CGEBench: Benchmarking Concept Generalization of Promptable Image Segmentation Models

Alexander von Recum
Ludwig Maximilian University of Munich
Recum.Al@campus.lmu.de

Christoph Schnabl
University of Cambridge
cs2280@cam.ac.uk

Abstract

Promptable image segmentation models have emerged as an evolution of image segmentation models with fixed classes. This approach allows for great flexibility during usage, enabling the user to prompt the model using a broad set of text prompts. However, it is unknown how robust these models are to generalizations of concepts, or "hypernyms". For example, if a model is prompted with "orange cat", "cat", and "animal", will the mask for the more specific concept be contained within masks for more general concepts? Has the model learned a consistent concept hierarchy, where each concept entails a broader concept? To evaluate this, we introduce ***CGEBench****, a modified version of SaCo-Gold for evaluating concept generalization of open-vocabulary image segmentation models. We evaluate SAM-3, a state-of-the-art image segmentation model, on* ***CGEBench*** *and show that it exhibits inconsistencies when generalizing to more abstract concepts, with concepts of increasing generality being labeled increasingly less consistently on average. After filtering to high-confidence prompt hierarchies, mean IoGT decreases from 0.98 at p_{base} to 0.85, 0.80, and 0.74 at p_1, p_2, and p_3, while cumulative agreement drops to 0.56 by p_3. Exact-zero IoGT becomes more common at higher abstraction levels, rising from 13.2% at p_1 to 22.7% at p_3, although partial-overlap cases remain the majority. These results position concept generalization of future promptable image segmentation models as an important area for benchmarking and improvement. The dataset is available at huggingface.co/datasets/avrecum/cgebench.*

1. Introduction

Open-vocabulary segmentation requires models to map free-form concepts to precise pixel-level masks. Early deep models operated over fixed class taxonomies. Mask R-CNN [10] and DeepLab [4] performed well on COCO [15] and ADE20K [31], but remained limited to categories seen during training. Panoptic segmentation [11] unified instance and semantic segmentation, but not the closed-vocabulary limitation.

Large-scale vision-language pretraining, including CLIP [21], enabled open-vocabulary segmentation. OpenSeg [8] aligned per-pixel features with CLIP embeddings using image-level supervision. Later methods extended this direction: OVSeg [14], FC-CLIP [29], and ODISE [27]. CLIPSeg [19], GroupViT [26], and LSeg [13] further connected language supervision and dense segmentation. These methods expanded label vocabularies, but usually predicted among candidate labels rather than handling fully free-form prompts.

SAM [12] introduced promptable segmentation from points, boxes, and masks. It showed strong zero-shot transfer on SA-1B but remained class-agnostic. SAM 2 [22] extended this design to video. Related systems such as SegGPT [25], SEEM [32], and Grounded SAM [17, 23] improved semantic grounding, often via multi-stage pipelines.

SAM-3 [3] introduces Promptable Concept Segmentation (PCS), combining text and exemplar prompts for detection, segmentation, and tracking. The accompanying SA-Co Gold set contains about 207K concepts, about 50× of LVIS [9]. In the reported PCS setting, SAM-3 achieves about 2× higher accuracy than prior systems.

Current evaluations ask whether a model segments a given concept, but not whether predictions remain consistent across abstraction levels. If a model segments "Labrador retriever," does it produce a containing mask for "dog," "canine," and "animal"? We call this property *concept generalization consistency*. We introduce **CGEBench**, derived from SA-Co Gold, to evaluate this property in open-vocabulary segmentation and analyze how SAM-3 behaves across semantic hierarchies.

1.1. Related Work

Semantic hierarchies in vision-language models. Prior work studies whether vision-language models capture taxonomy. SHiNe [16] reports variation across semantic granularity and proposes hierarchy-aware classifiers. MERU [5] uses hyperbolic embeddings for visual-semantic entailment.

Figure 1. Select failures of concept generalization in SAM-3. Ground truth is labeled in light green, SAM-3 generated mask in purple. IoGT represents the intersection of the ground truth and the generated mask divided by ground truth area.

CHiLS [20] and Hierarchy-CLIP [7] incorporate WordNet structure into CLIP-style classification. HIPIE [24] addresses multi-granularity segmentation. Baryshnikov and Ryabinin [2] evaluate hypernymy in text-to-image models. Xu et al. [28] study cross-modal taxonomic generalization. While these works focus on classification/detection behavior, they do not test spatial mask consistency across abstraction levels.

Segmentation evaluation metrics. Standard metrics such as mIoU and Panoptic Quality [11] do not model semantic proximity. SG-IoU [18] addresses this by reducing penalties for semantically close confusions (e.g., "table" vs "coffee table"). Our containment-based evaluation is complementary: it tests whether inclusion relations are preserved across a concept hierarchy.

2. CGEBench: Evaluating Concept Generalization in Promptable Segmentation Models

Our goal is to evaluate whether image segmentation models have learned a consistent visual hierarchy, where more general prompts result in masks that entirely contain more specific prompts. We call this ability "concept generalization".

We define the task of concept generalization as reliably labeling concept prompts that are superclasses of other, more specific prompts. That is, if an object is an instance of concept X, it must also be an instance of concept Y whenever Y is a superclass of X. For our dataset, we create a modified version of the SaCo-Gold benchmark [1, 6] which, in addition to Image-NP pairs in SaCo-Gold includes increasingly general text prompts which entail the specific prompt provided by SaCo-Gold. To obtain a sequence of prompts with increasing generality, we prompt an LLM to generate three increasingly general concepts, yielding a sequence $(p_{\text{base}}, p_1, p_2, p_3)$, where p_{base} is the base prompt from SaCo-Gold and p_1, p_2, p_3 are progressively more general prompts. We then apply a VLM filtering pass to remove hierarchies whose prompts are not concrete segmentation targets or whose generated concepts do not match the image. We provide an overview of the most frequently occurring concepts after this filtering step in Table 2 of Appendix 5. The SAM-3 authors do not publish a fixed concept hierarchy with SA-Co Gold, so we construct and evaluate the model only on our own prompt-generalization hierarchies. To validate data quality, after generating and filtering, all co-authors manually verified at least 200 randomly sampled prompt chains and labels for correctness. We run this filtering pass over 17,802 candidate hierarchies and retain 812 high-confidence prompt chains for the final SAM-3 evaluation. Several example generalization prompt chains can be found in Appendix 5. As of writing, SAM-3 is the most capable promptable

segmentation model available to us for this evaluation. We expect similar trends on the recent SAM-3.1 checkpoint, as SAM-3.1 primarily improves in the area of multi-object video tracking. As a result, we do not evaluate SAM-3.1.

2.1. Ground-Truth-Normalized Intersection Score (IoGT)

Given a prompt $p_k \in \{p_{\text{base}}, p_1, p_2, p_3\}$, let $M(p_k)$ be the predicted binary mask and let G be the ground-truth binary mask for the instance. We measure

$$s(p_k) = \frac{|M(p_k) \cap G|}{|G|}, \tag{1}$$

where $|\cdot|$ denotes mask area in pixels.

This score measures how much of the ground-truth object is covered by the mask produced from each prompt. We use it because if one prompt denotes a subclass and another denotes its superclass, then the superclass mask should include at least the subclass region. Therefore, for a superclass prompt p_j of p_i, we expect $s(p_j) \geq s(p_i)$.

In addition, we measure cumulative intersection with the ground truth up to hierarchy level i. Let $S_i = \bigcap_{k=0}^{i} M(p_k)$, where $p_0 = p_{\text{base}}$. We define

$$c(i) = \frac{|S_i \cap G|}{|G|} = \frac{|(\bigcap_{k=0}^{i} M(p_k)) \cap G|}{|G|}. \tag{2}$$

This metric captures agreement across levels of the learned concept hierarchy. While IoGT evaluates each prompt independently, $c(i)$ reveals whether predictions remain mutually consistent as concepts become more general. A sharp drop in $c(i)$ at some level indicates that the hierarchy collapses: masks at that level no longer preserve shared object support from earlier levels. Ideally, for a well-ordered hierarchy, all masks contain the true object region and both $s(p_k)$ and $c(i)$ stay close to 1 at every level.

3. Experiments

We evaluate SAM-3 on ***CGEBench***, our modified version of SaCo-Gold [3]. Figure 2 reports mean IoGT across prompt levels $(p_{\text{base}}, p_1, p_2, p_3)$. Mean IoGT decreases monotonically from 0.977 at p_{base} to 0.847, 0.799, and 0.745 at p_1, p_2, and p_3, respectively. The cumulative intersection score drops more sharply, from 0.977 at p_{base} to 0.836, 0.712, and 0.560.

Figure 3 shows that full-overlap predictions become rare at higher abstraction levels: 2.1%, 2.8%, and 3.1% of examples have IoGT $= 1$ at p_1, p_2, and p_3. Exact-zero IoGT rises from 13.2% at p_1 to 17.7% at p_2 and 22.7% at p_3, while partial-overlap cases remain the majority.

Together, these trends suggest that hierarchy consistency degrades as concept abstraction increases.

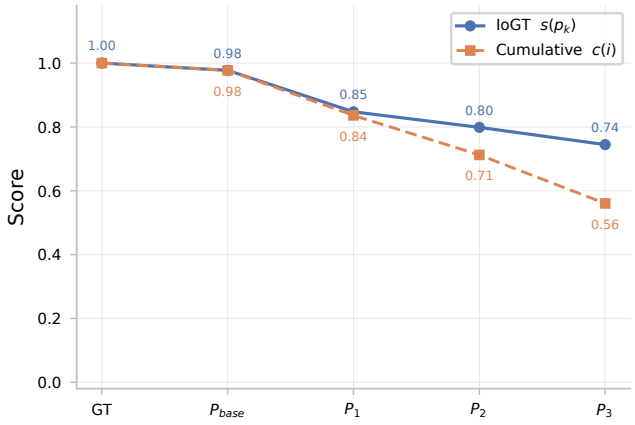

Figure 2. Mean IoGT and cumulative intersection score across generalization levels. Mean IoGT decreases from 0.98 at p_{base} to 0.74 at p_3, while cumulative agreement drops from 0.98 to 0.56.

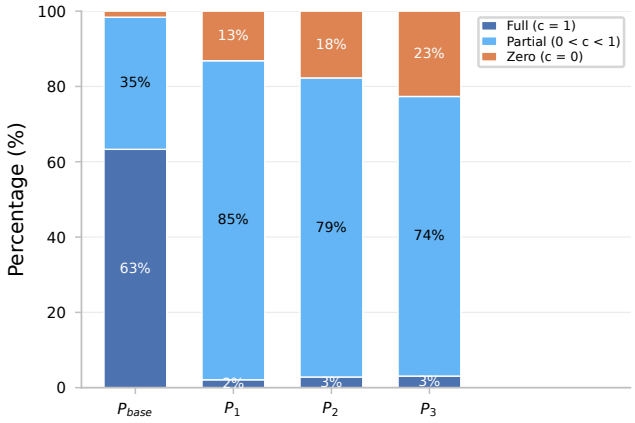

Figure 3. IoGT score distribution by level. Full-overlap predictions become rare after p_{base}, while exact-zero IoGT increases from 13.2% at p_1 to 22.7% at p_3. Partial-overlap predictions remain the majority at higher abstraction levels.

3.1. Concept Generalization across Semantic Categories

To analyze whether failures differ across semantic categories, we embed each source prompt p_{base} using `Qwen/Qwen3-Embedding-4B` [30]. We cluster the normalized embeddings with k-means, selecting $k = 28$ by cosine silhouette over candidate cluster counts. For visualization, we plot the largest 20 clusters after omitting the children/people cluster, and label clusters using semantic summaries of representative source prompts.

We define `failure` as $1 - \text{IoGT}$ and show failure rates across semantic categories in Figure 4.

Failure rates vary substantially by semantic category. At p_3, the highest-failure large clusters include floors/surfaces (74.0%), roads/paths (57.7%), flowers/florals

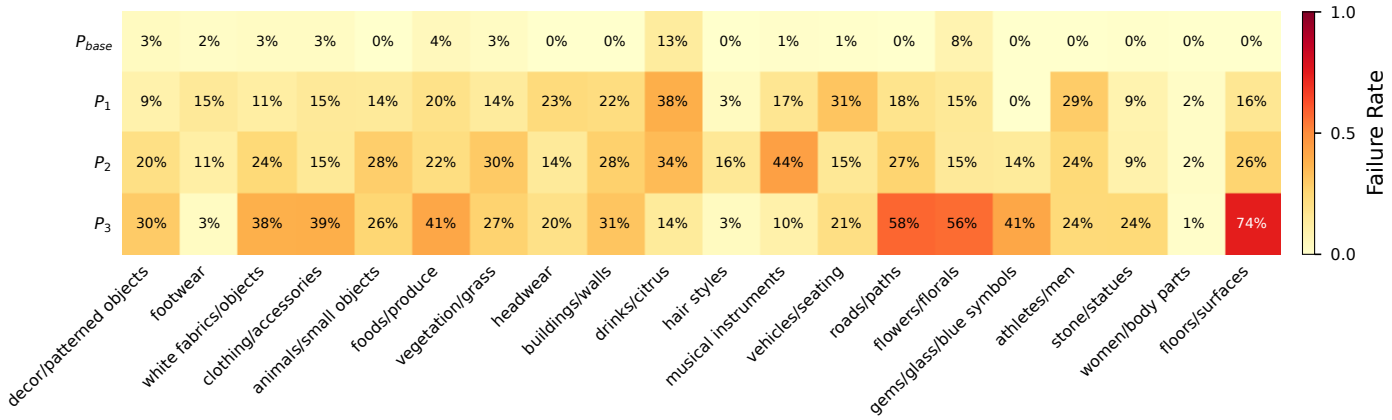

Figure 4. Failure heatmap across semantic categories and generalization levels. We plot $1 - \text{IoGT}$ in percent as the failure rate of concept generalization for a particular prompt level. Clusters sorted by cluster size in descending order from left to right.

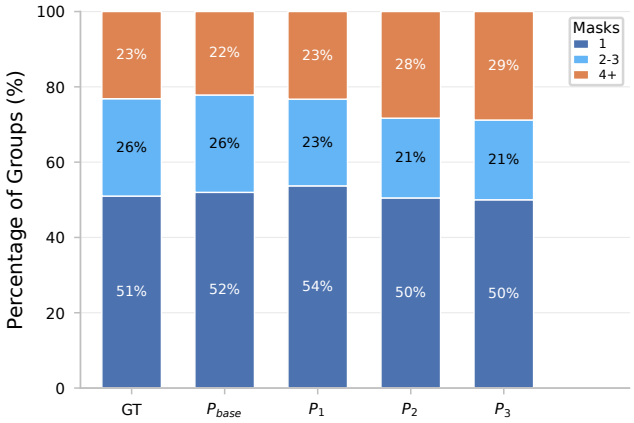

Figure 5. Distribution of the number of SAM-3 mask outputs per prompt level. The share of groups with four or more mask outputs increases from 22.2% at p_{base} to 28.8% at p_3, while single-mask outputs remain about half of examples.

(56.3%), foods/produce (41.4%), and gems/glass/blue symbols (41.0%). In contrast, several clusters remain comparatively stable, including women/body parts (0.9%), footwear (3.1%), hair styles (3.3%), and musical instruments (9.7%). These differences suggest that concept generalization failure is not uniform across semantic domains.

Overall, however, lack of concept generalization appears to be a consistent trend in current promptable image segmentation models.

4. Discussion and Future Work

Our experiments on CGEBench show a clear containment drop as concepts become increasingly abstract. Part of this degradation can come from inherently harder containment decisions at higher abstraction levels, but the dominant pattern is a generalization failure: full-overlap predictions become rare at higher abstraction levels, exact-zero failures increase, and most retained examples fall into partial-overlap behavior. To reduce annotation noise as a confounder, all co-authors manually checked at least 200 sampled prompt chains and labels. This supports our core argument that, for open-vocabulary segmentation, robust concept generalization should not depend on one specific hierarchy design. Beyond evaluation, CGEBench is also useful as a training signal to improve model robustness to semantic abstraction.

A natural next step is to specialize CGEBench with benchmark variants that stress different generalization axes, such as spatially conditioned prompts (e.g., “left”, “right”, “behind”), attribute-based prefixes, and other compositional prompt alterations. These extensions can help isolate where generalization breaks and guide model and training improvements.

Further improvements to open-vocabulary semantic segmentation models informed by the results of CGEBench could include hierarchically consistent datasets that reflect containment relationships. Additional experiments could include replacing the text encoder in SAM-3 with a more powerful one, which could potentially provide more general text representations helpful for more reliably recognizing concepts across semantic hierarchies.

References

[1] Sa-Co-Gold dataset (roboflow universe). https://universe.roboflow.com/sa-co-gold, 2026. Accessed: April 1, 2026. 2

[2] Anton Baryshnikov and Max Ryabinin. Hypernymy understanding evaluation of text-to-image models via WordNet hierarchy. *arXiv preprint arXiv:2310.09247*, 2023. 2

[3] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus

Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. 1, 3

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018. 1

[5] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Ramakrishna Vedantam. Hyperbolic image-text representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023. 1

[6] Facebook. SACo-Gold dataset (hugging face). https://huggingface.co/datasets/facebook/SACo-Gold, 2026. Accessed: April 1, 2026. 2

[7] Yunhao Ge, Jie Ren, Andrew Gallagher, Yuxiao Wang, Ming-Hsuan Yang, Hartwig Adam, Laurent Itti, Balaji Lakshminarayanan, and Jiaping Zhao. Improving zero-shot generalization and robustness of multi-modal models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11093–11101, 2023. 2

[8] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 540–557, 2022. 1

[9] Agrim Gupta, Piotr Dollár, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017. 1

[11] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9404–9413, 2019. 1, 2

[12] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 1

[13] Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 1

[14] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted CLIP. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7061–7070, 2023. 1

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 1

[16] Mingxuan Liu, Tyler L. Hayes, Elisa Ricci, Gabriela Csurka, and Riccardo Volpi. SHiNe: Semantic hierarchy nexus for open-vocabulary object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16634–16644, 2024. 1

[17] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1

[18] Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[19] Timo Lüddecke and Alexander S. Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7086–7096, 2022. 1

[20] Zachary Novack, Julian McAuley, Zachary C. Lipton, and Saurabh Garg. CHiLS: Zero-shot image classification with hierarchical label sets. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 26342–26362, 2023. 2

[21] 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 International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 1

[22] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1

[23] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded SAM: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 1

[24] Xudong Wang, Shufan Li, Konstantinos Kallidromitis, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Hierarchical open-vocabulary universal image segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[25] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. SegGPT: Segmenting everything in context. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. 1

[26] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18134–18144, 2022. 1

[27] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2955–2966, 2023. 1

[28] Tianyang Xu et al. Cross-modal taxonomic generalization in (vision-) language models. *arXiv preprint arXiv:2603.07474*, 2026. 2

[29] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional CLIP. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 32215–32234, 2023. 1

[30] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. 3

[31] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017. 1

[32] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1

CGEBench: Benchmarking Concept Generalization of Promptable Image Segmentation Models

Supplementary Material

5. Generalizations

Table 1. Example generalization chains from CGEBench.

$\mathbf{P_{base}}$	$\mathbf{P_1}$	$\mathbf{P_2}$	$\mathbf{P_3}$
black high heel	high heel	shoe	footwear
large clear diamond	large diamond	diamond	gemstone
wood guitar	guitar	string instrument	musical instrument
red rose	rose	flower	plant
association football ball	soccer ball	sports ball	ball

Table 2. Most frequently occurring concepts at each prompt level.

	$\mathbf{P_{base}}$		$\mathbf{P_1}$		$\mathbf{P_2}$		$\mathbf{P_3}$	
	Concept	n	Concept	n	Concept	n	Concept	n
1	woman's hair	13	female hair	13	shoe	32	footwear	45
2	association football ball	12	soccer ball	12	percussion instrument	20	plant	41
3	young girl	9	long wavy hair	10	person	18	hair	39
4	baseball shoes	8	sneaker	10	hat	16	headwear	30
5	running shoe	6	drum	9	human hair	13	musical instrument	29
6	golfer	6	athlete	8	sports ball	13	human	25
7	association football goal	5	diamond	8	gemstone	13	fruit	22
8	long wavy brown hair	5	sports shoes	8	child	12	person	17
9	long dark brown hair	5	girl	8	flower	11	ball	16
10	dirt path	5	baseball cap	7	postage stamp	10	drink	16