A VLM-Based Approach for 2D Mask Grouping Across Views

Saro Harutyunyan, Edoardo Calderoni, Valentin Brevet

Technical University of Munich

Abstract

We explore the use of large Vision-Language Models (VLMs) for the task of cross-view 2D instance mask grouping, aiming to associate object masks observed in multiple camera views into consistent 3D instances. Using the Replica indoor dataset, we formulate mask grouping as a binary visual reasoning problem: given two cropped instance masks from different views, determine whether they correspond to the same underlying 3D object. We establish a baseline using the pretrained Qwen3-VL model and subsequently fine-tune it using LoRA adapters, without modifying the base model weights. Our experiments on held-out scenes demonstrate that LoRA fine-tuning improves mask ID consistency highlighting the potential of VLMs without explicit 3D reasoning at inference time. The codebase is available at `https://github.com/saro2808/ML43Dproject`.

1. Introduction

Instance-level scene understanding is a core challenge in 3D perception [5]. While modern pipelines often rely on explicit geometric reasoning — such as point cloud clustering or hierarchical feature aggregation methods (e.g., PointNet++ [6]) — these approaches typically require dense reconstruction and carefully designed heuristics [8]. In contrast, recent Vision-Language Models (VLMs) have shown remarkable zero-shot reasoning abilities across images [3, 7], suggesting they may implicitly capture object identity cues such as shape, texture, and semantic consistency [4, 9].

In this work, we investigate whether a VLM can be trained to group 2D instance masks across views into consistent 3D object identities. Instead of predicting explicit correspondences or embeddings, we frame the problem as a binary question-answering task: Do these two image crops correspond to the same object?

Our contributions are:

1. a VLM-compatible formulation of multi-view 2D mask grouping using the Replica dataset,
2. a baseline evaluation of pretrained Qwen3-VL [1] on this task,
3. a lightweight LoRA fine-tuning strategy that improves consistency without updating base weights.

2. Dataset and Preprocessing

2.1. Replica Dataset

We use the Replica dataset, a photorealistic indoor dataset with high-quality meshes, camera trajectories, and per-vertex instance annotations. Replica provides an ideal testbed for studying multi-view instance consistency under controlled conditions.

Eight scenes are available: 5 offices and 3 rooms. Given that object distributions do not vary drastically across scenes (see Figure 1) we split the dataset as follows:

- 6 training scenes (4 offices and 2 rooms): used for LoRA fine-tuning,
- 2 evaluation scenes (1 office and 1 room): office4 and room2.

Inspecting Figure 1 one may observe that the distributions of office4 and room2 are pretty much in the middle of the others, which justifies their usage as evaluation scenes.

2.2. Multi-View Rendering and Instance Mask Generation

For each scene, we render multi-view RGB images and instance masks using PyTorch3D. Each view is generated from the original camera poses and intrinsics provided by Replica.

Key steps:

- load the textured mesh and camera parameters,
- render RGB images using a flat shader,
- render instance masks by mapping rendered faces to ground-truth instance IDs via majority voting over vertex labels (see Figure 2).

Each rendered view is saved as:

- `rgb.png`: the RGB image,
- `instance_mask.npy`: per-pixel instance IDs,
- `unique_instances.npy`: list of visible instance IDs.

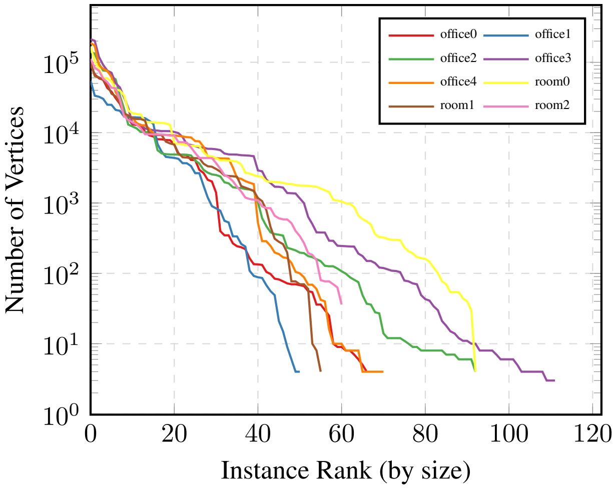

Figure 1. Distribution of instance sizes (vertex counts) across the Replica dataset. The logarithmic scale reveals a consistent long-tail distribution across both office and room environments.

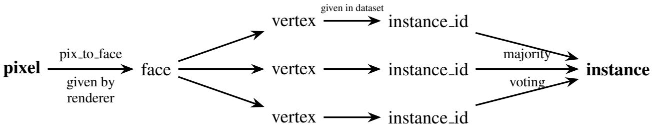

Figure 2. Hierarchical mapping from 2D pixels to 3D instance labels using renderer metadata and majority voting.

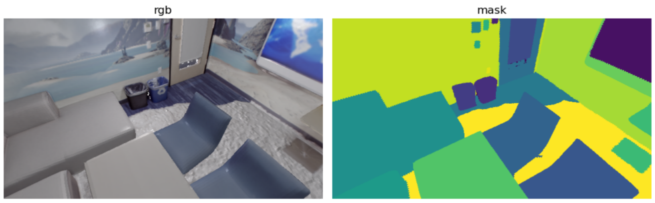

Figure 3. Visualization of the RGB image and instance mask.

2.3. Instance Pair Construction

To train and evaluate the VLM, we construct pairs of instance crops:

- two views are sampled;
- a shared instance ID defines a positive pair,
- a different instance ID defines a negative pair.

When constructing the dataset for a given scene we build a map giving for each instance ID the view pairs containing that instance. However, to combat instance imbalance we bound the number of view pairs for each instance by 10. Moreover, for the same reason dataset items are sampled uniformly over instance IDs.

Each instance crop is extracted by tightly bounding the corresponding mask region. We add a small padding of width 10 pixels to make the crops less abrupt and provide the VLM with more local context around the object boundaries. The resulting pair is packaged into a VLM-style prompt whether they refer to the same object in the 3D scene (see Figure 4). This formulation removes the need for explicit geometric reasoning during inference.

3. Model and Training

3.1. Base Model: Qwen3-VL

We use Qwen3-VL-8B, the Qwen variant with 8B total parameters, a large Vision-Language Model capable of multi-image reasoning. The base model is evaluated without any fine-tuning to establish a baseline for zero-shot performance.

3.2. LoRA Fine-Tuning

To adapt the model to our task efficiently, we fine-tune Qwen3-VL using Low-Rank Adaptation (LoRA) [2]:

- only LoRA adapters are trained; base model weights remain frozen;
- vision, language, attention, and MLP modules are adapted;
- 4 bit quantization is utilized.

This choice is motivated by:

- limited GPU memory (Colab environment, one Tesla T4 GPU with 15GB RAM),
- desire to preserve general VLM capabilities,
- faster training and reduced storage requirements.

Training is performed for 2 epochs with 1000 samples from each train scene, which we found sufficient to observe measurable improvements.

Table 1. LoRA Hyperparameters

Parameter	Value
LoRA rank (r)	16
LoRA alpha	16
LoRA dropout	0.05
Learning rate	2×10^{-4}
Batch size	2
Grad. accum.	4
Epochs	2
FP16	Yes

Table 2. Dataset Configuration

Parameter	Value
Max train samples	6×1000
Max eval samples	2×300
Neg. pair prob.	0.5

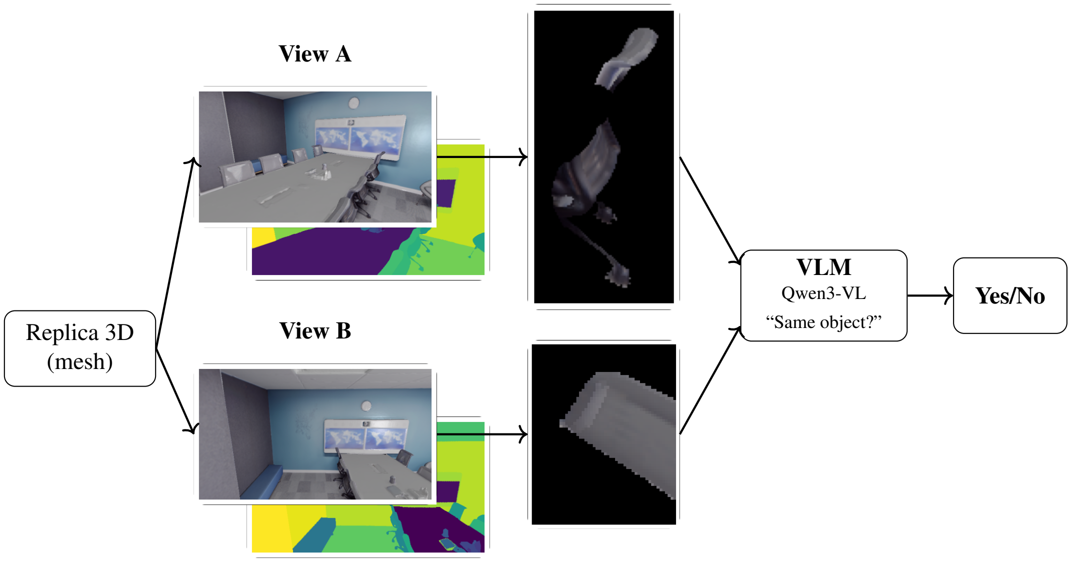

Figure 4. Overview of the proposed pipeline. Multi-view RGB images and instance masks are rendered from a 3D scene, cropped into instance-level regions, and evaluated by a Vision-Language Model via a binary query.

3.3. Training Stability and Checkpointing

Given the constraints of Colab (session crashes, limited disk space), we:

- periodically save checkpoints to Google Drive,
- overwrite the previous checkpoint to conserve storage,
- train for a fixed small number of epochs rather than early stopping.

While intermediate evaluation during training could be added, the limited training duration makes validation loops less critical in this setting.

4. Evaluation Protocol

4.1. Metrics

We evaluate performance by tracking accuracy, precision, recall and ambiguity rate (fraction of samples where the model fails to clearly answer "Yes" or "No"). Ambiguous outputs are handled explicitly:

- attempt to parse the model's response;
- retry with a forced instruction ("Answer with exactly one word: Yes or No.");
- if still ambiguous, count the sample as ambiguous and force an incorrect prediction.

This protocol avoids artificially inflated accuracy due to skipped samples.

Remarkably, phrasing the problem as binary classification makes evaluation much easier than if we were to give more than two crops to the VLM and ask to group them (a substantial difficulty could arise from parsing VLM's output).

4.2. Scenes and Setup

Evaluation is conducted on the scenes **office4** and **room2**. For each scene, 300 instance pairs are evaluated with resume support, ensuring robustness to interruptions.

Table 3. Performance Comparison. Acc: Accuracy, Prec: Precision, Rec: Recall, Amb.: Ambiguity Rate.

Scene	Model	Acc.	Prec.	Rec.	Amb.
office4	Base	0.74	0.67	0.88	0.00
office4	LoRA	0.82	0.76	0.92	0.00
room2	Base	0.80	0.76	0.91	0.00
room2	LoRA	0.92	0.91	0.94	0.00

5. Results and Discussion

5.1. Baseline Performance

The base Qwen3-VL model demonstrates acceptably good performance, suggesting that large VLMs encode rich visual priors about object identity. Surprisingly, the model is quite assertive in its answers — we observe zero ambiguous responses (possibly after retrying with more forceful prompt). This attests to the model's high confidence and lack of hedge-word bias, even when its predictions are technically incorrect. However, there is clearly room for improvement, especially with hard negatives.

5.2. Effect of LoRA Fine-Tuning

After LoRA fine-tuning:

- accuracy and precision improve considerably, especially for negative pairs,
- recall improves mildly (or even worsens, see Table 4).

This suggests that LoRA fine-tuning helps the model internalize task-specific cues, such as instance-level shape consistency across views. Particularly, as we could anticipate, after fine-tuning too (similar to the base model) when given the prompt Qwen3-VL is firm in its answers (again, possibly after the forceful second prompt); there is no any ambiguity whatsoever.

Table 4. Hyperparameter tuning results (Rank and LR).

R	LR	Scene	Acc.	Prec.	Rec.	Amb.
8	$2e^{-4}$	office4	0.82	0.75	0.92	0.00
		room2	0.92	0.91	0.94	0.00
8	$2e^{-5}$	office4	0.82	0.83	0.77	0.00
		room2	0.89	0.92	0.87	0.00
16	$2e^{-4}$	office4	0.82	0.76	0.92	0.00
		room2	0.92	0.91	0.94	0.00
32	$2e^{-4}$	office4	0.84	0.82	0.84	0.00
		room2	0.91	0.90	0.93	0.00
64	$2e^{-4}$	office4	0.85	0.83	0.85	0.00
		room2	0.92	0.93	0.91	0.00
64	$2e^{-5}$	office4	0.82	0.88	0.72	0.00
		room2	0.89	0.95	0.83	0.00

5.3. Failure Modes

Common failure cases include:

- symmetric or repetitive objects (e.g. chairs, walls),
- severe occlusions,
- very small instance crops with limited visual context.

Another thing worth mentioning is the model's bias toward answering "Yes" for difficult negative pairs (see Table 5):

Table 5. Mistake Rate on Negative Pairs (Rank-16)

Scene	Model	Samples	Neg. Errors	Share
office4	Base	78	61	0.78
office4	LoRA	53	42	0.79
room2	Base	61	47	0.77
room2	LoRA	25	15	0.60

We present some difficult pairs in Appendix A.

6. Conclusion

We present a VLM-based approach for grouping 2D instance masks across views, demonstrating that large Vision-Language Models can be adapted for geometric consistency tasks with minimal fine-tuning. Our results show that LoRA fine-tuning improves instance grouping accuracy and other related metrics, even with limited training data and compute. The key characteristic of our approach is that it avoids explicit 3D reasoning at inference time, relying instead on learned visual semantics.

Future work includes:

- extending beyond binary decisions to clustering,
- backprojecting the predicted 2D masks to reconstruct 3D instance masks.

A. Qualitative Visualizations

What follows are visualizations of pairs which were misclassified with the base Qwen3-VL but discriminated correctly with the fine-tuned one.

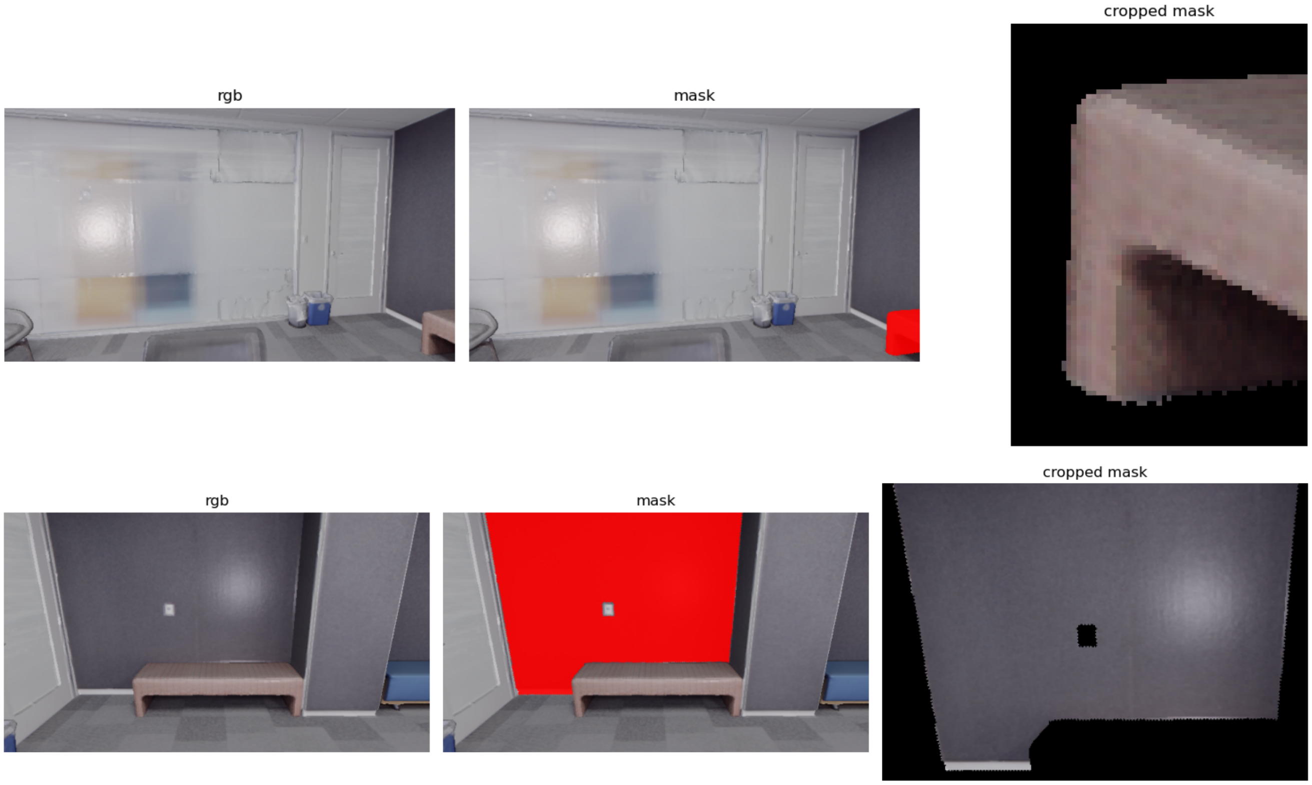

Figure 5. A negative pair. Assumed difficulty: shape similarity.

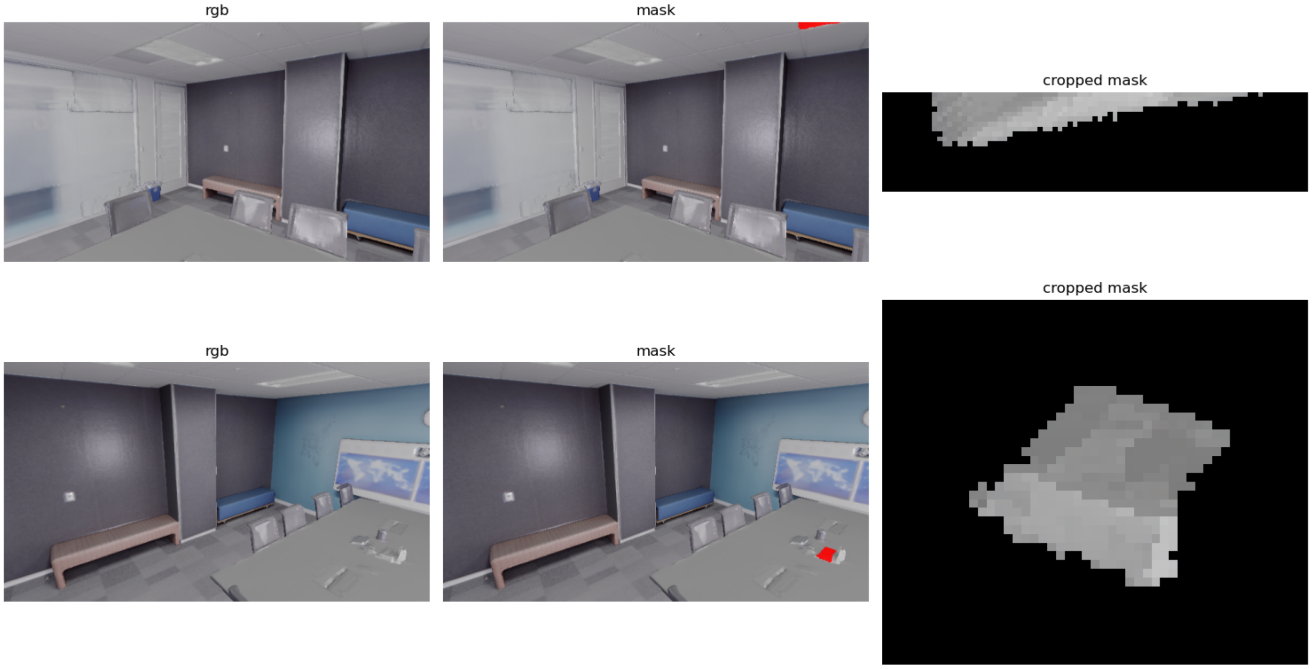

Figure 6. A negative pair. Assumed difficulty: smallness and occlusion.

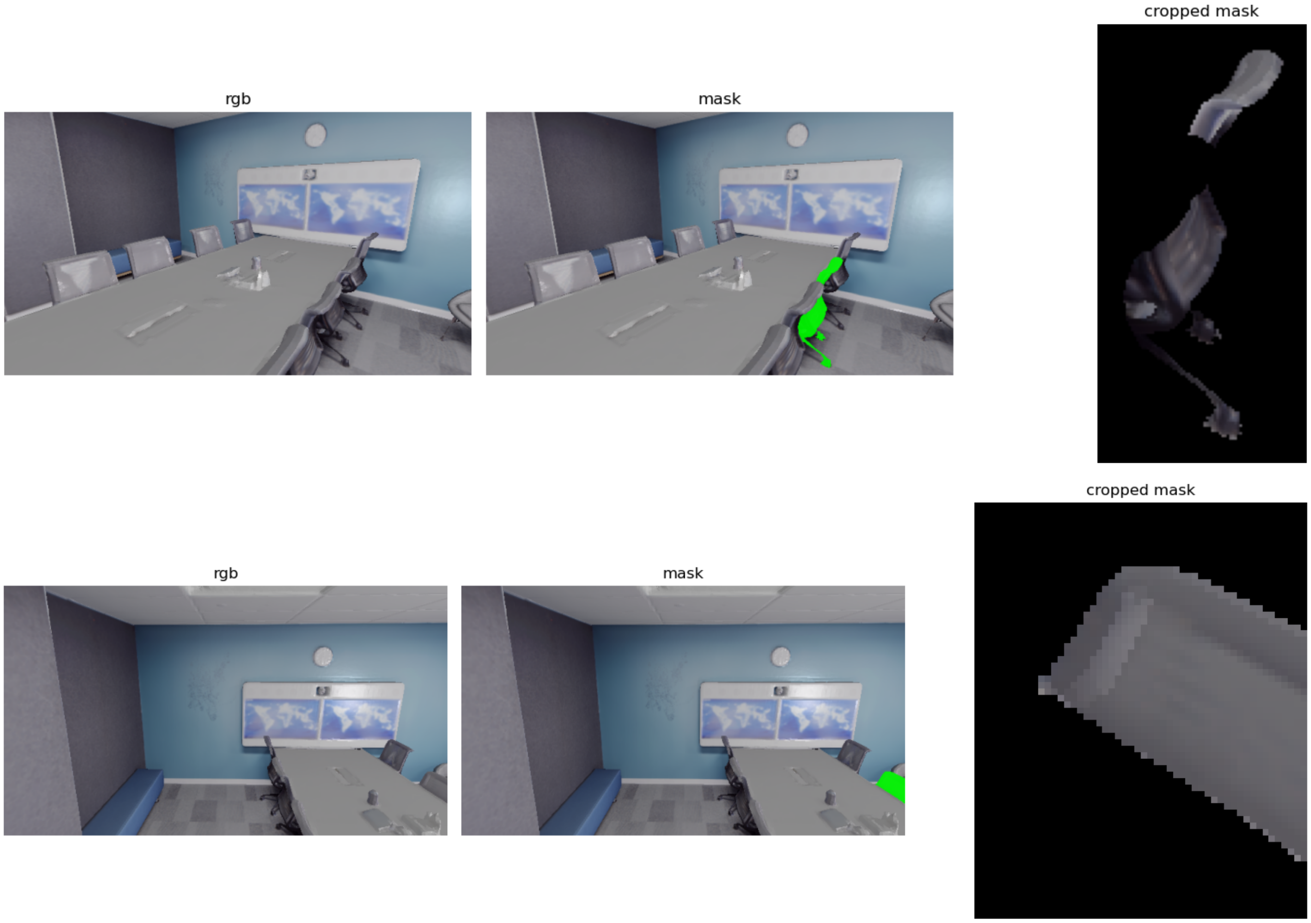

Figure 7. A positive pair. Assumed difficulty: symmetry and occlusion.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. 1

[2] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 2

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003, 2023. 1

[4] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification. 1

[5] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[6] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: deep hierarchical feature learning on point sets in a metric space. In *Proceedings of the 31st International*

Conference on Neural Information Processing Systems, page 5105–5114, Red Hook, NY, USA, 2017. Curran Associates Inc. 1

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[8] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223, 2023. 1

[9] Jianshu Zhang, Dongyu Yao, Renjie Pi, Paul Pu Liang, and Yi R. Fung. VLM2-bench: A closer look at how well VLMs implicitly link explicit matching visual cues. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7510–7545, Vienna, Austria, 2025. Association for Computational Linguistics. 1