

Supplementary Material

No Captions, No Problem:

Captionless 3D-CLIP Alignment with Hard Negatives via CLIP Knowledge and LLMs

Cristian Sbrolli

www.deib.polimi.it/eng/people/details/1273522

Matteo Matteucci

www.deib.polimi.it/eng/people/details/267262

Department of Electronics, Information

and Bioengineering

Politecnico di Milano

Milano, IT

1 ULIP 2 captions limitations

In the paper, we discussed the limitations of ULIP 2 [1] LMM-generated captions for 2D representations of 3D samples, which are publicly available at the ULIP github repository.

In particular, we identified two categories of issues: those associated with 2D features not discernible on the 3D sample, and those arising from the application of the LMM itself.

- The first category of issues arises because 2D view captions include descriptions of attributes such as color, texture, and material, which cannot be utilized to align text with 3D since point clouds only contain spatial data.
- The second category of issues is intrinsically linked to the use of an LMM, as it tends to fill in missing information, create hallucinations, and exhibit bias towards the object category or general use.

We illustrate these problems in Figure 1, emphasizing both types of limitations and demonstrating that they apply to the majority of their captions, even though we only present two random samples here. While these captions allows for the training on text-3D couples, beneficial for the multimodal understanding, these issues affect the accurate alignment of text and 3D by aligning false or 3D-imperceptible information. As presented in the paper, this lead us to define the $(I2L)^2$ method, which allows us to leverage LMMs more effectively. With $(I2L)^2$, we use LMMs not for captioning but for generating a much smaller number of landmark texts, avoiding the color-texture-material bias and using the inventive/hallucinatory power of LMMs to our advantage. Indeed, this was a problem in captioning 2D views as each caption had to perfectly depict the object in the image, leading to misalignments due to hallucinations. However, in our case, this characteristic of LMMs is advantageous. It enables us to generate a variety of detailed texts that do not need to correspond exactly to individual images. Instead, they act as landmarks that a sample can more or less align with, and collectively, they serve to represent an image.

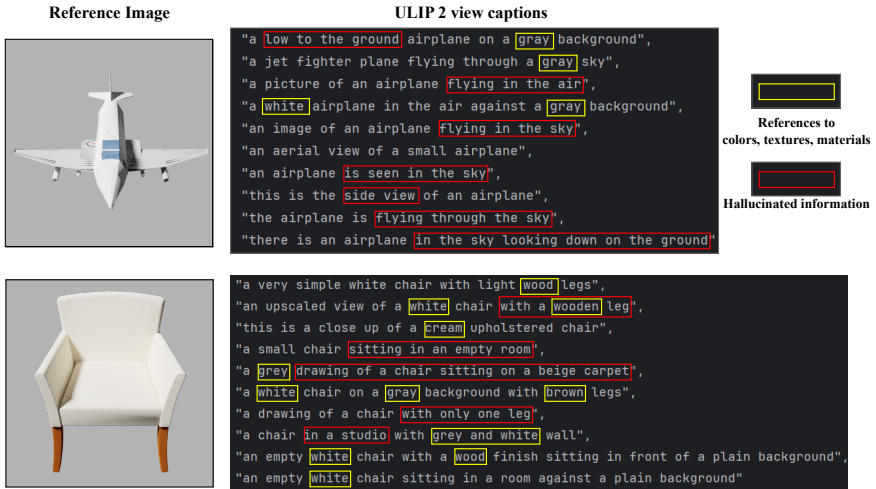


Figure 1: Limitations of ULIP 2 captions.

We report here few randomly sampled examples of our generated text landmarks for airplanes and chairs:

Airplanes

"A commercial jet with a sleek, aerodynamic body and two wing-mounted engines.",
 "A single-engine propeller plane with a high-wing design.",
 "A military fighter jet with a delta wing structure and afterburner capabilities.",
 "A vintage biplane with two stacked wings and open cockpit.",
 "A private jet with a low-wing configuration and retractable landing gear.",
 "A helicopter with a single main rotor and a tail rotor for directional control."

Chairs

"A four-legged seat with a straight back and no armrests, designed for one person.",
 "An office chair with a high back, adjustable height, and swivel base on wheels.",
 "A rocking chair with curved runners and a slatted back.",
 "A pub stool with a round, cushioned seat and circular footrest.",
 "A classic dining chair with a square seat, square back and four tapered legs.",
 "An elegant chair with a curved back and ornate carvings on the legs and arms."

These text landmarks illustrate a range of unique and relevant details that are specific to the individual object classes. Each landmark captures essential characteristics that distinguish one object type from another, whether through structural features (e.g., wing configuration in airplanes, leg design in chairs) or functional components (e.g., afterburners in jets, swivel bases in office chairs). This diversity allows the landmarks to collectively cover a broad spectrum of possible shapes and designs within each category.

2 EMD and CD for hard negative mining

In our initial studies, we tested Chamfer Distance (CD) and Earth Mover Distance (EMD) as 3D similarities for hard negative mining. As discussed in the paper, we were unable to achieve results as satisfactory as those we obtained with our proposed neural 3D similarities. In Figure 2, we present the results of the top-5 hardest retrieved shapes for the same example chair we discussed in the paper. While EMD and CD succeed in matching the overall shape

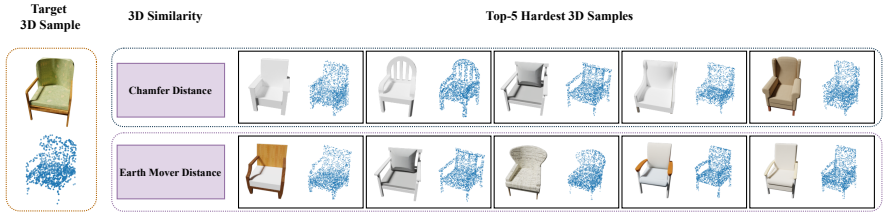


Figure 2: 3D-to-3D retrieval using EMD and CD on a chair sample from ShapeNet dataset..

structure of denser areas, they fail to retrieve category-specific relevant features, such as the shape and type of arms, legs, seat, and back in chairs. These metrics struggle to match elements like holes and curvatures, due to their inherent inability to capture finer-grained and less dense structures. Despite our dissatisfaction with the preliminary results on 3D hard negative mining, we trained our contrastive pipeline using these metrics following the same methodology outlined in the paper. We present the results in Tables 1 and 2. While the models derived using these metrics perform well in cross-modal retrieval, they fall short in both zero-shot and 3D classification when compared to state-of-the-art models. This performance gap served as a catalyst for the development of the similarities proposed in our study, which show competitive performance in classification and superior results in cross-modal retrieval.

	Zero-Shot 3D Classification				Standard 3D Classification	
	ModelNet40		ScanObjectNN		ModelNet40	ScanObjectNN
	top-1	top-5	top-1	top-5	top-1	top-1
Hard Negative Metric						
CD	62.1	81.0	50.7	76.2	93.5	88.8
EMD	61.6	80.8	49.9	77.4	93.1	87.3

Table 1: 3D classification results using CD and EMD to mine 3D hard negatives.

	No Background				Background			
	Image-to-Shape		Shape-to-Image		Image-to-Shape		Shape-to-Image	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
Hard Negative Metric								
CD	32.4	74.1	25.5	64.0	24.4	57.3	19.8	52.6
EMD	31.0	72.5	26.1	64.6	21.9	55.7	20.3	24.0

Table 2: Cross-modal retrieval results using CD and EMD to mine 3D hard negatives.

3 Ablation extended results

We report in Tables 3 and 4 the extended results for the ablation presented in the paper on the number of generated text landmarks L per category. As discussed in the paper, we observe a consistent trend of performance improvement with the increase in the quantity of texts. However, the improvement saturates when going beyond $L = 128$, and for this reason we adopt that value as it allows for an optimal balance between efficiency and performance.

	Zero-Shot 3D Classification				Standard 3D Classification	
	ModelNet40		ScanObjectNN		ModelNet40	ScanObjectNN
L	top-1	top-5	top-1	top-5	top-1	top-1
32	52.5	67.4	47.3	69.8	93.1	88.3
64	56.3	73.3	49.7	74.3	92.9	88.5
128	63.7	86.9	54.8	83.9	94.0	88.9
256	63.9	87.1	55.1	83.7	93.7	88.9
512	63.9	87.0	55.3	84.2	94.1	89.0

Table 3: Extended ablation results on landmark number L for 3D classification.

	No Background				Background			
	Image-to-Shape		Shape-to-Image		Image-to-Shape		Shape-to-Image	
L	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
32	20.5	61.9	16.7	52.1	18.1	47.1	14.6	46.6
64	24.8	65.7	19.9	58.4	19.7	56.5	20.5	51.8
128	33.3	74.4	26.8	66.8	25.3	59.9	22.2	55.5
256	34.0	74.8	27.5	66.0	25.7	59.9	22.5	56.4
512	34.4	74.6	28.1	67.1	25.3	59.7	22.5	56.2

Table 4: Extended ablation results on landmark number L for cross-modal retrieval.

References

- [1] Le Xue, Ning Yu, Shu Zhang, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. *arXiv preprint arXiv:2305.08275*, 2023.