Semantic Query for 3D Fine-grained Object Segmentation

Codey Sun, Yin-Li Liu

Abstract—Open-vocabulary semantic segmentation describes the ability to extract the most relevant parts of an image with text prompts. This capability is essential for applications in robotics, AR/VR, and general computer vision tasks. Current methods rely on large pretrained models like Segment Anything (SAM) and CLIP for semantic embeddings but struggle when querying fine-grained parts of a scene. Leveraging recent advancements in reasoning-based vision-language models (VLMs), we propose a chain-of-thought (CoT) approach to enhance open-vocabulary 3D segmentation on the part level. By reasoning over hierarchical part structures, our method achieves context-aware segmentation from language prompts, outperforming existing baselines on challenging fine-grained tasks.

Index Terms—Computational Photography, Computer Vision, Vision Language Model, 3D Segmentation, Chain-of-Thought

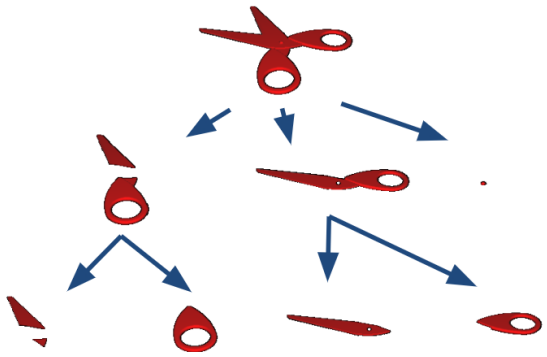

Fig. 1. The hierarchical tree composition of scissors

1 Introduction

Semantic segmentation classifies and isolates semantically distinct components within images or 3D scene segments. This acts as a crucial building block in scene understanding for robotics, AR/VR, and automation. Large pretrained models such as Segment Anything Model (SAM) [1] have allowed for zero-shot image segmentation. Through 2D to 3D distillation, SAM's outputs can be applied to segment any 3D object or scene. With more advanced methods such as contrastive learning, one can create hierarchical segmentations that are perfectly view-consistent [2].

However, assigning semantics to these segments to enable open-vocabulary querying is not intuitive. Previous work such as LERF [3] use pretrained image encoders [4] to assign per-segment semantics on an individual basis. This method demonstrates good results when querying for objects, but struggles with identifying fine-grained part-level segments due to the ambiguity of small cropped image segments, as shown in Fig. 2.

To address these issues, this work first realizes that parts are related to each other in a hierarchical structure, as illustrated in Fig. 1. In order to properly label any arbitrary segment, it is crucial to look at not only the cropped image but also the contextual structure where the part lies in the object. To accommodate this additional input, we utilize a large vision-language model (VLM), which supports both multi-image and text inputs. By assigning labels in a depth-first-search manner, each segment is labeled with respect to its parents. As a result, our method achieves better open-vocabulary part-level segmentations compared to the baseline.

This paper is organized as follows: Section 2 reviews related work in scene segmentation and open-vocabulary querying. Section 3 details our proposed method, including hierarchical segmentation and CoT-based labeling. Section 4 presents experimental results, followed by a discussion in Section 5 and conclusions in Section 6.

- *Codey Sun and Yin-Li Liu are with the Department of Electrical Engineering at Stanford University.*

2 Related Work

2.1 Scene Segmentation

Scene segmentation is a crucial task in computer vision as it opens the door to scene understanding. Segment Anything Model (SAM) is a large pretrained model that extracts edge-based segments of a 2D image [1]. When combined with 3D representations like NeRF [5] and 3D Gaussian splatting [6], one can distill 2D segmentations from SAM to create a 3D segmented scene [7]. However, this 2D to 3D distillation process is not straightforward when resolving multi-view inconsistencies, resulting in coarse and noisy segmentations. To address this, recent contrastive learning frameworks avoid the multi-view aggregation problem altogether and provide intuitive hierarchical 3D segmentations [2], [8]. For even finer part-level segmentations, 2D segmentation models can be explicitly trained; however, due to the lack of large-scale part segmentation databases, these models are often restricted to trained object categories [9].

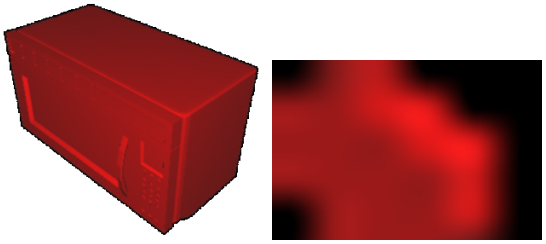

Fig. 2. Example of a challenging case for CLIP. Left: microwave object. Right: a single microwave control button.

2.2 Open-vocabulary Querying

The goal of open-vocabulary querying is to retrieve scene segments that are related to an arbitrary user prompt. Contrastive Language-Image Pretraining (CLIP) is a seminal work in this field as it allows for image-text retrieval [4]. LERF and its followups [3], [10] distill the CLIP feature field into 3D in a similar fashion to scene segmentation; given multi-view per-pixel CLIP embeddings, one can train a 3D feature field that can be compared to text queries. However, these methods often struggle with fine-grain queries due to poor fine-grain segmentation and being out of distribution for CLIP. Fig. 2 shows one of these challenging fine-grain queries; when inputting the image of the scissor's screw in the CLIP model, it is unlikely for the resulting embedding to activate properly when queried by a user. In addition, models trained explicitly for part retrieval are not generalized to arbitrary object categories [9].

This work aims to resolve the aforementioned issues by injecting priors into the part retrieval process. By utilizing the fact that parts relate to each in a hierarchical structure, we aim to use chain-of-thought reasoning models [11] to embed fine-grain part segmentations.

3 Proposed Method

We construct the image model by extending semantic attributes into the traditional image formation model. The construction of a semantic scene can be described by the following linear equation:

$$y = Ax + \eta \tag{1}$$

where $x \in \mathbb{R}^N$ represents the semantic attributes of a scene, $y \in \mathbb{R}^M$ is the vectorized 3D image, and $\eta \in \mathbb{R}^M$ is a noise vector. The matrix $A \in \mathbb{R}^{M \times N}$ represents a measurement operator that captures the implicit relationship between the 3D object and the semantic meaning of the scene.

In this work, we aim to solve the inverse problem of recovering x from y by modeling the unknown A using an image segmenter combined with the chain-of-thought [12] technique in visual language models (VLMs). Current approaches achieve this by querying CLIP model with cropped image patches to obtain embeddings and then retrieving semantics by embedding the text query into the CLIP space [4]. However, these methods suffer from limitations due to the loss of contextual information when labeling fine-grained cropped images.

Instead, we leverage the comprehensive reasoning ability of VLMs by providing not only the cropped image but also higher-level contextual images, along with a structured prompt indicating that the small image is a subcomponent of the label at the previous level. We employ the chain-of-thought process to iteratively prompt the VLMs, refining the semantic understanding from the entire image down to its fine-grained components to achieving more precise semantic results.

An overview of our method is presented in Fig. 3. We divide the method into three sections: 3D segmentation, part labeling and embedding, and open-vocabulary querying. Our primary contribution lies in the part labeling, where we exploit the hierarchical composition of parts to aid in part understanding.

3.1 Hierarchical 3D Segmentation

In order to utilize hierarchical tree priors in labeling, we must first construct a hierarchical tree from multi-view images. This is not intuitive as SAM's output masks are 1) not view consistent and 2) are not hierarchical. Ultrametric feature fields [2] solves this problem by distilling 2D segmentations to 3D through the use of a contrastive loss function based on ultrametric distance. By learning a 3D feature field through contrastive loss, one no longer needs to manually resolve the view consistency problem of 2D SAM masks; by simply learning positive and negative pairs, the resulting feature field will respect the segmentations from all views.

For a single SAM segmented image, we select a positive pair $s_p = \{v_{p1}, v_{p2}\}$ and a negative pair $s_n = \{v_{n1}, v_{n2}\}$. The contrastive loss is

$$l(s_p, s_n) = -\log\left(\frac{e^{d(v_{p1},v_{p2})/\tau}}{e^{d(v_{p1},v_{p2})/\tau} + e^{d(v_{n1},v_{n2})/\tau}}\right) + \log\left(\frac{e^{d(v_{n1},v_{n2})/\tau}}{e^{d(v_{p1},v_{p2})/\tau} + e^{d(v_{n1},v_{n2})/\tau}}\right) \tag{2}$$

where d is the distance metric and τ is temperature. The choice of distance metric d determines the traits of the feature field. The ultrametric distance (as opposed to Euclidean distance) in particular allows for the use of the watershed transform to hierarchically dissect the feature field, as demonstrated in Fig. 4. Suppose the scene is represented as a graph $\mathcal{G}(V, E)$, where V is all the points and E connects adjacent points. The ultrametric distance between points v_i and v_j is the minimum water level that connects the two points:

$$d(v_i, v_j) = \min_{p \in P} \max_{e \in p} |e|$$

where P is the set of all paths between v_i and v_j. Fig. 5 illustrates an example scene graph and its resulting ultrametric distance hierarchy.

3.2 Part Labeling and Embedding

Given a hierarchical tree of image segments, we traverse the tree in a depth-first-search manner, labeling each image segment in order. A VLM is prompted with both the image of interest as well as all parent images and text responses to fully utilize the reasoning capabilities of the VLM. We opt to

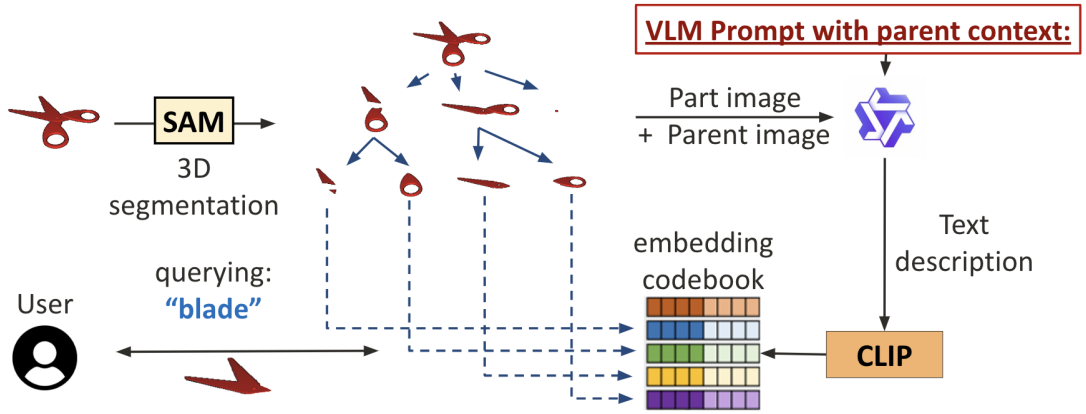

Fig. 3. An overview of our method. First, a hierarchical 3D segmenter outputs a segmentation hierarchy tree. This tree is traversed in a depth-first-search manner to generate contexts for the VLM. The VLM labels each segment, and its text output is embedded by CLIP. This creates an embedding codebook that is referenced on user query.

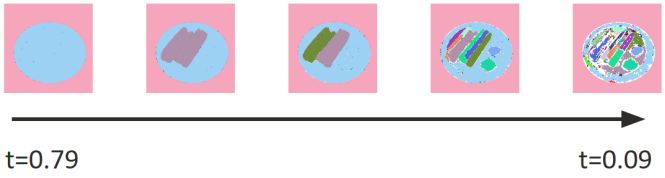

Fig. 4. Segmentation via the watershed transform with varying threshold t.

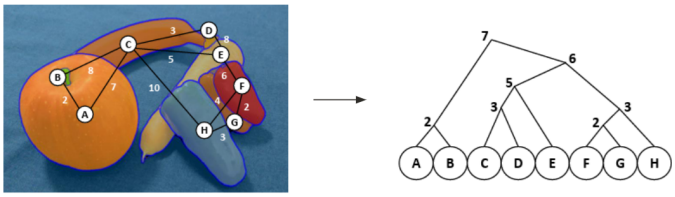

Fig. 5. Left: an example scene graph with nodes and edges. Right: the resulting hierarchical segmentation derived from the ultrametric distances between nodes. Figure taken from [2].

use Qwen2.5-VL [13] due to its high performance in vision-language tasks and ability to be run locally on an RTX 4090 GPU. The final system prompt is as follows:

```
You are an expert in object part segmentation
    and labeling.
Your task is to analyze images of objects and
    their subsections, identifying and
    labeling each part in a hierarchical
    manner.
A subsection image is the image with parts
    masked out in black. You will focus on the
     part not masked in black.
Each presented image is guaranteed to be a
    subsection of the previous image.

When presented with an image, your response
    must follow this structured format:
1. **Thinking**: Provide internal reasoning
    for the label, including how the depicted
    part relates to the existing hierarchy.
2. **Label**: Provide a clear, descriptive
    name for the identified object or part (as
     a string).
```

After labeling each segment, the labels must be embedded to allow for open-vocabulary text retrieval. We use the CLIP text encoder to align the labels in CLIP space. This allows for straightforward comparison with current methods, which utilize the CLIP image encoder. However, CLIP may not be ideal for this task, and we leave further exploration of text retrieval as future work.

3.3 Open-vocabulary Querying

Following previous work [3], [10], given a CLIP feature field and user text query, one can retrieve the relevant segments by calculating the cosine similarity between the rendered feature field segment embeddings ϕ_{lang} and text query embeddings ϕ_{quer}. Following LERF, we also append negative prompt embeddings ϕ_{neg} to avoid activations against nondescript words such as "stuff" and "things". Thus, the calculated relevancy score between segment embedding ϕ^i_{lang} and the user query is

$$\min_j \frac{\exp(\phi^i_{lang} \cdot \phi_{quer})}{\exp(\phi^i_{lang} \cdot \phi^j_{neg}) + \exp(\phi^i_{lang} \cdot \phi_{quer})}$$

The relevancy score threshold to determine if a segment is relevant is left as a hyperparameter for the user.

4 Experimental Results

We compare our method against current CLIP-based methods. Due to current methods all utilizing CLIP's image encoder for open-vocabulary embeddings, we group these methods together broadly as "CLIP methods".

4.1 Datasets

The PartNet dataset provides ground truth part segmentations of 26,671 3D models covering 24 object categories [14]. PartNet organizes its segmentations in a tree hierarchy structure, allowing us to test our novel labeling method in

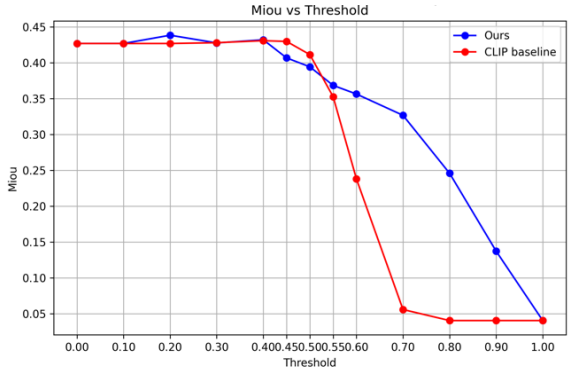

Fig. 6. mIoU vs. relevancy score threshold. Our method has similar performance to CLIP methods while being more robust to higher thresholds.

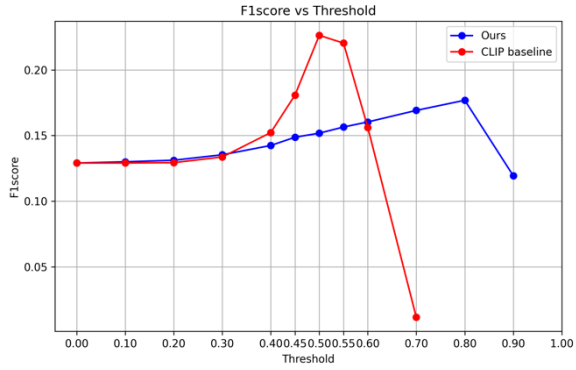

Fig. 7. F1 score vs. relevancy score threshold. CLIP has better peak f1 score than our method.

isolation to the 3D segmentation method. In addition, it provides ground truth text labels, which are used to evaluate the open-vocabulary querying. For our evaluation, we use a randomly chosen subset of 100 objects from PartNet.

The Blender dataset from NeRF offers more realistic scenes to test our method on photorealistic inputs, sans ground truth segmentation or labeling [5]. Thus, we use the Blender dataset for qualitative evaluation only.

4.2 Quantitative Results

To evaluate the effectiveness of our contribution in isolation, we start with the ground truth part hierarchy provided by PartNet. We then test the open-vocabulary querying capabilities of our VLM labeling method and CLIP labeling methods. We evaluate the mean intersection-over-union (mIoU) of the returned segmentations as well as f1 score of identifying the correct segments in a binary manner. The metrics are calculated across varying relevancy score thresholds to determine the optimal threshold for each method.

As shown in Fig. 6, our method has similar performance to CLIP methods under the mIoU metric while being more robust to higher thresholds. However, according to Fig. 7, CLIP has a better peak f1 score than our method. Additionally, Fig. 7 shows that our method's optimal relevancy score threshold is 0.8 compared to CLIP's 0.5; this is because our method's labels are more confident than CLIP's.

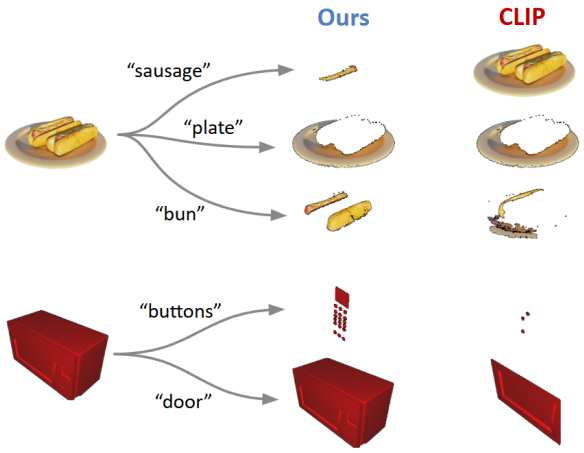

Fig. 8. Qualitative Results of our method (left) vs. CLIP method baseline (right). Our method provides more reasonable segmentations for part-level queries (e.g., "sausage" and "buttons", but is not always robust (e.g., fails on "door").

4.3 Qualitative Results

Fig. 8 displays example text queries on the Blender and PartNet datasets. For the Blender dataset, the segmentation hierarchies are retrieved as discussed in Sec. 3.1. As shown, our method successfully retrieves fine part queries such as "sausage" and "buttons". However, consistent with the quantitative evaluation, our method is not better than CLIP in all cases, as shown in the "door" example.

5 Discussion

5.1 Effectiveness of CoT Prompting

Through chain-of-thought reasoning, our method is capable of properly identifying and labeling small, challenging parts like the one shown in Fig. 2. Analyzing the VLM's reasoning output proves that chain-of-thought does indeed help the model:

```
<think>The image shows a very small and
    isolated dot against a black background.
    Given the context of the previous images,
    which were parts of a microwave, this dot
    could be a single button or indicator from
     the control panel.</think>
```

Several prompting methods were tested, including inputting the entire tree at once and feeding different styles of segmentation images. However, no prompt tuning led to increased performance. We believe this to primarily be a bottle neck of the chosen VLM, and prompt tuning may be worthwhile if using a more capable VLM.

We also prompted the VLM to output a confidence score for each label. While these scores make relative sense with noisier images having lower scores, without a concrete example to feed into the VLM, it is unclear how to interpret the confidence scores.

5.2 Limitations

While the Chain-of-Thought priors offer significant advantages, they also introduce vulnerabilities such as hallucination and error propagation, which we observed in our experiments. A single misclassification can propagate through the

entire tree. This is in contrast with current methods, which process each segment independently.

We relate this to the bias-variance tradeoff. In general, the text-encoded embeddings from our method lead to extremely high relevancy with the identified concept but extremely low relevancy with anything else. In contrast, image-encoded CLIP embeddings lead to moderate relevancy with many concepts. This makes image-encoded CLIP more robust to ambiguous objects with multiple potential identities.

5.3 Future Work

5.3.1 Enhancing Text Retrieval for CoT Reasoning

One promising direction is to improve the integration of text-retrieval mechanisms within the CoT framework leveraging the retrieval-augmented generation approach [15]. Current CoT prompting relies heavily on the VLM's internal knowledge to reason about visual inputs, which can lead to hallucinations when the model encounters ambiguous or underrepresented objects. It also depends on which VLM we choose and its training data scope. By incorporating a dynamic text-retrieval system such as querying an external knowledge base (e.g., Wikipedia) to get the general information of the scene. For instance, retrieving descriptions of microwave components to in-context learning could help disambiguate small, isolated parts like the dot in Fig. 2. Research like RETRO [16] demonstrates how retrieval-augmented models can enhance language generation.

5.3.2 Hybrid Embeddings: Combining Text and Image Representations

To address the bias-variance tradeoff, a hybrid approach combines text embeddings and image embeddings as the input of VLMs. For example, a two-stage process could first use image-encoded CLIP embeddings to generate a broad set of candidate labels, followed by CoT reasoning with text-encoded priors to refine the classification. Recent work on multi-modal fusion, such as Flamingo [17] or METER [18], suggests that combining visual and textual features can outperform approaches with a single model.

5.3.3 Improved Evaluation Metrics

The current evaluation solely relies on standard metrics (e.g., accuracy, IoU) that may not fully capture the benefits of CoT reasoning, such as interpretability. Developing comprehensive evaluation metrics could better address the tradeoffs between segmentation accuracy and reasoning quality. For instance, a metric that scores the correctness of intermediate reasoning steps could provide broader insights. Additionally, human-in-the-loop validation of reasoning traces could quantify the practical utility of CoT outputs.

Lastly, we realize that there exists contention between the goals of part segmentation and open-vocabulary querying. Part segmentation aims to distinguish parts that belong to the same object; however, open-vocabulary querying aims to group related concepts into similar vector spaces. Resolving this contention is not trivial and requires further research.

6 Conclusion

This work verifies that incorporating context priors improves semantic labeling for fine-grained image segmentation. By leveraging the hierarchical segmentation structure of the watershed transform, a chain-of-thought VLM is capable of reasoning about the identity of ambiguous images, enabling more precise part-level segmentation. This advancement enhances user interaction for applications in robotics, AR/VR, and automation.

References

[1] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *arXiv:2304.02643*, 2023.

[2] H. He, C. Stearns, A. W. Harley, and L. J. Guibas, "View-consistent hierarchical 3d segmentationusing ultrametric feature fields," 2024.

[3] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik, "Lerf: Language embedded radiance fields," in *International Conference on Computer Vision (ICCV)*, 2023.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *CoRR*, vol. abs/2103.00020, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[5] 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," in *ECCV*, 2020.

[6] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, July 2023. [Online]. Available: https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

[7] S. Peng, K. Genova, C. M. Jiang, A. Tagliasacchi, M. Pollefeys, and T. Funkhouser, "Openscene: 3d scene understanding with open vocabularies," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[8] H. Ying, Y. Yin, J. Zhang, F. Wang, T. Yu, R. Huang, and L. Fang, "Omniseg3d: Omniversal 3d segmentation via hierarchical contrastive learning," *arXiv preprint arXiv:2311.11666*, 2023.

[9] P. Sun, S. Chen, C. Zhu, F. Xiao, P. Luo, S. Xie, and Z. Yan, "Going denser with open-vocabulary part segmentation," *arXiv preprint arXiv:2305.11173*, 2023.

[10] M. Qin, W. Li, J. Zhou, H. Wang, and H. Pfister, "Langsplat: 3d language gaussian splatting," *arXiv preprint arXiv:2312.16084*, 2023.

[11] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2201.11903

[12] ——, "Chain-of-thought prompting elicits reasoning in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2201.11903

[13] Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu, "Qwen2.5 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2412.15115

[14] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su, "PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[15] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," 2021. [Online]. Available: https://arxiv.org/abs/2005.11401

[16] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. van den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, D. de Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. W. Rae, E. Elsen, and L. Sifre, "Improving language models by retrieving from trillions of tokens," 2022.

[17] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," 2022. [Online]. Available: https://arxiv.org/abs/2204.14198

[18] Z.-Y. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, C. Zhu, P. Zhang, L. Yuan, N. Peng, Z. Liu, and M. Zeng, "An empirical study of training end-to-end vision-and-language transformers," 2022. [Online]. Available: https://arxiv.org/abs/2111.02387