PhraseStereo: The First Open-Vocabulary Stereo Image Segmentation Dataset

Thomas Campagnolo[1,2] , Ezio Malis[1] , Philippe Martinet[1] , Gaetan Bahl[2]

[1]Centre Inria d'Universite Cote d'Azur, France [2]NXP Semiconductors, France

{thomas.campagnolo, ezio.malis, philippe.martinet}@inria.fr,

{thomas.campagnolo, gaetan.bahl}@nxp.com

Abstract

Understanding how natural language phrases correspond to specific regions in images is a key challenge in multimodal semantic segmentation. Recent advances in phrase grounding are largely limited to single-view images, neglecting the rich geometric cues available in stereo vision. For this, we introduce PhraseStereo, the first novel dataset that brings phrase-region segmentation to stereo image pairs. PhraseStereo builds upon the PhraseCut dataset by leveraging GenStereo to generate accurate right-view images from existing single-view data, enabling the extension of phrase grounding into the stereo domain. This new setting introduces unique challenges and opportunities for multimodal learning, particularly in leveraging depth cues for more precise and context-aware grounding. By providing stereo image pairs with aligned segmentation masks and phrase annotations, PhraseStereo lays the foundation for future research at the intersection of language, vision, and 3D perception, encouraging the development of models that can reason jointly over semantics and geometry. The PhraseStereo dataset will be released online upon acceptance of this work.

1. Introduction

Interpreting natural language phrases with the corresponding specific regions in images is a core challenge in multimodal AI, for tasks such autonomous robot navigation, human-robot interaction, and more. While recent advances in phrase grounding have improved performance on single-view images, these methods often ignore the rich geometric information available in stereo vision. This limitation restricts their ability to reason about spatial relationships and depth, a key aspect of real-world understanding.

Despite the progress in grounding natural language phrases to image regions, existing works are predominantly designed for mono RGB images with datasets such as ReferIt [3], RefCOCO [13], Google Referring Expressions [7], and PhraseCut [11]. However, these datasets lack detailed geometric representations.
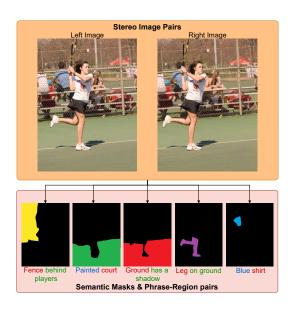

Figure 1. Stereo image pairs with corresponding semantic segmentation masks from PhraseStereo dataset. Each phrase-region pair is color-coded to highlight its linguistic meaning: attributes in blue, categories in red, and relationships in green.

The main contribution of this paper is the introduction of a new dataset about phrase grounding segmentation with stereo image pairs, enabling models to leverage stereo geometry for more accurate segmentation of referred objects and regions. PhraseStereo dataset most extends PhraseCut [11] to the stereo domain. It focuses on two important aspects: (1) the use of stereo images and (2) the consequently geometric spatial relationships between image pairs. PhraseStereo contains 77,262 stereo images and 345,486 phrase-region annotations, with multiple annotations per image pair, as illustrated in Fig. 1, following the structure and properties of PhraseCut dataset [11]. The right-view images are generated using GenStereo [9], a stereo generative model applied to the original single images. PhraseStereo's task involves segmenting regions in stereo image pairs based on natural language phrases, leveraging geometric information to improve spatial segmenta-

tion and address challenges inherent to single-image approaches, such as occlusions and depth ambiguities.

Our contributions are summarized as follows:
- We introduce PhraseStereo, the first dataset for open-vocabulary stereo semantic segmentation. PhraseStereo extends PhraseCut [11] to the stereo domain, enabling models to exploit geometric context from stereo vision for more precise segmentation of phrase-referred objects and regions.
- We address the challenge of hallucinations, relative to the right-view image generation step. To this end, we conducted a detailed analysis using perceptual similarity metrics, including SSIM and LPIPS.

2. Related Work

Table 1 presents a comparison of datasets related to grounding referring expressions. Kazemzadeh *et al*. [3] introduced the ReferItGame, a two-player interactive framework designed to collect the ReferIt dataset. In this setup, one participant generates natural language expressions to refer to specific objects within real-world scene images, while the other identifies the target object by clicking on its location. RefCOCO [13] also uses the ReferItGame, but focuses specifically on appearance-based descriptions (e.g., "the man in the yellow polka-dotted shirt") rather than positional or relational cues (e.g., "the second man from the left"), aiming to isolate visual grounding from spatial reasoning. Mao *et al*. [7] introduced Google RefExp, which builds upon the methodology of ReferIt [3] while utilizing the MSCOCO dataset [6]. Unlike ReferIt, MSCOCO provides instance-level segmentation annotations for 80 predefined object categories.

The SUN-Spot dataset [8] provides both RGB images and depth maps to support the localization of objects using spatial referring expressions. However, its use of depth is limited to interpreting spatial language involving prepositions like *"behind"* and *"in front of"*. Chen *et al*. [1] introduced the ScanRefer dataset, which builds upon ScanNet [2] by incorporating RGB-D scans where depth information is encoded as part of the 3D point cloud representation. Designed for 3D object localization from natural language descriptions, ScanRefer is limited to indoor environments, potentially restricting its generalization to more diverse scene types.

The PhraseCut dataset [11] is derived from Visual Genome [4] and grounds natural language phrases to image regions. It includes segmentation masks for the regions corresponding to each phrase, enabling fine-grained phrase-level supervision. In contrast, PhraseStereo encodes geometric information directly within stereo image pairs. To our knowledge, PhraseStereo is the only dataset that combines referring expression annotations with stereo image-based semantic segmentation, while also offering diverse

scenes across both indoor and outdoor scene images.

Dataset	Data Format	Tasks	Geometric Context
ReferIt [3]	Image	S	No
RefCOCO [13]	Image	S	No
Google RefExp [7]	Image	S	No
SUN-Spot [8]	Image	L	Yes (Depth map)
ScanRefer [1]	Image	L	Yes (3D point cloud)
PhraseCut [11]	Image	S	No
PhraseStereo (ours)	Stereo Images	S	Yes (Stereo encoded)

Table 1. Comparison between competing datasets. S indicates semantic segmentation tasks, L indicates localization tasks. PhraseStereo uniquely integrates referring expression segmentation with stereo-based 3D geometric vision.

3. PhraseStereo dataset

In this section, we describe the composition of PhraseStereo. Our (left) images and annotations are derived from the PhraseCut [11], which provides natural language phrases and corresponding segmentation masks. To extend this into the stereo domain, we generate the right-view images using GenStereo [9]. The complete data collection pipeline, illustrated in Fig. 2, consists of four main stages: image pre-processing, stereo image generation, right image post-processing, and data transferring and composition.

Image pre-processing The images of PhraseCut dataset [11] vary widely in resolution, ranging from small to very high dimensions. To ensure consistency and compatibility, in the pre-processing stage we resize the image to 512×512 pixels. This resolution offers a practical compromise: it preserves sufficient visual detail while standardizing input size across diverse image resolutions. Although GenStereo originally employed cropping or, more recently, patching strategies to meet its input size requirements [9], these methods either exclude parts of the image or introduce artifacts that compromise the visual integrity of the generated right-view image. In contrast, our resizing approach preserves the entire image content, ensuring better alignment and visual coherence in the resulting stereo pairs.

Stereo image generation To generate the right images in PhraseStereo, our pipeline employs GenStereo [9], a diffusion-based framework for stereo image synthesis. Given the (left-view) image from PhraseCut [11], I_L, the goal is to generate a high-quality right-view image I_R that preserves both visual fidelity and geometric consistency. GenStereo treats the left image as the reference and synthesizes the right image as the target. The corresponding disparity D is obtained from a predicted depth map using Depth Anything V2 [12], a monocular depth estimation model. This disparity information is used to compute disparity-aware coordinate embeddings, which, along with
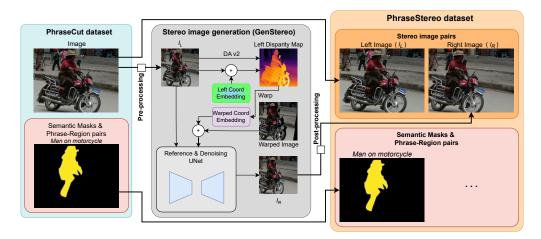
Figure 2. Overview of the PhraseStereo dataset generation pipeline. Starting from a PhraseCut [11] image I_L (left), we apply a pre-processing step including image resizing. GenStereo [9], is then used to generate the corresponding right-view image I_R (center). Post-processing includes resizing I_R back to the original resolution of I_L, ensuring stereo consistency. The resulting stereo pair (I_L, I_R), along with the phrase-region annotations pairs, forms the PhraseStereo dataset (right).

a warped version of the input image, conditions the diffusion process for improved stereo alignment. Additionally, GenStereo incorporates an adaptive fusion mechanism that blends the diffusion-generated image with the warped input, enhancing both realism and disparity consistency in the final stereo pair.

Right image post-processing After generating the right-view image I_R, the output resolution corresponds to the resized input image 512×512 pixels. To restore the original spatial alignment with the left-view image I_L, we perform a final resizing step. Knowing the original resolution of I_L, we resize I_R back to match it, ensuring that the stereo pair maintains consistent dimensions.

PhraseStereo composition PhraseStereo is composed of the original left-view image I_L from the PhraseCut dataset [11], the right-view image I_R generated by GenStereo [9], and the corresponding phrase-region annotations. Our final dataset includes 345,486 referring expressions across 77,262 stereo image pairs. Following the original PhraseCut splits, PhraseStereo is divided into 71,746 stereo image pairs for training with the corresponding 310,816 phrases, 2,971 stereo image pairs for validation (with the related 20,316 phrases), and 2,545 stereo image pairs with 14,354 phrases for testing. The phrase-region annotations are directly transferred from PhraseCut [11], preserving the original structure.

4. Results and Analysis

The PhraseStereo task involves generating a binary segmentation mask for a given input image pairs, conditioned on a referring natural language phrase and informed by stereo geometric cues.

To evaluate the quality of PhraseStereo dataset, a comparative analysis across multiple scale factors was conducted. In the context of stereo image generation, the parameter scale factor, which we denote as β controls the magnitude of disparity between the left and right views, effectively simulating the baseline distance between stereo cameras. A larger scale factor corresponds to a wider baseline. While this enhances depth cues, it also introduces several challenges. One of the most critical issues is the emergence and intensification of hallucinations, regions in the generated right image that do not correspond to any real content in the left image. These hallucinations typically arise when the model attempts to infer unseen areas, extrapolating geometry that is not visible in the input. However, due to the absence of corresponding real stereo data as ground-truth, i.e. stereo pairs where both views are captured from the same scene using a real stereo setup, we are unable to perform a direct analysis of hallucinations.

Instead, we focus on evaluating the perceptual and geometric quality of the generated stereo pairs. As the scale factor increases, the model struggles to maintain both perceptual realism and geometric consistency, leading to distorted object shapes, misaligned edges, and even the invention of non-existent structures. This compromises the overall realism of the stereo pair and poses difficulties in preserving semantic coherence.

For this analysis, we created three versions of the PhraseStereo dataset using scale factors $\beta \in \{0.05, 0.15, 0.25\}$. To assess the quality of the generated right images and address the issues described above, we adopted two complementary evaluation metrics: SSIM (Structural Similarity Index Measure) [10] and LPIPS (Learned Perceptual Image Patch Similarity) [14], using the *AlexNet* [5] backbone network. These metrics were chosen for their relevance in the
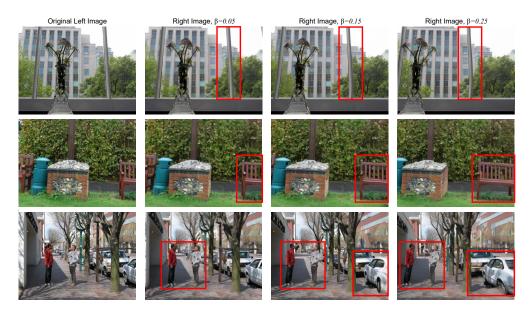
Figure 3. Qualitative analysis of right images in the PhraseStereo dataset across different scale factors. From left to right, each row shows the original left image [11], the generated right images at β increasing. Red rectangles highlight perceptual and geometric inconsistencies. PhraseStereo adopts $\beta = 0.15$ as a balanced configuration.

prior work GenStereo [9]. Table 2 presents the scores obtained across the different scale factors.

Scale Factor	SSIM ↑	LPIPS ↓
0.05	0.679	0.227
0.15 (final choice)	0.601	0.352
0.25	0.571	0.412

Table 2. Quantitative analysis of generated right images across different GenStereo [9] scale factors using SSIM and LPIPS metrics. Higher SSIM and lower LPIPS indicate better structural and perceptual similarity, respectively.

Our experiments show that as the scale factor increases, both perceptual and structural similarity to the reference image decrease in performance. Specifically, with $\beta = 0.05$, the generated images are perceptually close to the reference but exhibit minimal disparity, which limits their effectiveness for stereo learning. Higher scale factor, $\beta = 0.25$, introduces more pronounced disparity, but the generated images suffer from increased hallucinations and semantic drift, as reflected in the lower SSIM and higher LPIPS scores. Figure 3 presents the qualitative results. In the first row, the model struggles in geometric consistency, with distortions more evident at $\beta = 0.25$. The second row highlights fine textures, where missing details appear at $\beta = 0.05$ and $\beta = 0.25$, while $\beta = 0.15$ preserves them well. The third row features strong semantic cues, making hallucinations more noticeable at higher scale factors. We selected $\beta = 0.15$ as the optimal scale factor for PhraseStereo, a balanced value determined through both quantitative metrics and qualitative observations. This configu-

ration provides a realistic simulation of stereo camera systems, where moderate baselines are commonly employed to balance depth perception and image coherence. This choice ensures that PhraseStereo provides stereo pairs that are both challenging and realistic, making the dataset suitable for training and evaluating models in stereo vision and referring expression semantic segmentation.

5. Conclusion

We presented PhraseStereo, a novel dataset for phrase grounding segmentation in stereo image pairs. By enabling models to leverage stereo geometry, PhraseStereo can facilitate more accurate segmentation of referred objects and regions. It provides a foundation for exploring multimodal architectures that integrate vision and language in a stereo context, paving the way for advances in both geometric reasoning and semantic understanding.

The performance on the PhraseStereo dataset, particularly in the generation of stereo image pairs, remains limited due to the introduction of hallucinated content when synthesizing right-view images from monocular images. These hallucinations and the geometry consistency underscore the current gap between synthetic stereo generation and the real stereo data. To address this, a promising future direction of research is the integration of a stereo disparity matching module into the pipeline. After generating the right view, a stereo matching model can evaluate its consistency with the left view and compute a disparity loss to guide training. This stereo supervision encourages geometrically consistent outputs and improves depth prediction, helping bridge the gap between synthetic and real stereo data.

References

[1] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, pages 202–221, 2020. 2

[2] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 2

[3] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014. 1, 2

[4] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017. 2

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 25, 2012. 3

[6] 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 *ECCV*, pages 740–755, 2014. 2

[7] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 1, 2

[8] Cecilia Mauceri, Martha Palmer, and Christoffer Heckman. Sun-spot: An rgb-d dataset with spatial referring expressions. In *ICCVW*, pages 0–0, 2019. 2

[9] Feng Qiao, Zhexiao Xiong, Eric Xing, and Nathan Jacobs. Genstereo: Towards open-world generation of stereo images and unsupervised matching. *arXiv preprint arXiv:2503.12720*, 2025. 1, 2, 3, 4

[10] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 3

[11] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *CVPR*, pages 10216–10225, 2020. 1, 2, 3, 4

[12] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *NeurIPS*, 37:21875–21911, 2024. 2

[13] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85, 2016. 1, 2

[14] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 3