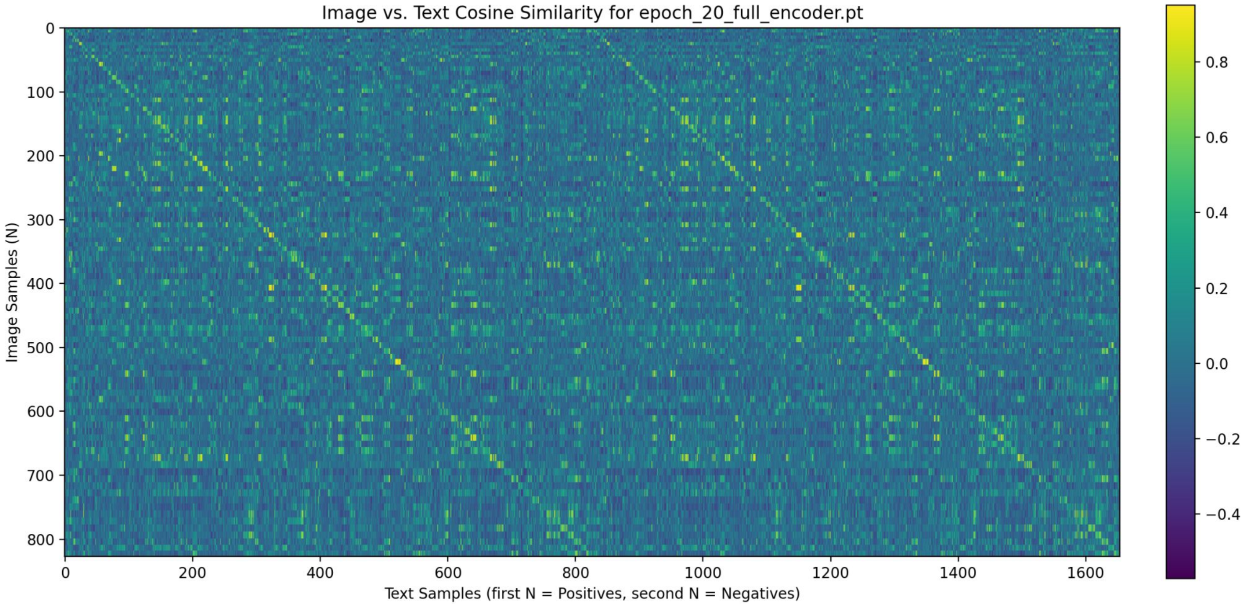
Random negative score -> smaller

Hard negative score ->larger then smaller





The test dataset score for fine-tuned model



Because the hard negative captions still contain a lot of image information.
So HNC scores are higher than random negative scores.

Test data scores and comparison

Model	Avg_Pos	Avg_Neg	Margin	Avg_Rand_Neg	t_1	t_1.1	t_1.2	t_1.5	t_2	t_3
CLIP	0.2471	0.2449	0.0022	0.1817	0.5393	0.1161	0.0459	0.0000	0.0000	0.0000
HNC_L2_vision	0.3661	0.3584	0.0077	0.0115	0.4982	0.2273	0.1439	0.0762	0.0314	0.0169
HNC_L2_text	0.0386	0.0150	0.0236	-0.2713	0.4268	0.3265	0.2696	0.1959	0.1258	0.0713
HNC_L2_last	0.4479	0.4376	0.0103	0.0182	0.526	0.2588	0.1753	0.0943	0.0508	0.0302
HNC_L2_full	0.3080	0.2902	0.0178	0.0158	0.4389	0.3362	0.2805	0.1850	0.1149	0.0701

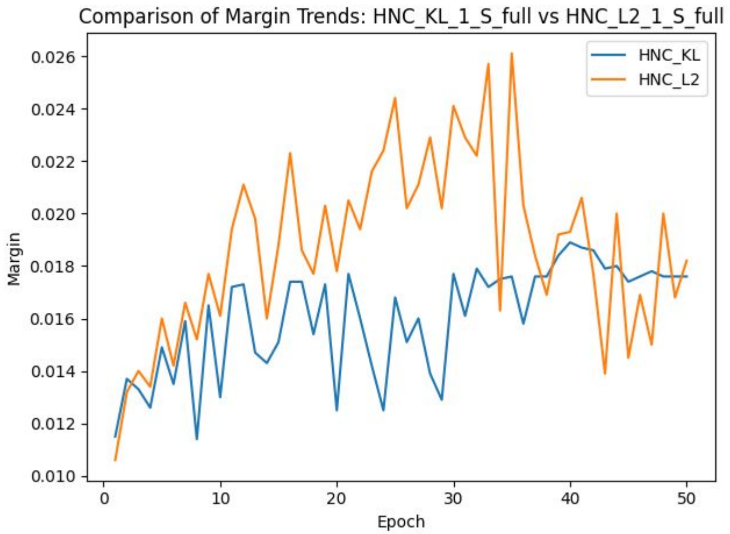
Here, T_1 means threshold = 1

L2 regularization vs KL divergence

$$\mathrm{Reg}_{L2} = \sum_i \left(\theta_i - \theta_i^{(CLIP)}\right)^2$$

$$s_j = \frac{\exp\left(s_j/T\right)}{\sum_k \exp\left(s_k/T\right)}, \quad t_j = \frac{\exp\left(t_j/T\right)}{\sum_k \exp\left(t_k/T\right)},$$

$$D_{\mathrm{KL}}(t \,\|\, s) = \sum_{j=1}^{2B} t_j \, \log \frac{t_j}{s_j},$$



L2 regularization vs KL divergence

Model	Avg_Pos	Avg_Neg	Margin	Avg_Rand_Neg	t_1	t_1.1	t_1.2	t_1.5	t_2	t_3
CLIP	0.2471	0.2449	0.0022	0.1817	0.5393	0.1161	0.0459	0.0000	0.0000	0.0000
HNC_L2_full	0.3080	0.2902	0.0178	0.0158	0.4389	0.3362	**0.2805**	**0.1850**	**0.1149**	**0.0701**
HNC_KL_ful	**0.3959**	**0.3741**	**0.0218**	0.0211	**0.4958**	**0.3482**	0.2769	**0.1850**	0.1125	0.0556

Loss function2: HNC+DPO

$$\max_{\pi_\theta} \; E_{x\sim D,\, y\sim\pi_\theta(y|x)}\left[r_\phi(x,y)\right] \;-\; \beta\, D_{\mathrm{KL}}\left[\pi_\theta(y\mid x)\,\|\,\pi_{\mathrm{ref}}(y\mid x)\right]$$

$$p(y_1 > y_2) = \frac{e^{r(x,y_1)}}{e^{r(x,y_1)} + e^{r(x,y_2)}}$$

$$\mathcal{L}_{\mathrm{DPO}}(\theta) = -E_{(x,y^+,y^-)\sim D}\left[\log\sigma\left(\beta\left[\log\frac{\pi_\theta(y^+\mid x)}{\pi_{\mathrm{ref}}(y^+\mid x)} - \log\frac{\pi_\theta(y^-\mid x)}{\pi_{\mathrm{ref}}(y^-\mid x)}\right]\right)\right]$$

Scoring function : logit_scale × cosine_similarity

Average over generated responses: Expectation → Empirical mean

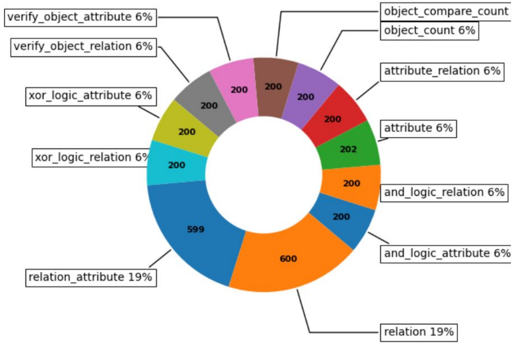
$$\boxed{\mathcal{L}_{\mathrm{DPOCLIP}} = -\frac{1}{B}\sum_{i=1}^{B}\log\sigma\left(\beta\left[\left(r_\theta(x_i,y_i^+)-r_{\mathrm{ref}}(x_i,y_i^+)\right)-\left(r_\theta(x_i,y_i^-)-r_{\mathrm{ref}}(x_i,y_i^-)\right)\right]\right)}$$

HNC+DPO

- Only using DPO ($L_{DPOCLIP}$)
- DPO + contrastive loss ($L = L_{DPOCLIP} + L_{contrastive}$)
- DPO + contrastive loss, using L2 regularization instead of KL divergence

Model	Avg_Pos	Avg_Neg	Margin	Avg_Rand_Neg	t_1	t_1.1	t_1.2	t_1.5	t_2	t_3
CLIP	0.2471	0.2449	0.0022	0.1817	**0.5393**	0.1161	0.0459	0.0000	0.0000	0.0000
DPO_KL_full	0.3735	0.2559	**0.1176**	0.6572	0.3591	0.0387	0.0278	0.0109	0.0036	0.0036
C_DPO_KL_full	0.3861	0.3737	0.0125	0.0124	0.4692	0.3083	**0.2563**	**0.1596**	**0.1016**	0.0496
C_DPO_L2_full	**0.3901**	**0.3749**	0.0152	0.0195	0.4752	**0.3144**	0.2455	0.1584	0.1004	**0.0508**

Test set sample 1, type='relation'



Pos:

The trees are behind the fence

Neg:

The trees are in front of the fence

Test set sample 2, type='relation'

Pos:
The chair is in front of the computer desk

Neg:
The couch is in front of the computer desk

Test set sample 3, type='relation'

Pos:

The wall is behind the elephant

Neg:

The wall is behind the trees

Test set samples and relative score

Pos: The trees are behind the fence
Neg: The trees are in front of the fence

Pos: The chair is in front of the computer desk
Neg: The couch is in front of the computer desk

Pos: The wall is behind the elephant
Neg: The wall is behind the trees

	1_pos	1_neg	1_margin	2_pos	2_neg	2_margin	3_pos	3_neg	3_margin
ViT-B/32	0.1833	0.1907	-0.0073	0.2214	0.2534	-0.0320	0.2343	0.1693	0.0649
HNC_L2_1_text	-0.0925	-0.0930	0.0005	0.0591	0.2440	-0.1849	0.4158	-0.2013	0.6171
HNC_L2_1_vision	0.3240	0.3191	0.0049	0.4709	0.5215	-0.0505	0.1744	0.2239	-0.0494
HNC_L2_1_last	0.1215	0.0958	0.0257	0.6099	0.6699	-0.0601	0.6548	0.4517	0.2031
HNC_L2_1_full	0.2223	0.1902	0.0321	0.5420	0.5254	0.0166	**0.5879**	**0.0791**	**0.5088**
HNC_KL_1_full	**0.3765**	**0.2402**	**0.1362**	**0.3577**	**0.2898**	**0.0679**	0.4258	0.1886	0.2372

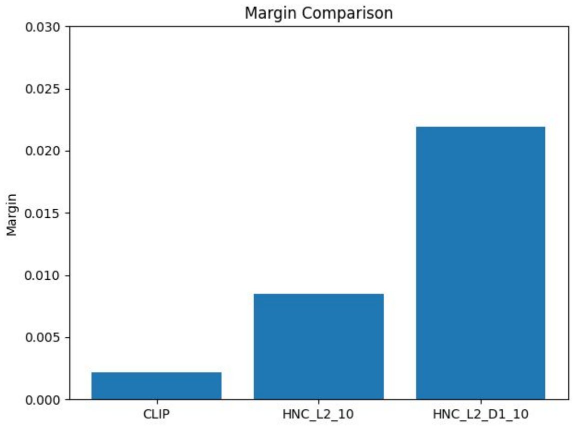
Performance using Coco test dataset

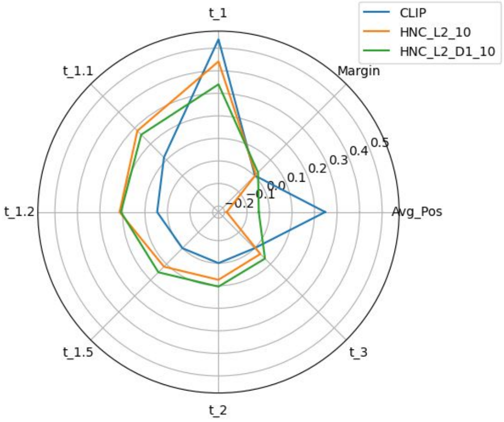
Model	Avg_Pos	Avg_Neg	Margin	Avg_Rand_Neg	threshold_1	threshold_1.1	threshold_1.2	threshold_1.5	threshold_2	threshold_3
CLIP	0.3046	0.2972	0.0074	0.1498	**0.7345**	0.0812	0.0116	0.0001	0.0000	0.0000
HNC_L2_1_full	0.5415	0.5098	**0.0317**	0.0406	0.602	**0.321**	**0.224**	**0.125**	**0.061**	**0.029**
HNC_KL_1_full	**0.6253**	**0.6089**	0.0165	**0.0299**	0.616	0.187	0.119	0.041	0.018	0.008

Limitations:

1. Large HNC weight:

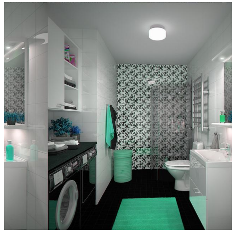
- Using larger HNC weight doesn’t performance well.
- Using dynamic weight (small at first, increasing the weight during training), better than fixed large weight, but still performance bad.



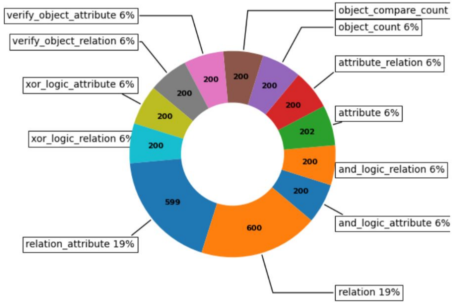


Limitations:

2. Unstable/bad performance in other ‘type” e.g. verify_object_attribute:

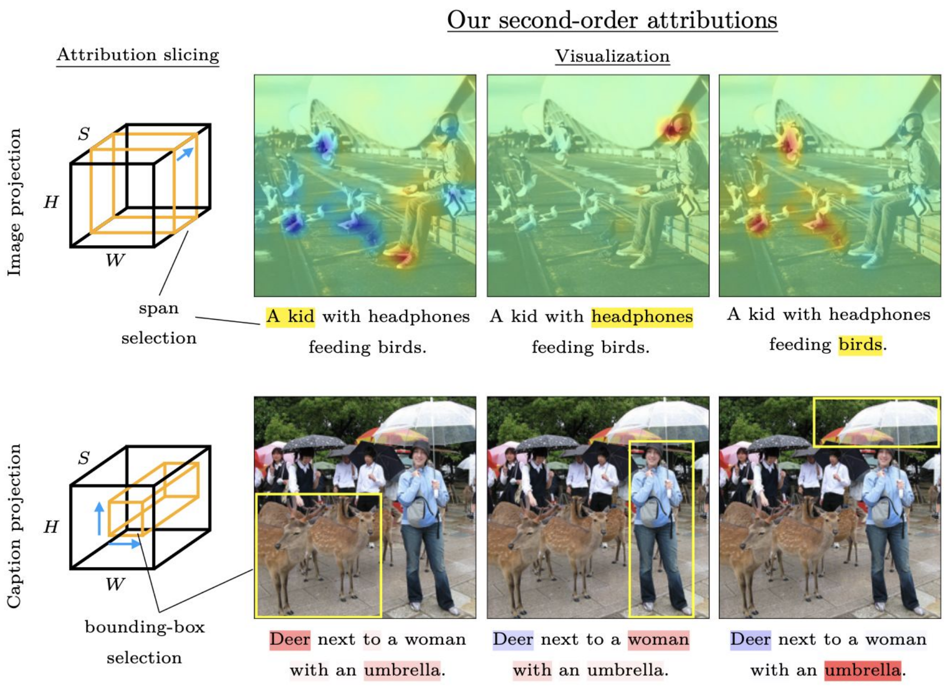


	ViT-B/32	HNC_L2_1_full	HNC_KL_1_full
Pos: There is at least one washing machine that is black.	0.2050	0.5151	0.4817
Neg: There is no washing machines that is black.	0.2017	0.5825	0.5205
Margin	0.0033	**-0.0674**	**-0.0388**



	ViT-B/32	HNC_L2_1_full	HNC_KL_1_full
Pos: There is at least one couch that is black.	0.2419	0.5244	0.3022
Neg: There is no couch that is black.	0.2365	0.2074	0.1110
Margin	0.0055	**0.3170**	**0.1912**

Explanation of Image-HNC(Code-to-be-released)



Explaining Caption-Image Interactions in CLIP models with Second-Order Attributions (Pascal Tilli et al.)

- Image, caption → second-order attribution pipeline → which image patches and which caption tokens drive their similarity score in a CLIP-style dual-encoder→ Visualization

Explanation of Image-HNC

Pos: The trees are behind the fence

Here fine-tuned CLIP is *HNC_KL_1_S_full*

Explanation of Image-HNC

Pos: The **trees** are behind the fence.

Base CLIP: “trees”

Fine-tuned CLIP: “trees”

Explanation of Image-HNC

Pos: The trees are behind the **fence**.

Base CLIP: “fence”

Fine-tuned CLIP: “fence”

→ Object detection get improved. E.g. trees, fence

Explanation of Image-HNC

Pos: The trees are **behind** the fence.

Base CLIP: “behind”

Fine-tuned CLIP: “behind”

→ For relational words (e.g. behind) can also focus on the right place in image.

Thank you