Semantic-Augmented 3D Gaussian Splatting for Visual Localization in Complex Indoor Environments

Ba-Tuan-Hoang Chu and Gon-Woo Kim, *Member, IEEE*

***Abstract*—This letter presents a new visual localization framework for complex indoor environments under dynamic scene change conditions. Conventional visual localization methods often struggle to maintain accuracy and robustness in such environments, where frequent scene changes, occlusions, diverse object categories, and intricate scene structures significantly affect feature consistency and matching reliability. These challenges highlight the need for a more adaptive and semantically aware localization approach. By proposing an algorithm that integrates semantic information with a Gaussian map as input, the method enhances the algorithm's environmental awareness. This allows robust objects to be identified and extracted, thereby improving feature extraction performance and consequently enhancing pose estimation precision. Furthermore, a new coarse-to-fine matching strategy has been developed that takes an overview of the Gaussian map, from which suitable viewpoints are generated to produce the best matching images. Rendered images produced from the Gaussian map are employed in subsequent stages to improve comparison effectiveness, thereby enabling the determination of the most accurate camera pose. Finally, the capability of the proposed methodology is confirmed through experiments on different types of datasets.**

***Index Terms*—Visual localization, gaussian splatting, place recognition.**

I. Introduction

NOWADAYS, autonomous navigation has advanced significantly in the field of mobile robotics [1], where many challenging indoor tasks require high accuracy of robot trajectory and system stability. Accordingly, the capability to determine a robot's position with high precision is essential for the effectiveness of autonomous robotic operations. To track the location of the robots, visual localization emerges as a promising and robust solution for indoor applications. One of the effective approaches in vision-based methods is image retrieval techniques. A query image is compared against a pre-provided database of images with known poses, allowing the estimation of the camera position that captured the query image. A representative image retrieval technique is Absolute Pose Regression (APR), which employs deep neural networks to directly estimate the camera position from input images. A notable method is Marepo [2], which divides the training network into two parts: one for image feature extraction and the other for learning the relationship between the image and the camera pose. This modular design allows for better supervision over each component. As a result, the system avoids excessive storage and processing of information, thereby reducing its dependency on the scale of the environment. However, the extracted features are employed exclusively to learn the relationship between images and their corresponding camera poses, rather than being utilized as structural information for geometric reasoning about the scene. Thus, the algorithm lacks geometric awareness of the environment, which adversely affects the accuracy of pose estimation. Furthermore, its dependency on the training process makes the method prone to significant performance degradation when applied to complex indoor environments characterized by large spatial layouts, intricate structural compositions, diverse object types, and dynamic scene variations.

A structure-based localization method is Scene Coordinate Regression (SCR), which leverages the geometric structure of the environment for localization. This approach enables more accurate pose estimation compared to APR methods. A typical method of SCR is ACE [3], which designs a Convolutional Neural Network (CNN) to extract features and represent an implicit map. This allows for establishing 2D-3D correspondences between the query image and the implicit map. The enhanced understanding of the environment's geometry through the implicit map has contributed to improved localization accuracy. Nevertheless, the performance of the method depends heavily on the training dataset for each scene, resulting in limited generalization capability. Moreover, relying solely on the geometric structure of the environment is insufficient, particularly in complex indoor settings, where the diversity of scene composition can significantly affect the performance of feature extraction models due to the inherent limitations of the training dataset and CNN capacity. These factors collectively hinder the algorithm's ability to achieve accurate pose estimation under complex environmental conditions.

Furthermore, with the emergence of rendering techniques such as NeRF [4], several localization methods have integrated them to refine the output pose, including pNeRFLoc [5] and Crossfire [6]. These NeRF-based algorithms employ either APR or SCR methods to generate a coarse pose for the query image. Then, an iterative loop is subsequently applied, combining pose evaluation techniques with NeRF-based rendering to refine the initial pose and produce a more accurate final result. The integration of NeRF-based maps, which enrich the representation of appearance information from the environment, leads to theoretical improvements in localization performance over traditional

Received 28 August 2025; accepted 24 November 2025. Date of publication; date of current version. This article was recommended for publication by Associate Editor D. Belter and Editor S. Behnke upon evaluation of the reviewers' comments. This work was supported in part by the National Research Foundation of Korea (NRF) through the Korea government (MSIT) under Grant RS-2025-00561031, 50, in part by the Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation (IITP), and in part by the Korea government (MSIT) under Grant IITP-2025-RS-2020-II201462, 50. *(Corresponding author: Gon-Woo Kim.)*

The authors are with the Intelligent Robotics Laboratory, Department of Intelligent Systems and Robotics, Chungbuk National University, Cheongju 28644, South Korea (e-mail: 2024298018@chungbuk.ac.kr; gwkim@cbnu.ac.kr).

Digital Object Identifier 10.1109/LRA.2025.3643336

SCR and APR approaches. Despite the refinement process, using APR or SCR as initial inputs leaves unresolved issues related to limited generalization and suboptimal performance in complex indoor environments.

Although NeRF-based rendering enables the generation of novel-view images, it is a technique that is susceptible to noise and sensitive to environmental lighting conditions, which results in inconsistent image quality. Besides, the rendering process of NeRF requires substantial computational resources, limiting its applicability in localization tasks. A recently developed rendering technique, Gaussian Splatting (GS) [7], has demonstrated significant improvements in rendered image quality compared to NeRF. The method encodes the map as modeled 3D Gaussians, enabling optimization and rendering processes to be executed directly on this compact Gaussian mode. Representative GS-based methods, such as GS-CPR [8], STDLoc [9], iComMa [10], and SplatLoc [11], have effectively leveraged the advantages of GS within their algorithms. In particular, by utilizing high-quality rendered images for localization feature extraction, these approaches have demonstrated a significant improvement in pose estimation performance compared to NeRF-based methods. Thus, the substitution of NeRF with GS can be viewed as a promising direction for enhancing localization frameworks. However, issues related to generalization and handling complex environments remain unresolved when relying solely on GS.

Motivated by the aforementioned problems, the objective of this letter is to develop a robust framework for the complex indoor environment. In this study, a novel visual localization technique relied on an effective combination of Gaussian maps with semantic information and a new robust coarse-to-fine matching technique. The main contributions of this work are summarized as follows:

- This work presents a visual localization framework that integrates semantic information and employs GS techniques to enhance environmental awareness. By utilizing semantic information, the algorithm can avoid relying on training models, enabling more general applicability and effective operation in complex environments.
- A robust coarse-to-fine matching strategy has been developed to enhance the preliminary pose estimation of the reference database. This approach enables the identification and selection of optimal viewpoints for rendering images that closely correspond to the query image, which in turn significantly enhances the accuracy and robustness of the localization model.
- The performance of the proposed pipeline is thoroughly investigated on different datasets of complex indoor environments, ranging from simple to complicated scenes. The results demonstrate superior performance compared to state-of-the-art (SOTA) methods.

The rest of this article is organized as follows; Section II provides an overview of the pipeline; Section III details the preprocessing module; Section IV describes the place recognition module; Section V discusses the alignment module. The experimental results are presented in Section VI. Finally, Section VII concludes the article.

II. Overview

This section presents the visual localization pipeline, which takes input from a current query image and a set of pre-provided RGB images in complex indoor environments. However, it is difficult to achieve high-accuracy localization in such complicated scenes. An innovative solution is to classify and identify robust objects to enhance the perception of surrounding spaces. The Gaussian map is applied to augment the understanding of the geometry and appearance of the observed area. Additionally, semantic information is integrated into the Gaussian map to prioritize objects that are distinctive, stable, and exhibit minimal change over time. Then, a large language model (LLM) [12] is employed to perceive the scene context. Based on this, a robust Gaussian submap is established to estimate the initial position of the query image. This serves as the first step of the place recognition module. By leveraging the continuous scene representation capability of GS [7], a coarse-to-fine matching strategy is designed to generate a preliminary pose for the query image. Finally, the pose refinement process executes an iterative loop of position estimation using the Perspective-n-Point (PnP) algorithm [13], combined with image rendering from the Gaussian map. This loop minimizes the error between the predicted and actual poses, resulting in a final pose with high accuracy. The algorithm flowchart is illustrated in Fig. 1.

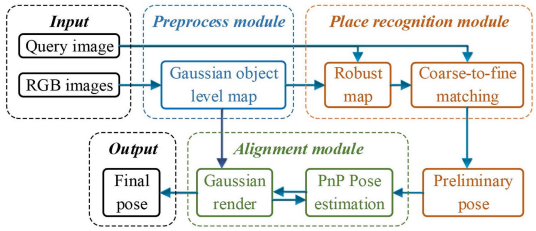


Fig. 1. Proposed framework.

III. Preprocessing Module

The preprocessing module constructs a Gaussian object-level environmental map to enhance contextual understanding by integrating both geometric and semantic information.

The Open-Gaussian algorithm [14] is utilized to generate a Gaussian semantic map from the provided RGB images. This map represents the scene as a collection of 3D GSs.

The object-level map is built to enable efficient retrieval of semantic information from the Gaussian map. The GSs are clustered by spatial and semantic similarity to form a Gaussian object-level map, represented by the following equation:

$$M_o = \{o_j\} \tag{1}$$

where $o_j = (G_k^o | k \in C_j)$ is the object j^{th} constructed by corresponding GSs G_k^o; C_j represents the index set of Gaussians associated with a specific object j^{th}.

IV. Place Recognition Module

A. Build Robust Map

As previously mentioned, complex indoor environments contain diverse objects, repetitive structures, and dynamic changes, which can degrade visual localization if objects are unidentified or unclassified. Therefore, a robust map is created to recover stable objects. Given a query image, the LLM [12] interprets the environment and identify stable objects within it. The resulting semantic information is then incorporated into the Gaussian object-level map to retrieve the corresponding 3D objects, thereby forming a robust map representation.

A list of robust objects from the query image is obtained based on contextual information processed by the LLM through

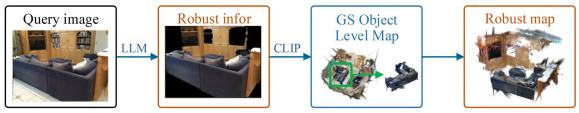


Fig. 2. Robust map is created based on the Gaussian object-level map and the query image.

a prompt-based setup. Using this list, the object-level map is queried to identify and extract the corresponding robust 3D objects with the support from CLIP [15], thereby constructing a robust submap. The process of developing the robust map is illustrated in Fig. 2 for better visualization.

B. Coarse-to-Fine Matching

The robust map provides the query's scene location, but this alone is insufficient for the PR module to estimate its preliminary pose. Traditional image-based localization, which relies on 2D feature matching, often fails in complex environments due to limited scene understanding. To address this, the GS rendering is utilized to generate reference images from the location indicated by the robust map. These are then compared with the query to estimate the preliminary pose. A coarse-to-fine matching strategy is introduced to improve both computational efficiency and output quality.

In the coarse matching phase, overview-shot and viewpoint filtering are designed to create viewpoints for GS rendering. The reference images produced by these steps are then compared with the query image to obtain the best coarse match.

In the fine matching phase, view enhancement refines the viewpoint based on the coarse best match to produce a new set of reference images that closely approximate the query image.

C. Coarse-to-Fine: Coarse-Overview Shot

The overview shot step generates evenly distributed viewpoints around the robust map to identify the direction of the view closest to the query image's pose.

To achieve the distributed viewpoints, a spherical observation model is made to cover around the robust map. This model is expressed as follows:

$$Sphere = (\Omega, r) \tag{2}$$

in this context, $\Omega = (|\gamma|)^{-1} \sum \mu_\gamma$ is the center coordinate of the sphere; $r = \max \|\mu_\gamma - \Omega\| + \delta$ is the radius of the sphere; γ is the index value of GSs belong to robust map; $\mu_\gamma \in \mathbb{R}^3$ is the mean of GSs belong to robust map; $\delta \in \mathbb{R}$ is a small offset.

To obtain a set of uniformly distributed viewpoints on the sphere, the Fibonacci sphere method [16] is employed. The rendering reference images derive from these viewpoints. In order to minimize operational resources, the number of viewpoints is adjusted to a sparsified value, while remaining sufficiently dense to identify a view direction that closely aligns with the query image's pose.

Remark 1: When the 3D GS map quality is degraded, features extracted from rendered images may be unreliable, reducing pose estimation accuracy. Observing the map through an overview shot helps identify viewpoints closer to the query, enabling more accurate feature extraction and mitigating map quality issues. This step significantly improves the algorithm's robustness.

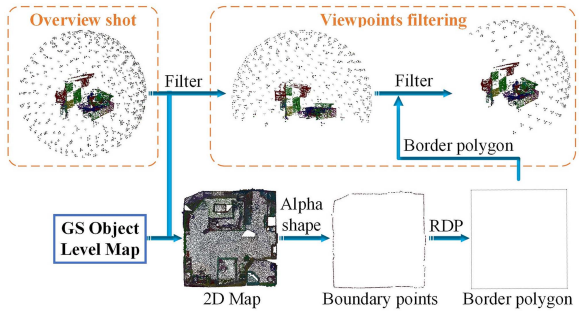


Fig. 3. Process of filtering outlier viewpoints.

D. Coarse-to-Fine: Coarse-Viewpoint Filtering

The coarse-overview shot generates viewpoints around the robust map, but some are misaligned or even opposite to the pre-provided poses, producing low-quality rendered images. Therefore, a viewpoint filtering method is developed to remove such outlier viewpoints, ensuring stable rendering quality. Viewpoints that violate map boundaries, such as the ground plane or spatial perimeter, are excluded based on boundary constraints to maintain the reliability of the analysis process.

Semantic ground information is extracted from the object-level map to estimate the ground plane. The map is then projected onto the ground plane to form a 2D representation, where the Alpha Shape [17] and Ramer–Douglas–Peucker (RDP) [18] algorithms are applied to identify and simplify boundary points, resulting in an enclosing polygon. This ground plane and the derived polygon are used to filter out inappropriate viewpoints through spatial vector orientation analysis. The process effectively eliminates viewpoints located outside the map domain or below the ground plane. Fig. 3 illustrates the detailed process of filtering outlier viewpoints.

The polygon filter is formulated as follows:

$$\begin{aligned} \partial : (F_{v_x} - \psi_x)N_{\partial_x} + (F_{vy} - \psi_y)N_{\partial_y} + (F_{v_z} - \psi_z)N_{\partial_z} < 0 \\ \forall A \in P : (F_{v_x} - \rho_x)N_{A_x} + (F_{v_y} - \rho_y)N_{A_y} < 0 \end{aligned} \tag{3}$$

where $\partial \in \mathbb{R}^3$ is the ground plane; N_∂ is the normal vector perpendicular with plane ∂; ψ is the center of the plane ∂; P is polygon; $A \in \mathbb{R}^2$ is edge of polygon; $F_v \in \mathbb{R}^2$ is the viewpoint coordinate; $\rho \in \mathbb{R}^2$ is midpoint of edge A; N_A is the normal vector perpendicular with edge A.

E. Coarse-to-Fine: Coarse-Correspondence Comparison

The correspondence comparison follows query-driven visual localization principles using extracted image features. Two types of features are used to determine similarity: local and global features. ALIKE [19], trained on large-scale datasets, provides robust local feature detection across diverse environments, extracting numerous features from both reference and query RGB images. To match these features accurately and efficiently, LightGlue [20] is employed, offering high-speed feature matching while preserving accuracy.

After matching, corresponding 2D keypoints are identified and used to compute a local feature similarity score based on the number of reliable matches. Although ALIKE is trained on large-scale datasets, it can still extract unstable features in complex scenes, leading to errors. To address this, a robust

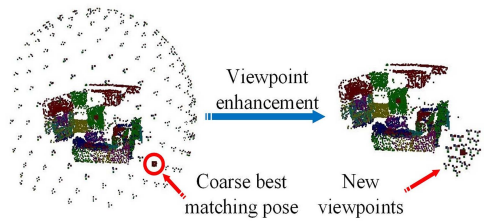


Fig. 4. Viewpoint enhancement process.

masking strategy removes low-confidence features, enhancing the reliability of the results.

In the robust map construction stage, robust information is reused to guide segmentation using Fast-SAM [21] which generates a robust mask for the query image. Matched key points are then classified as reliable or unreliable based on their position within this mask, with those in robust regions given higher weights to strengthen the impact of reliable features in computing local feature similarity.

The score for the local feature is calculated by:

$$L_\varepsilon = \sum_{\kappa=1}^{n_\varepsilon} w_\kappa \text{where} w_\kappa = \begin{cases} w_{in} = 2 \\ w_{out} = 1 \end{cases} \tag{4}$$

in this context, $n_\varepsilon \in \mathbb{N}$ is the total number of matching point pairs; $L_\varepsilon \in \mathbb{N}$ is the local features; w_κ is the weight of each κ^{th} matching pair; w_{in} and w_{out} indicate the matching point is in the mask region or not, respectively.

Local features capture regional similarities between image pairs but can become ambiguous in complex scenes with many candidates. Thus, global feature matching is also incorporated to provide a more holistic image representation and improve overall matching accuracy.

Features from the last layer of ALIKE's backbone are extracted as global descriptors, and their similarity is assessed using the cosine method. Then, the similarity score for each image pair is computed as follows:

$$Score_\varepsilon = \alpha \frac{L_\varepsilon}{\arg\max(L)} + (1-\alpha)\frac{H_\varepsilon}{\arg\max(H)} \tag{5}$$

where $L_\varepsilon \in \mathbb{N}$ is the score of local features in (4); $H_\varepsilon \in \mathbb{R}$ is the score of the global features; $\alpha \in (0;1)$ is the gain value.

The best matching pair is the one with the highest score, from which the initial pose of the query image is obtained.

F. Coarse-to-Fine: Fine-Viewpoint Enhancement

From the coarse matching stage, the best matching pair is identified to estimate the query image's initial pose. However, since reference images are rendered from sparsely distributed viewpoints around the robust map, the best match may still differ from the query pose, causing errors in later stages. The fine matching stage refines this pose by generating additional reference images from viewpoints near the initial estimate, increasing visual similarity to the query and improving overall matching accuracy.

To perform this refinement, new viewpoints are distributed around the initial pose estimate, forming a hemispherical pattern using the Fibonacci sphere algorithm [16]. This approach resembles the overview shot strategy but uses denser sampling within a smaller spatial region. Reference images rendered from these viewpoints are then compared with the query to estimate the refined pose, as illustrated in Fig. 4.

V. Alignment Module

In this section, an alignment module is introduced to perform the final refinement of the preliminary pose, ultimately producing the final pose for the query image.

The alignment process is an iterative loop involving pose estimation using the PnP algorithm [13] and environmental information obtained from rendered images of the GS map. Firstly, the 2D points on the best match rendered image are projected back to 3D space using the depth image:

$$X_\kappa = {R_{coarse}}^{-1}.\left(D\left(\chi_\kappa\right).K^{-1}.[\chi_\kappa \quad 1]^T - \tau_{coarse}\right) \tag{6}$$

where $X_\kappa \in \mathbb{R}^3$ represents the 3D coordinate corresponding to the 2D pixel $\chi_\kappa \in \mathbb{R}^2$ of each κ^{th} matching pair; $R_{coarse} \in \mathbb{R}^{3\times3}$ denotes the rotational component of the input preliminary pose; $D(\chi_\kappa)$ is the depth value extracted from depth image; $\tau_{coarse} \in \mathbb{R}^3$ is the translation matrix.

The depth value per pixel is rasterized by alpha-blending according to the following formula:

$$D = \sum T_\kappa G'_\kappa(\chi_\kappa|\mu',\Sigma')\sigma_\kappa z_\kappa \tag{7}$$

where $z_\kappa \in \mathbb{R}$ is the distance to the mean of Gaussian; T_κ is the transmittance coefficient; G'_κ is the 2D gaussians; $\mu' \in \mathbb{R}^2$ and $\Sigma' \in \mathbb{R}^{2\times2}$ are 2D mean and covariance of Gaussian.

The PnP algorithm utilizes 3D–2D point correspondences to estimate an updated and more precise pose for the query image. This refined pose is then used to re-render the scene, from which new 2D–3D correspondences are extracted for subsequent PnP estimation. Through this iterative procedure, a pose alignment loop is established for the query image.

VI. Experimental Result

This section evaluates the performance of the proposed method in solving the visual localization problem under complex indoor conditions. The method is implemented and performed by a computer equipped by a Core-i7 12th generation CPU, 32 GB ram, and NVIDIA Geforce RTX 4070 GPU. The evaluation focuses on two key criteria: robustness and accuracy. A detailed description of the evaluation setup is provided in the following subsection.

A. Evaluation Setup

Dataset: Experiments are conducted on three datasets: 7Scenes [22], ScanNet [23], and a custom dataset, covering indoor environments from simple to complex conditions, including static small-scale, static large-scale, and semi-dynamic scenes.

Evaluation metrics: Two types of metrics are used to evaluate performance: Average translation error (ATE) and average rotation error (ARE); Success rate: This metric indicates the percentage of query images whose estimated poses fall within standard error thresholds, 5 cm / 5° and 2 cm / 2°, commonly used in visual localization benchmarks.

Baseline: To validate the effectiveness of the proposed method, comparison against representative visual localization approaches, including: APR method: Marepo [2]; SCR method: ACE [3]; NeRF-based methods: CROSSFIRE [6] and

TABLE I
MEDIAN ROTATION AND TRANSLATION ERRORS (°/CM) ACROSS 7SCENE ↓

Methods	Chess	Fire	Heads	Office	Pumpkin	Redkitchen	Stairs	Avg. Error
Proposed[1]	0.15/**0.38**	0.26/0.61	**0.24/0.37**	**0.2/0.48**	**0.18/0.72**	0.2/0.75	0.43/**1.35**	**0.24/0.67**
Proposed[2]	0.16/0.5	0.27/0.64	0.34/0.48	0.26/0.9	0.21/0.9	0.2/0.85	0.53/1.6	0.28/0.84
Marepo	0.83/1.9	0.92/2.3	1.24/2.1	0.93/2.9	0.88/2.5	0.98/2.9	1.48/5.9	1.04/2.93
ACE	0.18/0.5	0.33/0.8	0.33/0.5	0.29/1	0.22/1	0.2/0.8	0.81/2.9	0.34/1.07
CrossFire	0.4/1	1.9/5	2.3/3	1.6/5	0.8/3	0.8/2	1.9/12	1.39/4.43
pNeRFLoc	0.8/2	0.88/2	0.83/1	1.05/3	1.51/6	1.54/5	5.73/32	1.76/7.29
GS-CPR[1]	0.15/0.5	0.25/0.6	0.28/0.4	0.26/0.9	0.23/1	**0.17**/0.7	0.42/1.4	0.25/0.8
SplatLoc[1]	0.71/1.89	0.9/2.3	0.52/0.92	0.82/1.91	1.12/2.54	2.2/4.21	4.78/17.2	1.57/4.42
STDLoc[1]	0.15/0.46	**0.24/0.57**	0.26/0.45	0.24/0.86	0.21/0.93	0.19/**0.63**	**0.41**/1.42	**0.24**/0.76
iComMa[1]	**0.14**/0.66	0.26/0.6	0.28/0.53	0.3/0.87	0.22/0.93	0.18/0.75	0.56/1.83	0.28/0.88
GS-CPR[2]	0.21/1.62	0.37/1.12	0.37/0.78	0.68/1.47	0.31/2.3	0.22/1.75	0.68/2.64	0.41/1.67
SplatLoc[2]	1.23/3.67	1.66/3.75	0.57/1.53	1.56/2.72	1.55/3.85	2.45/5.6	4.54/21.6	1.94/6.1
STDLoc[2]	0.16/1.08	0.33/1.65	0.43/0.92	0.48/1.37	0.36/2.08	0.54/1.12	0.91/2.57	0.46/1.54
iComMa[2]	0.17/1.35	0.42/0.98	0.45/1.23	0.66/1.74	0.31/2.12	0.32/1.9	0.76/2.59	0.44/1.7

(*)[1]: Indicates GS-based methods with high-quality GS map input; (*)[2]: Indicates GS-based methods with low-quality GS map input; Bold values indicate the best results; ↓: indicates that lower values correspond to better accuracy.

TABLE II
SUCCESS RATE (%) OF FRAMES MEETING A [5°, 5CM], [2°, 2CM] POSE ERROR THRESHOLD ACROSS 7SCENE↑

Threshold	Proposed	Marepo	ACE	CrossFire	pNeRFLoc	GS-CPR	SplatLoc	STDLoc	iComMa
[5°, 5cm]	**100[1]/98.8[2]**	84	97.1	81.6	73.3	**100[1]**/97.9[2]	76.3[1]/71.1[2]	99.1[1]/96.7[2]	97.5[1]/94.4[2]
[2°, 2cm]	**96.7[1]/93.3[2]**	33.7	83.3	31.3	26.6	93.1[1]/92.2[2]	28.8[1]/25.5[2]	90.9[1]/87.8[2]	88.8[1]/85.6[2]

(*)[1]: Indicates GS-based methods with high-quality GS map input; (*)[2]: Indicates GS-based methods with low-quality GS map input; Bold values indicate the best results; ↑: indicates that higher values correspond to better accuracy.

pNeRFLoc [5]. GS-based methods: GS-CPR [8], STDLoc [9], iComMa [10] and SplatLoc [11].

B. First Scenario: 7Scene Dataset

This scenario evaluates the proposed method in a small, complex indoor environment and analyzes the impact of GS map quality on GS-based localization, with results reported in Tables I and II. NeRF-based methods such as CROSSFIRE and pNeRFLoc exhibit low accuracy and success rates due to their reliance on additional components (e.g., descriptors and warping losses), whose limited robustness constrains localization performance. In contrast, ACE and Marepo demonstrate higher effectiveness owing to meta-learning, which improves generalization to unseen scenes, particularly on the 7Scenes dataset. GS-based methods are further evaluated using two map qualities: a high-quality GS map built from about 1000 RGB images and a low-quality one from around 300. As shown in Table I, all methods benefit from high-quality maps, with the proposed method achieving the best results in five sequences. When using low-quality maps, SplatLoc suffers a significant performance drop because it does not leverage differentiable rendering. In contrast, GS-CPR, STDLoc, and iComMa achieve better accuracy by exploiting GS rendering. Notably, the proposed method consistently outperforms other GS-based approaches even with low-quality maps, demonstrating the robustness of the proposed pipeline.

As discussed above, GS-based methods show degraded performance with low-quality maps. To highlight the robustness of the proposed method, the second and third scenarios are therefore evaluated exclusively using low-quality maps.

The rendered image from the final pose estimated by the proposed method is shown in Fig. 5. As previously noted, the proposed method achieves high accuracy; consequently, the rendered images closely resemble the query images, even in scenes containing multiple objects.

As discussed above, GS-based methods exhibit reduced performance when using low-quality map inputs. Therefore, to highlight the robustness of the proposed method, the experiments in the second and third scenarios are conducted exclusively with low-quality map inputs.

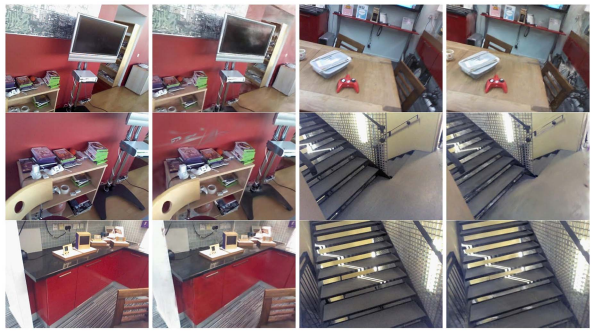

(a) Query RGB (b) Final pose (a) Query RGB (b) Final pose

Fig. 5. Results of the Proposed method - 7scene dataset.

The rendered image from the final pose estimated by the proposed method is shown in Fig. 5. As previously discussed, the proposed method achieves a high level of accuracy. Consequently, the rendered output images closely resemble the query images. Even in cases where the scene contains multiple objects, the proposed algorithm consistently produces rendered images with high accuracy.

C. Second Scenario: ScanNet Dataset

This section evaluates the algorithm's localization performance in a more complex environment with greater diversity and larger scale than the 7-Scenes dataset. Since NeRF-based methods show subpar performance in the first scenario, they are excluded from this comparison, and the average localization errors are shown in Table III. Marepo exhibits the highest error due to inaccurate scene coordinate predictions caused by noisy map construction and frequent occlusions in ScanNet, as well as the absence of explicit 2D.3D correspondences. Although ACE shares the same backbone and achieves better results, it also suffers from the lack of explicit 2D.3D matching, limiting its performance in large-scale environments. SplatLoc shows large

TABLE III
THE MEDIAN ROTATION AND TRANSLATION ERRORS (°/CM) ON SCANNET ↓

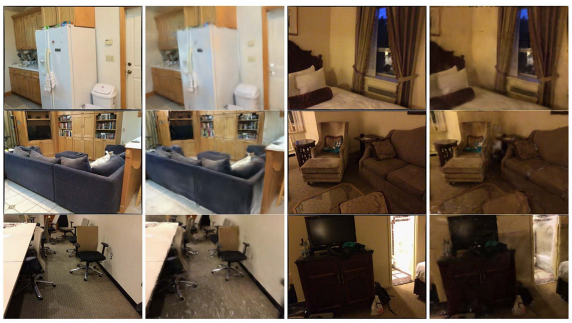
Methods	Scene0000	Scene0140	Scene0645	Avg. Error
Proposed[2]	**0.49/2.53**	0.77/**1.78**	**0.25/0.91**	**0.5/1.74**
Marepo	3.9/24.63	2.47/31.9	2.1/26.73	2.82/27.75
ACE	1.2/15.22	0.5/22.08	0.43/18.33	0.71/18.54
GS-CPR[2]	1.1/13.7	0.5/14.23	0.52/14.67	0.71/14.2
SplatLoc[2]	2.32/17.9	3.14/22.71	3.47/25.87	2.98/22.16
STDLoc[2]	1.1/12.4	**0.48**/13.21	0.53/12.1	0.7/12.57
iComMa[2]	1.2/14.88	**0.48**/19.8	0.41/17.81	0.7/17.5

(*)[2]: Indicates GS-based methods with low-quality GS map input; Bold values indicate the best results; ↓: indicates that lower values correspond to better accuracy.

TABLE IV
THE SUCCESS RATE (%) OF FRAMES MEETING A [5°, 5CM], [2°, 2CM] POSE ERROR THRESHOLD ON SCANNET↑

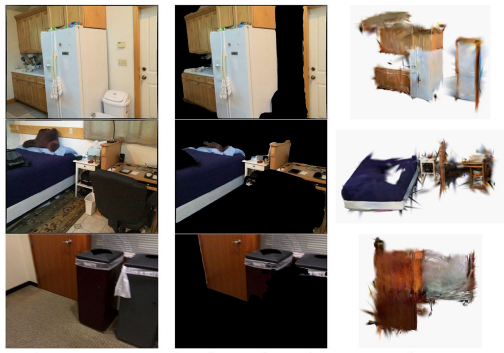
Threshold	Pro-posed[2]	Mare po	ACE	GS-CPR[2]	Splat Loc[2]	STDL oc[2]	iCom Ma[2]
[5°, 5cm]	**97.5**	68.3	85.0	89.4	76.7	91.3	88.3
[2°, 2cm]	**95.0**	31.7	51.7	72.1	52.2	74.6	68.3

(*)[2]: Indicates GS-based methods with low-quality GS map input; Bold values indicate the best results; ↑: indicates that higher values correspond to better accuracy.



(a) Query RGB (b) Final pose (a) Query RGB (b) Final pose

Fig. 6. Results of the Proposed method across ScanNet dataset.



(a) Query RGB (b) Robust Information (c) Robust map

Fig. 7. Robust map generation results of the proposed method across ScanNet dataset.

(a) Scene at mapping time (b) Scene at query time

Fig. 8. Describe the environment and scene change.

TABLE V
MEDIAN ROTATION AND TRANSLATION ERRORS (°/CM) ACROSS CUSTOM DATASET ↓

Methods	Pro-posed[2]	Mare po	ACE	GS-CPR[2]	Splat Loc[2]	STD Loc[2]	iCom Ma[2]
Avg. Error	**2.4/ 5.87**	7.93/ 22.6	5.3/ 16.2	4.1/ 15.3	8.02/ 20.7	3.83/ 15.8	4.9/ 16.1

(*)[2]: Indicates GS-based methods with low-quality GS map input; Bold values indicate the best results; ↓: indicates that lower values correspond to better accuracy.

positional errors because sparse and geometrically inconsistent 3D features lead to inaccurate 2D–3D correspondences. While GS-CPR and iComMa achieve improved localization through enhanced feature matching but still suffer from residual errors due to the lack of explicit 3D geometry modeling during feature construction. In contrast, STDLoc improves robustness through hierarchical geometric feature extraction and correspondence strategy. The proposed method achieves the highest accuracy by integrating geometric information from the GS map with semantic cues, enabling more accurate 2D.3D correspondences and superior localization performance.

The success rates of the compared algorithms are presented in Table IV. ACE, Marepo, and SplatLoc exhibit low success rates due to their limitations in accurately establishing 2D–3D correspondences, as discussed above. In contrast, GS-CPR, iComMa, and STDLoc achieve improved performance by leveraging GS, although they still face challenges in feature processing in complex environments. Notably, the proposed algorithm attains a superior success rate exceeding 95%.

The image rendered from the final pose estimated by the proposed method is shown in Fig. 6. As highlighted earlier, this method demonstrates a notably high success rate. Accordingly, the rendered images closely resemble the corresponding query images, even on datasets with large-scale, diverse, and complex scenes.

An illustration in Fig. 7 highlights the robust map construction process. Stable elements such as wardrobes, beds, doors, and shelves are successfully extracted, while objects prone to positional changes (e.g., trash bins, pillows) are excluded. Robust maps constructed from this information achieve stable accuracy using the GS technique, thereby contributing to improved localization performance, as described in the proposed method.

The image generated from the final pose estimated by the proposed approach is shown in Fig. 6. As highlighted earlier, this method demonstrates a notably high success rate. Accordingly, the rendered images exhibit a strong resemblance to the corresponding query images, even on datasets with large areas, diverse scenes, and complex spatial configurations.

D. Third scenario: Custom Dataset

The custom dataset was collected in a 75 m classroom, as shown in Fig. 8, under semi-dynamic conditions, where query images were captured after object rearrangements. Compared to ScanNet, these changes reduce feature consistency in robust regions and increase localization difficulty, as summarized in Table V. ACE, Marepo, and SplatLoc show suboptimal performance on ScanNet, which further degrades on the more dynamic and challenging custom dataset. GS-CPR, iComMa, and STDLoc show improved results but remain limited by purely geometric perception. In contrast, the proposed method achieves the highest accuracy by leveraging semantic information to focus on robust regions.

The success rates of the compared methods are reported in Table VI. ACE, Marepo, and SplatLoc exhibit the lowest performance, while GS-CPR, STDLoc, and iComMa achieve moderate improvements. In contrast, the proposed method delivers significantly superior results, exceeding an 80% success

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE VI
THE SUCCESS RATE (%) OF FRAMES MEETING A [10°, 10CM], [5°, 5CM], [2°, 2CM] POSE ERROR THRESHOLD ACROSS CUSTOM DATASET ↑

Threshold	Pro-posed[2]	Mare po	ACE	GS-CPR[2]	Splat Loc[2]	STD Loc[2]	iCom Ma[2]
[10°, 10cm]	**93.3**	55.1	71.6	76.3	63.3	77.5	75.5
[5°, 5cm]	**81.67**	38.3	61.7	63.8	52.2	68.8	62.2
[2°, 2cm]	**61.7**	21.2	23.3	38.8	22.2	41.3	32.2

(*)[2]: Indicates GS-based methods with low-quality GS map input; Bold values indicate the best results; ↑: indicates that higher values correspond to better accuracy.

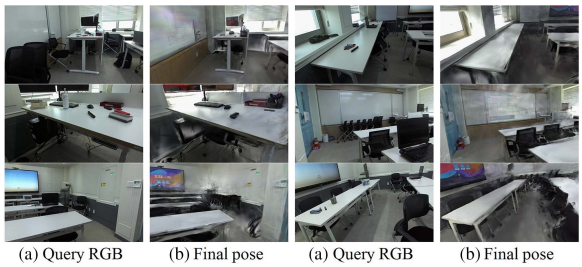


Fig. 9. Results of the Proposed method across custom dataset.

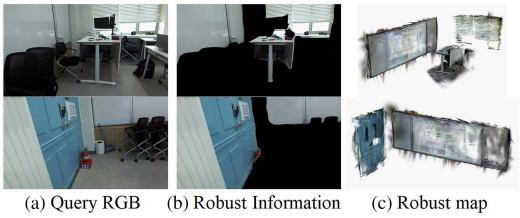


Fig. 10. Robust map generation results of the proposed method across custom dataset.

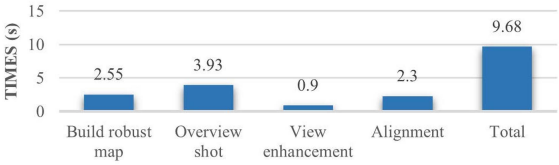


Fig. 11. Runtime analysis for the proposed method.

rate under the 5 cm and 5° threshold, demonstrating strong robustness to scene changes.

Fig. 9 illustrates the robust map construction process, showing that the system successfully identifies stable objects such as whiteboards, windows, desks, and doors using semantic information, while excluding changeable items like chairs, pens, and books. The resulting robust map enables consistent information extraction in dynamic scenes and provides reliable initialization for subsequent matching and alignment stages, directly impacting final pose estimation. Fig. 10 further confirms that the rendered views closely align with the query images despite scene changes.

E. Ablation Study

This section presents ablation studies to evaluate the contributions of key components in this pipeline on ScanNet dataset. Seven configurations are designed to isolate the main factors:

Running time: Although real-time performance is not the primary goal, runtime analysis is conducted to assess efficiency and identify bottlenecks. Fig. 11 presents a detailed runtime breakdown. The proposed algorithm, designed for accuracy and robustness in offline settings, has an average execution time of 9.68 s. The main computational costs arise from the overview shot stage and robust map construction. Therefore, reducing the processing time of these stages is critical. One possible direction is to leverage geometric information from objects to estimate relative poses more efficiently, instead of relying on the overview shot stage. In addition, the current LLM and CLIP implementations used for robust map construction are not yet fully optimized. Future work will focus on improving their integration and efficiency to reduce the overall runtime.

TABLE VII
ABLATION STUDY RESULTS FOR THE PROPOSED METHOD WITH SCANNET

Configurations	Avg. Err. (°/cm)↓	Success rate [5°, 5cm]↑	Success rate [2°, 2cm]↑	Runtime (s)↓
Avg. Error	**0.5/1.74**	**97.5**	**95**	9.68
W/o semantic	7.6/33.7	63.4	48.3	**2.79**
W/o c2f strategy	1.3/7.23	91.3	88.9	16.23
W/o alg. loop	2.3/14.56	87.6	83.52	13.4

↓: indicates that lower values correspond to better accuracy; ↑: indicates that higher values correspond to better accuracy.

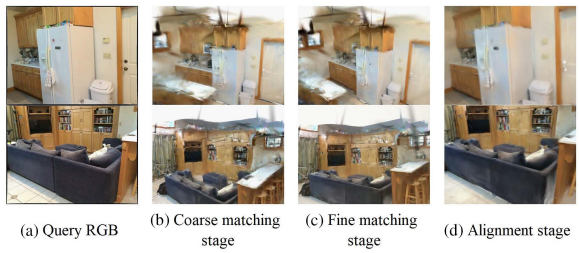


Fig. 12. Query image pose evaluation through algorithm stages.

TABLE VIII
MEDIAN ROTATION AND TRANSLATION ERRORS (°/CM) ACROSS PROCESSING STAGES ON THE SCANNET DATASET ↓

Stage	Scene0000	Scene0140	Scene0645	Avg. Err.
Coarse matching	8.24/45.7	11.6/58.3	9.42/47.9	9.75/50.6
Fine matching	6.87/31.6	9.12/41.3	8.9/40.2	8.3/37.7
Alignment	0.49/2.53	0.77/1.78	0.25/0.91	0.5/1.74

↓: indicates that lower values correspond to better accuracy.

Method without semantic information: To assess the impact of semantic information, semantic elements were removed from the Gaussian map while keeping the alignment module unchanged. This leads to reduced localization accuracy and success rate, with a slight runtime improvement, highlighting the importance of semantic context, as shown in Table VII.

Method without coarse-to-fine (c2f) matching strategy: To evaluate the role of view filtering and enhancement, both steps were removed and all viewpoints were directly used for alignment. This results in higher localization errors and increased runtime, demonstrating that the C2F strategy is crucial for accuracy and efficiency, as reported in Table VII.

Method without alignment (alg.) loop: To evaluate the role of the render-PnP loop, it was replaced with a single-shot PnP step using only the preliminary pose. This change led to higher errors and lower success rate, but improved runtime. The results in Table VII confirm that the render-PnP loop plays a key role in refining pose accuracy, albeit with added running time.

Evaluation of localization accuracy on processing stages: Three ScanNet sequences are used to evaluate pose estimation across the coarse and fine matching stages of the PR module and final alignment. Fig. 12 and Table VIII show the results. Coarse matching recovers an approximate view direction, while fine matching refines it, providing a strong basis for the final

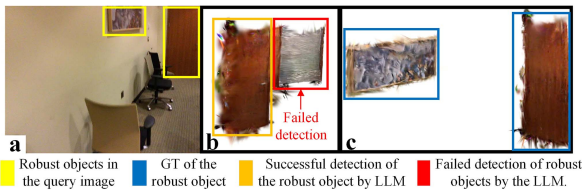


Fig. 13. Failure case of robust object detection by the LLM model. (a) Query image, (b) Failed detection, (c) GT of the robust object.

TABLE IX
QUANTITATIVE ANALYSIS OF FINAL POSE ACCURACY UNDER ROBUST OBJECT DETECTION FAILURES IN SCENE0140 OF THE SCANNET DATASET ↓

Scene	Translation Err. (cm)	Rotation Err. (°)	Failure rate (%)
0140	134	37.8	1

TABLE X
SUCCESS RATE (%) OF LLM MODULES ACROSS SCANNET DATASET↑

Items	LLava	Phi-3	GPT-4V
Module type	Local	Local	Cloud-based
Success rate	99%	**100%**	**100%**

alignment. Consequently, the combined stages enable high-precision pose estimation.

Unsuccessful detection of robust objects by the LLM model: Although the LLM achieves high accuracy in identifying robust objects for constructing the robust map, it may still fail when object shapes are ambiguous or the query image has poor visual quality. Fig. 13 illustrates such a failure: in Fig. 13(a), the robust objects are a wall painting and a door, while in Fig. 13(b), the LLM incorrectly detects a window and a door, with the window being a false positive. Ideally, it should match the GT in Fig. 13(c). This misidentification results in a less accurate final pose, as shown in Table IX. Nonetheless, the LLM maintains a low failure rate of about 1%, preserving the overall robustness of the proposed method.

Evaluate the impact of alternative LLMs on semantic extraction: This subsection evaluates the semantic extraction capabilities of three LLMs: LLaVA [12], Phi-3 [24], and GPT [25]. LLaVA is an open-source lightweight model that integrates easily into the framework, providing reasonable accuracy with good efficiency. Phi-3 offers stronger scene understanding and balanced performance but is less suitable for direct embedding due to its larger size. GPT delivers the most accurate comprehension thanks to large-scale training but relies on cloud access, limiting adaptability. As shown in Table X, LLaVA achieves about 99% accuracy, while Phi-3 and GPT both reach 100%, reflecting the simplicity of the task involving common, well-represented objects.

VII. CONCLUSION

In this study, we propose a robust visual localization framework that demonstrates significant improvements over existing SOTA methods across three distinct scenarios. Based on the experimental results, the key contributions of our approach are summarized as follows: 1) enhanced localization accuracy by integrating 3D GS with semantic information, 2) the ability to identify and select optimal viewpoints for rendering images that closely match the query image, and 3) robust state estimation in complex indoor environments.

REFERENCES

[1] Q. H. Hoang and G. W. Kim, "IMU augment tightly coupled LiDAR-visual-inertial odometry for agricultural environments," *IEEE Robot. Automat. Lett.*, vol. 9, no. 10, pp. 8483–8490, Oct. 2024.
[2] S. Chen et al., "Map-relative pose regression for visual relocalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 20665–20674.
[3] E. Brachmann, T. Cavallari, and V. A. Prisacariu, "Accelerated coordinate encoding: Learning to relocalize in minutes using RGB and poses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5044–5053.
[4] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, 2021.
[5] B. Zhao et al., "PNeRFLoc: Visual localization with point-based neural radiance fields," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 7450–7459.
[6] A. Moreau et al., "Crossfire: Camera relocalization on selfsupervised features from an implicit representation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 252–262.
[7] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139:1–139:14, 2023.
[8] C. Liu et al., "GS-CPR: Efficient camera pose refinement via 3D Gaussian splatting," in *Proc. 13th Int. Conf. Learn. Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=mP7uV59iJM
[9] Z. Huang et al. "From sparse to dense: Camera relocalization with scene-specific detector from feature Gaussian splatting," *Proc. Comput. Vis. Pattern Recognit. Conf.*, 2025, pp. 27059–27069.
[10] Y. Sun et al., "iComMa: Inverting 3D Gaussian splatting for camera pose estimation via comparing and matching," 2023, *arXiv:2312.09031*.
[11] H. Zhai et al. "SplaTloc: 3D Gaussian splatting-based visual localization for augmented reality," *IEEE Trans. Visual. Comput. Graph.*, vol. 31, no. 5, pp. 3591–3601, May 2025.
[12] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Adv. neural Inf. Process. Syst.*, vol. 36, pp. 34892–34916, 2023.
[13] G. Xiao-Shan, H. Xiao-Rong, T. Jianliang, and C. Hang-Fei, "Complete solution classification for the perspective-three-point problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 8, pp. 930–943, Aug. 2003.
[14] Y. Wu et al., "Opengaussian: Towards point-level 3D Gaussian based open vocabulary understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 19114–19138.
[15] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[16] J. S. Brauchart and P. J. Grabner, "Distributing many points on spheres: Minimal energy and designs," *J. Complexity*, vol. 31, no. 3, pp. 293–326, 2015.
[17] H. Edelsbrunner and E. P. Mücke, "Three-dimensional alpha shapes," *ACM Trans. Graph. (TOG)*, vol. 13, no. 1, pp. 43–72, 1994.
[18] U. Ramer, "An iterative procedure for the polygonal approximation of plane curves," *Comput. Graph. Image Process.*, vol. 1, no. 3, pp. 244–256, 1972.
[19] X. Zhao, X. Wu, W. Chen, P. C. Chen, Q. Xu, and Z. Li, "Aliked: A lighter keypoint and descriptor extraction network via deformable transformation," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–16, 2023, Art. no. 5014016.
[20] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "LightGlue: Local feature matching at light speed," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 17627–17638.
[21] X. Zhao et al., "Fast segment anything," 2023, *arXiv:2306.12156*.
[22] B. Glocker, S. Izadi, J. Shotton, and A. Criminisi, "Real-time RGB-D camera relocalization," in *Proc. IEEE Int. Symp. Mixed Augmented Reality*, Adelaide, SA, Australia, 2013, pp. 173–179.
[23] A. Dai et al., "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5828–5839.
[24] M. I. Abdin, S. A. Jacobs, A. A. Awan, J. Weerasinghe, A. Hassan, and Dutta, "Phi-3 technical report: A highly capable language model locally on your phone," Microsoft, Tech. Rep. MSR-TR-2024-12, Aug. 2024. [Online]. Available: https://www.microsoft.com/en-us/research/publication/phi-3-technical-report-a-highlycapable-language-model-locally-on-your-phone/
[25] Z. Yang et al., "The dawn of LMMs: Preliminary explorations with GPT-4V (ision)," 2023, *arXiv:2309.17421*.