Context-Aware 3D Object Retrieval Through VLLM-Based Scene Understanding

Kim Nguyen
Falcuty of Infomation Technololgy
University of Science
VNU-HCM
Ho Chi Minh City, Vietnam
0009-0001-4609-6381
23122040@student.hcmus.edu.vn

Quan Nguyen Hung
Faculty of Computer Science
University of Information Technology
VNU-HCM
Ho Chi Minh City, Vietnam
0009-0001-7294-5543
23521263@gm.uit.edu.vn

Dat Phan Thanh
Faculty of Computer Science
University of Information Technology
VNU-HCM
Ho Chi Minh City, Vietnam
0009-0001-3694-4623
23520267@gm.uit.edu.vn

Hoang Tran Van
Faculty of Computer Science
University of Information Technology
VNU-HCM
Ho Chi Minh City, Vietnam
0009-0007-9029-4442
23520542@gm.uit.edu.vn

Tien Huynh Viet
Faculty of Computer Science
University of Information Technology
VNU-HCM
Ho Chi Minh City, Vietnam
0009-0004-4356-2211
23521570@gm.uit.edu.vn

Nhan Nguyen Viet Thien
Faculty of Computer Engineering
University of Information Technology
VNU-HCM
Ho Chi Minh City, Vietnam
0009-0000-7068-6798
23521086@gm.uit.edu.vn

***Abstract*—Retrieving objects in cluttered 3-D scenes from a short natural language description remains difficult because simple queries are ambiguous and panoramic images conceal crucial spatial cues. We address this problem with a pipeline centred on a frozen vision language large model (vLLM) that directly reasons over the masked panorama and candidate objects, producing scene aware plausibility scores that capture stylistic harmony, functional fit, and geometric consistency. A lightweight cross-modal retrieval step supplies a small candidate pool, after which the vLLM re-ranks the objects by jointly analysing the scene context and the refined query. This scene-level assessment bridges the gap between underspecified language and complex visual layouts without any model fine-tuning or heavy 3-D reconstruction. The proposed method secured 4th place in the ROOMELSA Grand Challenge, confirming that a vLLM-centred design can deliver high accuracy while remaining efficient and easily reproducible. Ours code available on `Github`.**

***Index Terms*—Vision Language Model, Retrieval, Scene Understanding**

I. Introduction

Retrieving a single object in a cluttered 3-D scene from a brief natural language description is a deceptively difficult task. Short queries often omit crucial detail can map to several look-alikes while panoramic RGB-D views compound the problem with occlusion, depth distortion, and a vast field of distractors. These factors render straightforward keyword matching and appearance-only similarity unreliable, motivating techniques that can interpret both linguistic intent and spatial context.

To meet this need, we present a pipeline centred on a frozen vision language large model (vLLM) that treats the panorama, the candidate object, and the user query as a single reasoning unit. By prompting the vLLM with paired images and text, we obtain a plausibility score that reflects style, function, and geometric fit—capturing cues that simpler embeddings overlook.

Robustness is achieved through a compact yet powerful combination of modules: CLIP supplies an initial cross-modal shortlist, a lightweight linguistic filter distils the query and spawns paraphrases, and a frozen vLLM then judges each candidate in the full panoramic context. By fusing CLIP similarity with the vLLM's scene aware plausibility score, the pipeline rewards objects that are both visually consistent with the description and contextually coherent with the 3-D environment—delivering strong, competition-level accuracy without any model finetuning or heavy 3-D reconstruction. The proposed system secured 4th place in the 2025 ROOMELSA Grand Challenge, confirming that the thoughtful orchestration of frozen vision language models, targeted linguistic processing, and spatially informed reranking can deliver competitive performance against far heavier baselines in embodied object retrieval.

II. Methodology

Our retrieval system stands out for fully frozen, and context-aware: it requires no finetuning, and combines panoramic scene reasoning with precise text–image matching to resolve ambiguous queries. Concretely, we first sanitise the user description—stripping opinion adjectives and purpose phrases—and generate several paraphrases to capture alternative phrasings; both the original text and its paraphrases are embedded with a frozen CLIP encoder alongside every candidate image to produce an initial top-k shortlist in a shared feature space. Each short listed object is then evaluated by

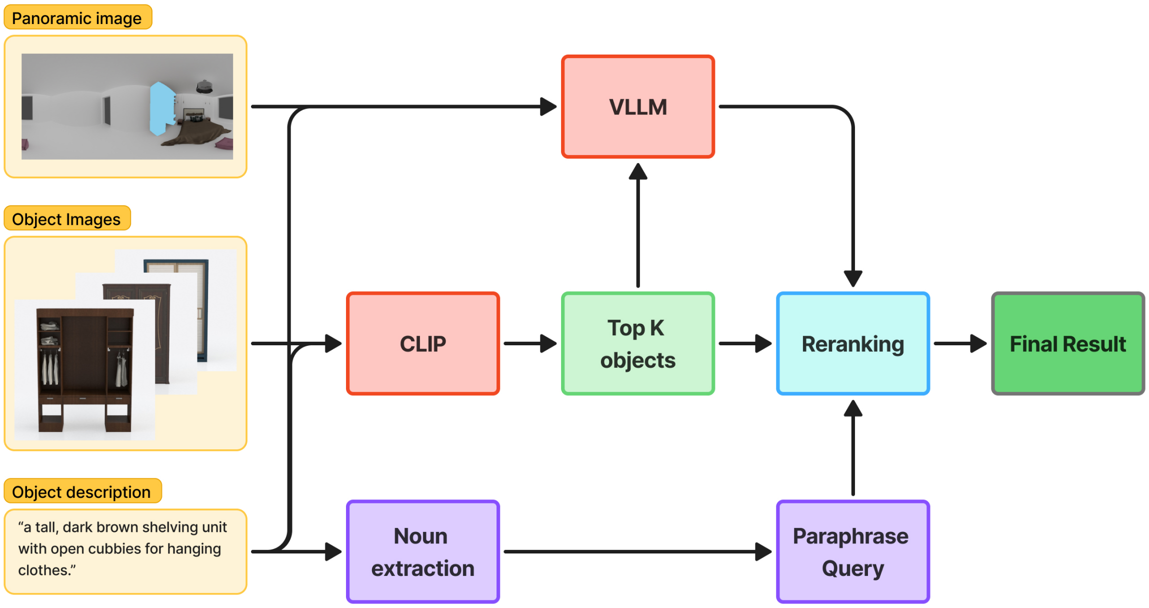

Fig. 1. Framework Architecture Overview

a frozen vLLM, which reads the masked panorama together with the candidate image in a single forward pass and outputs a plausibility score reflecting stylistic harmony, functional suitability, and spatial fit with the textual cues. A fusion of this VLLM score with the CLIP similarity yields the final ranking, marrying global scene context with local visual evidence while adding only minimal latency and consistently boosting mean reciprocal rank.

A. *CLIP-based retrieval module*

We adopt the OpenCLIP [1] ViT-SO400M-14-SigLIP-384 model, pretrained on the WebLI corpus, as our first layer of cross-modal reasoning. All catalogue images are embedded offline with this frozen encoder, and the user query—along with its paraphrases—is encoded at runtime into the same space. Computing a simple cosine similarity between text and image vectors yields a parameter-free relevance score for every object, from which we derive a top-k shortlist that captures coarse semantic alignment between language and vision. Because both weights and embeddings remain fixed, this stage scales linearly with catalogue size yet introduces only negligible latency, providing an efficient front-end before the vLLM-based scene re-ranking.

B. *Noun extraction and query paraphrasing*

Before retrieval, a shallow dependency parser removes opinion adjectives and purpose clauses from the query, isolating the head nouns that truly define the target object. These nouns seed a lightweight paraphrasing routine that generates several syntactically diverse yet semantically equivalent variants of the description. Encoding both the original sentence and its paraphrases widens lexical coverage, mitigates vocabulary mismatch with CLIP's training corpus, and measurably boosts recall in the top-k list all without introducing additional trainable parameters or appreciable computation overhead.

C. *VLLM-based scene understanding*

To eliminate the ambiguity that persists after appearance-driven retrieval, we employ the InternVL [2] 2.5-2B vision-language large model as a frozen, scene-level assessor. After the CLIP stage produces a shortlist of candidates, each object image is paired with the masked panorama and passed through InternVL [2] in a single forward pass. The model returns a scalar plausibility score that captures (i) stylistic harmony with surrounding materials and colours, (ii) functional suitability for the masked region, and (iii) geometric consistency with the spatial cues implied by the query. These scores are min–max normalised and linearly fused with the corresponding CLIP cosine similarities, yielding a final ranking that favours objects coherent with both the language description and the full 3-D context. The fusion consistently boosts mean reciprocal rank in our experiments, showing that—even in a frozen state—InternVL [2] 2.5-2B delivers the high-level scene reasoning needed to resolve underspecified queries without ad-

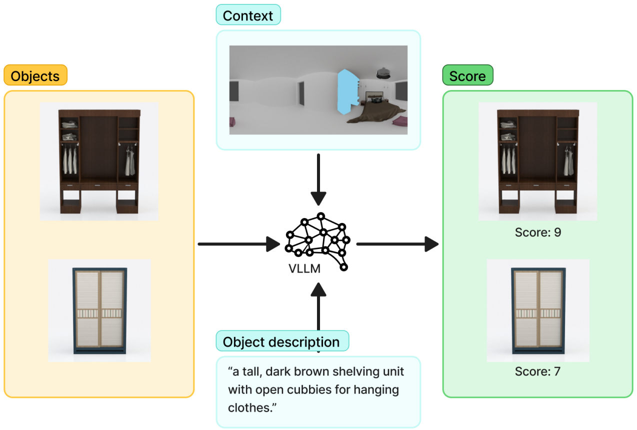
Context
Objects
Score
VLLM
Score: 9
Score: 7
Object description
"a tall, dark brown shelving unit with open cubbies for hanging clothes."

Fig. 2. Example of how we using VLLM and its result

ditional training, large memory overhead, or slow multi-stage processing.

III. Results

The ROOMELSA Grand Challenge is a groundbreaking competition that introduces a novel and impactful task in 3D environments. The challenge consists of retrieving a masked object in a 3D scene using only a natural language description and spatial context. Ours system has perform succesfull and achieve 4th place in the competetion

TABLE I
Results of the ROOMELSA Grand Challenge

Team Name	**R@1**	**R@5**	**R@10**	**MRR**
Stubborn_Strawberries	0.94	1.00	1.00	0.97
Ai-Yahh	0.92	1.00	1.00	0.96
MealsRetrieval	0.92	1.00	1.00	0.96
BUCCI_GANG (ours)	0.90	1.00	1.00	0.95
NoResources	0.88	1.00	1.00	0.93

IV. Conclusion

Our team presented a effective pipeline for 3D object retrieval using a vision-language large model (VLLM), addressing challenges like query ambiguity and the complexity of panoramic images. By combining CLIP-based retrieval, noun extraction, and query paraphrasing, followed by spatial-semantic reranking, our approach significantly improved retrieval accuracy, that achieving a 4th place placement in the ROOMELSA Grand Challenge with an MRR score of 0.95. Our method outperformed several more complex baselines, demonstrating that frozen VLLMs can achieve competitive performance while maintaining efficiency. This work paves the way for further advancements in real-time, resource-efficient 3D object retrieval.

References

[1] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image learning," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2023, p. 2818–2829. [Online]. Available: http://dx.doi.org/10.1109/CVPR52729.2023.00276

[2] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu, L. Gu, X. Wang, Q. Li, Y. Ren, Z. Chen, J. Luo, J. Wang, T. Jiang, B. Wang, C. He, B. Shi, X. Zhang, H. Lv, Y. Wang, W. Shao, P. Chu, Z. Tu, T. He, Z. Wu, H. Deng, J. Ge, K. Chen, K. Zhang, L. Wang, M. Dou, L. Lu, X. Zhu, T. Lu, D. Lin, Y. Qiao, J. Dai, and W. Wang, "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," 2025. [Online]. Available: https://arxiv.org/abs/2412.05271