Robustness of Generative Models using Language Guidance for Low-Level Vision Tasks: Findings from Depth Estimation

Agneet Chatterjee[♢] Tejas Gokhale[♠] Chitta Baral[♢] Yezhou Yang[♢]
[♢]Arizona State University [♠]University of Maryland, Baltimore County
agneet@asu.edu, gokhale@umbc.edu, chitta@asu.edu, yz.yang@asu.edu

Abstract

Text to image generative models have recently been leveraged to perform monocular depth estimation, by incorporating natural language as additional guidance. Although yielding impressive results, the impact of the language prior, particularly in terms of generalization and robustness, remains unexplored. In this paper, we address this gap by quantifying the impact of this prior and introduce methods to benchmark its effectiveness across various settings. We generate "low-level" sentences that convey object-centric, three-dimensional spatial relationships, incorporate them as additional language priors and evaluate their downstream impact on depth estimation. Our key finding is that current language-guided depth estimators perform optimally only with scene-level descriptions and counter-intuitively fare worse with low level descriptions. Despite leveraging additional data, these methods are not robust to directed adversarial attacks and decline in performance with an increase in distribution shift. Finally, to provide a foundation for future research, we identify points of failures and offer insights to better understand these shortcomings. With an increasing number of generative models using language for depth estimation, our findings highlight the opportunities and pitfalls that require careful consideration for effective deployment in real-world settings.

1. Introduction

Breakthroughs in large-scale vision–language pretraining [10, 14, 19] and diffusion-based modeling techniques [15, 16] have been effective in significantly improving the state-of-the-art in *higher-level* semantic visual understanding tasks. *Lower-level* vision has a different perspective on image understanding and seeks to understand images in terms of geometric and physical properties of the scene such as estimating the depth and surface normals of each pixel in an image. Until now, state-of-the-art techniques for low-level vision tasks [11, 20] did not use natural language. Recently, generative models originally developed for high-level vision tasks have started to demonstrate exceptional results as for pixel-level dense prediction tasks such as semantic segmentation, as well as low-level tasks such as depth estimation [8, 9, 23]. Through the use of natural language, these methods seek to bridge the gap between high and low-level vision tasks.

In this paper, we walk this bridge and investigate generative models that perform language-guided monocular depth estimation, and ask a simple question - **what is the impact of the natural language prior, introduced by text to image generative models, in such a setting?** Our study is positioned to complement early exploration of generative models for low-level tasks, especially given the emerging evidence of state-of-the-art performance on tasks such as depth estimation. We create multi-modal transformations to evaluate the true low-level understanding of these models . We construct natural language sentences that encode low-level object-specific spatial relationships and generate image captions using pixel-level ground truth annotations. We perform image-level adversarial attacks implementing object-level masking, comparing vision-only and language-guided depth estimators on varying degrees of distribution shift.

Our contributions and findings are summarized below:

- We quantify the language guidance used by current generative methods for monocular depth estimation . We find that existing approaches possess a strong scene-level bias, and become less effective at localization when low-level information is provided. We additionally offer analysis grounded in foundation models to explain these shortcomings.
- Through a series of supervised and zero-shot experiments, we demonstrate that existing language-conditioned generative models are less robust to distribution shifts than vision-only models.
- We develop a framework to generate natural language sentences that depict low-level spatial relationships in an image by leveraging pixel and segmentation annotations.

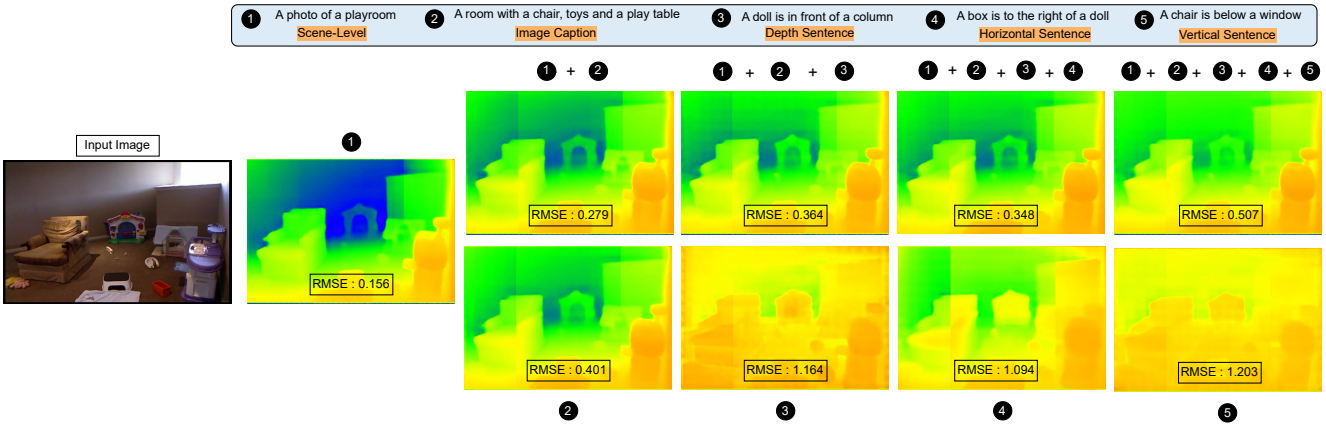

Figure 1. An illustration of language-guided depth estimation, showing depth maps generated by VPD (**zero-shot**) with the corresponding language guidance that we use as part of our study. The first row shows the effect of progressively adding descriptions as input, while the second row shows depth maps generated by single sentence inputs.

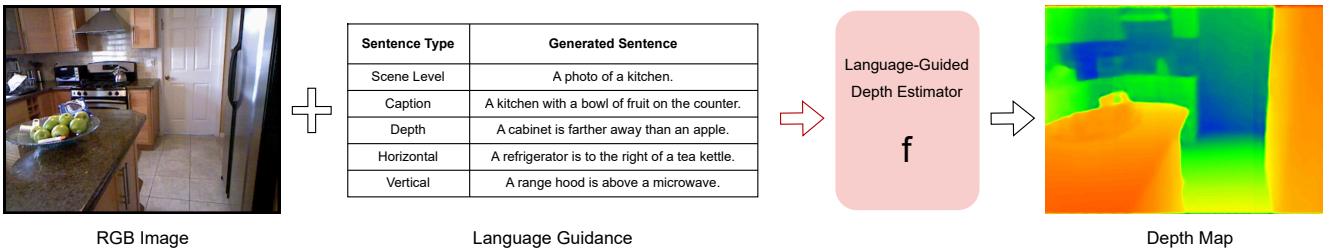

Figure 2. We systematically create additional knowledge for the depth estimator by leveraging intrinsic and low-level image properties. For each image we derive scene addendums, object and spatial level sentences and in supervised and zero-shot settings, quantify the effect of these sentences on monocular depth estimation.

2. Language-Guided Depth Estimation by Generative Models

2.1. Preliminaries

The use of natural language descriptions, by text to image generative models, to address low-level tasks is a new research direction. Although at a nascent stage, early evidence from depth estimation suggests that language can indeed improve the accuracy of depth estimators. This evidence comes from 3 recent approaches: VPD [23], TADP [8] and EVP [9] that show state-of-the-art results on standard depth estimation datasets such as NYU-v2 [12]. Our experiments are based on VPD thanks to open-source code.

The VPD model f takes as input an RGB image I and its ***scene-level*** natural language description S, and is trained to generate depth map D_I of the input image. VPD has a traditional encoder-decoder architecture where, the encoding block consists of:

(a) a frozen CLIP text encoder which generates text features of S, and

(b) a frozen VQGAN [3] encoder which generates features of I in its latent space.

The cross-modal alignment is learnt in the U-Net of the *Stable Diffusion* [16] model, which generates hierarchical feature maps. The prediction head, implemented as a Semantic FPN [7], is fed these feature maps for downstream depth prediction, optimizing the Scale-Invariant Loss.

2.2. Diverse Sentence Creation

Figure 2 depicts our workflow of sentence generation.

Sentences Describing Spatial Relationships : For a given image, our goal is to generate sentences that represent object-centric, low-level relationships in that image. Humans approximate depth through pictorial cues which includes relative relationships between objects. We focus on generating pairwise relationships between objects without creating complex sentences that would necessitate the model to engage in additional, fine-grained object grounding. These descriptions, which mirror human language patterns, explicitly contain depth information and could be potentially beneficial for improving depth estimation. Specifically, for all images I, we have semantic and depth ground-truth annotations at an instance and object-level across the dataset. Given this information, we generate sentences that

Sentence Type	δ_1 (↑)	δ_2 (↑)	δ_3 (↑)	RMSE (↓)	Abs. REL (↓)	Log_{10} (↓)
Scene-Level (Baseline)	**0.861**	**0.977**	**0.997**	**0.382**	**0.122**	**0.050**
Scene-Level + Low-Level	0.819	0.964	0.993	0.440	0.149	0.059
Only Low-Level	0.844	0.969	0.994	0.424	0.135	0.055

Table 1. Counter-intuitively, training with spatial sentences impairs performance compared to training with scene-level descriptions, limiting the efficacy of language-guided depth estimation.

Sentence Type	δ_1(↑)	δ_2(↑)	δ_3(↑)	RMSE (↓)	Abs. REL (↓)	Log_{10} (↓)
Scene-Level	**0.962**	**0.994**	**0.999**	**0.252**	**0.068**	**0.029**
Scene-Level + Caption *	0.950	0.993	0.998	0.279	0.076	0.033
Scene-Level + Caption + Depth *	0.932	0.992	0.998	0.311	0.084	0.037
Scene-Level + Caption + Depth + 2D *	0.864	0.973	0.993	0.403	0.109	0.050
Scene-Level + Caption $^{\oplus}$	0.916	0.986	0.997	0.347	0.092	0.041
Scene-Level + Caption + Depth $^{\oplus}$	0.878	0.980	0.994	0.399	0.105	0.048
Scene-Level + Caption + Depth + 2D $^{\oplus}$	0.849	0.973	0.994	0.443	0.115	0.053
Caption Only	0.827	0.961	0.988	0.474	0.127	0.059
Depth Only	0.372	0.696	0.878	1.045	0.284	0.153
Vertical Only	0.260	0.583	0.824	1.223	0.329	0.185
Horizontal Only	0.332	0.633	0.838	1.148	0.306	0.170

Table 2. In a zero-shot setting, VPD's performance is highest with baseline scene-level sentences. However, performance drops when more detailed, low-level information is introduced, as indicated by an increase in RMSE.

describe the spatial relationship between a pair of objects, in an image. We consider 3D relationships, i.e. depth-wise, horizontal and vertical relationships between an object pair, and thus the set of all spatial relationships R is defined as $\{front, behind, above, below, left, right\}$. Given I, and two objects A and B present in it, we create templatized sentences, which we share in the Appendix.

Image Captions : We generate captions corresponding to each image, which can be characterized as providing information in addition to scene level description. The goal is to test the model's performance with a detailed scene interpretation, more extensive than the baseline sentence S.

3. Measuring the Effect of Language Guidance

3.1. Supervised Experiments

In this setting, we answer, **does training on low-level language help?** We find that when trained and evaluated with additional low-level language, model performance decreases (Table 1). Apart from the baseline model, we train two more models s.t. for each I

(a) baseline sentence S and 1-3 supplementary sentences containing low-level relationships are used, and
(b) 4-6 sentences where only spatial relationships are used.

Compared to only low-level sentences, combining low-level with scene-level sentences deteriorates performance. This indicates that current approaches interpret language only when it is coarse-grained and require scene-level semantics for optimal performance.

# of Depth Sentences	δ_1 (↑)	δ_2 (↑)	δ_3 (↑)	RMSE (↓)	Abs. REL (↓)	Log_{10} (↓)
1	0.410	0.745	0.903	0.995	0.265	0.140
2	0.457	0.774	0.927	0.929	0.248	0.129
3	0.538	0.841	0.951	0.819	0.218	0.109
4	**0.582**	**0.869**	**0.965**	**0.770**	**0.205**	**0.101**

Table 3. VPD's performance when provided with multiple number of depth sentences. Overall performance is lower in comparison to baseline but iteratively improves as more sentences are provided, conveying better scene-level alignment.

3.2. Zero-Shot Findings

All zero-shot experiments are performed on the open-source VPD model. Language embeddings are generated via CLIP with an embedding dimension of 768, and image captions are generated using the BLIP-2-OPT-2.7b model [10].

Impact of Sentence Types: We evaluate VPD on our created sentences as shown in Table 2. To avoid ambiguity, we only consider sentences between unique objects in a scene. The original method averages out multiple scene-level descriptions, which are created using 80 ImageNet templates [22], and leverages the mean CLIP embedding as high level information. Following the original method, * in Table 2 represents the set-up, where for every I, we generate embeddings by stacking the mean baseline embedding and our sentence embeddings while in $^{\oplus}$, for every sentence $T \in$ the ImageNet Template, we concatenate T and our sentences, and compute its CLIP embedding. The distinction lies in the weighting : the former treats baseline and additional sentences equally, whereas the latter gives more weight to low-level sentences by virtue of them being added for each sentence in the template.

We re-affirm our initial findings through Table 2. the method maintains its optimal performance only when presented with scene-level sentences. The performance gradually worsens as additional knowledge (both high and low-level) is provided. Other forms of high-level language also seem to deteriorate performance. We observe a clear bias towards scene level description. For example, (Baseline + Caption) and (Caption Only) always outperform (Baseline + Caption + X) and (Depth/2D Only). This claim can be further underlined by the Δ decrease in performance from * to $^{\oplus}$, showing a distinct proclivity towards scene-level descriptions. In Figure 1, we present a visual illustration of the scene bias that persists in these methods.

Does Number of Sentences Matter? We find that using multiple low-level sentences, each describing spatial relationships, helps performance – performance is correlated with number of such sentences used. This can be attributed to more sentences offering better scene understanding. We find *again*, that model needs enough "scene-level" representation to predict a reasonable depth map as observed in Table 3. When the number of sentences is increased from one to four we observe a 41% increase and a 30% decrease

Relationship	Original Sentence	Relationship Switch	Object Switch	$\Delta_{orig.-rel.}$	$\Delta_{orig.-obj.}$
Horizontal	25.675	25.665	25.699	0.009	-0.024
Vertical	23.138	23.161	23.206	-0.023	-0.068
Depth	23.613	23.562	23.537	0.050	0.075

Table 4. CLIP struggles at differentiating between various spatial sentences, often producing higher scores for incorrect sentences spatial relationships.

in δ^1 and RMSE, respectively.

3.3. Potential Explanations for Failure Modes

The lack of understanding of spatial relationships of Diffusion-based T2I models is well studied by VISOR [4] and T2I-CompBench [6]. Studies [1] show that the cross-attention layers of Stable Diffusion lack spatial faithfulness to the input prompt; these layers itself are used by VPD to generate feature maps which could explain the current gap in performance. Similarly, to quantify CLIP's understanding of low-level sentences, we perform an experiment (Table 4) where we generate the CLIPScore [5] between RGB Images from NYUv2 and our generated ground-truth sentences. We compare the above score, by creating adversarial sentences where we either switch the relationship type or the object order, keeping the other fixed. We find that **a)** CLIPScore for all the combinations are low but more importantly, **b**) the Δ difference between them is negligible; with the incorrect sentences sometimes yielding a higher score.

4. Robustness and Distribution Shift

To assess performance under adversarial conditions, we setup the following experiments where we compare vision-only methods with VPD :

Masking : We perturb the image I in this setup, by masking an object in the image space. To offset the image-level signal loss, we include a language-level input specifying the precise relative position of the masked object with another object. We find that **vision-only models are more resilient to masking** in comparison to language-guided depth estimators. We compare AdaBins and VPD (Table 5) and find that the latter's Δ drop in performance is significantly more in comparison to its baseline performance. Despite leveraging additional information about the relative spatial location, VPD is less resilient in comparison to AdaBins. Following previous trends, we also find that the performance deteriorates significantly when scene-level information is removed. In these experiments, we compare VPD with AdaBins [2], MIM-Depth [21] and IDisc [13].

Scene Distribution Shift under the Supervised Setting: We define a new split of the NYUv2 dataset, where the train and test set have 20 and 7 non-overlapping scenes, with a total of 17k and 6k training and testing images. With this configuration, we train all the corresponding models and benchmark their results and adhere to all of the methods' original training hyper-parameters, only slightly reducing the batch size of IDisc to 12. Although VPD follows MIM-Depth as the 2nd-best performing model, we find that VPD has the largest performance drop amongst its counterparts, **107%**, when compared to their original RMSE (Table 6) . Since training is involved, we also allude to the # of trainable parameters to quantify the trade-off between performance and efficiency of the respective models. We present additional results in the Appendix.

This difference in performance between the two categories of models likely occurs because in language guided depth estimators, the model is forced to learn correlations between an in-domain and its *high-level* description. It cannot, therefore, map its learned representation to new data when an out-of-domain image with an unseen description is presented. On the contrary, vision-only depth estimators are not bound by any *language* constraints, and hence learn a distribution which better maps images to depth.

Model, Image	Sentence	$\Delta\ \delta_1$ (↓)	Δ RMSE (↓)	Δ Abs. REL (↓)
VPD	Scene-Level + Depth	0.062	0.093	0.024
VPD	Depth	0.586	0.794	0.213
AdaBins	N/A	**0.008**	**0.007**	**0.002**

Table 5. Under the masked image setting, we compare Δ decrease of VPD with AdaBins (vision-only depth estimator). AdaBins is significantly more robust to masked objects than VPD.

Method	Parameters (in Millions)	δ_1 (↑)	RMSE (↓)	$\Delta_{RMSE(original)}$%(↓)	Abs. REL (↓)
AdaBins	78	0.763	0.730	100.54	0.168
MIM-Depth	195	**0.872**	**0.527**	**83.62**	**0.115**
IDisc	209	0.836	0.609	94.56	0.129
VPD	872	0.867	0.547	107.48	0.121

Table 6. Comparison of VPD and Vision-only models in the supervised, scene distribution setting. When evaluated on novel scenes, VPD has the largest drop in performance, compared to its baseline.

5. Conclusion

The use of language guidance by text to image models opens new possibilities for bridging language and low-level vision. However, we find that current methods only work in a restricted setting with scene-level description. They do not perform well with low-level language, lack understanding of semantics and possess a strong scene-level bias. Compared to vision-only models, current language guided estimators are less resilient to directed adversarial attacks and show a steady decrease in performance with an increase in distribution shift. As low-level systems are actively deployed in real-world settings, it is imperative that these failures are addressed and the role of language is better understood. Our findings are a first step towards this and can be used as a means for better utilization of generative models in perception tasks.

References

[1] Anonymous. Spade : Training-free improvement of spatial fidelity in text-to-image generation. In *Submitted to The Twelfth International Conference on Learning Representations*, 2023. under review. 4

[2] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. AdaBins: Depth estimation using adaptive bins. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021. 4

[3] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2021. 2

[4] Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*, 2022. 4

[5] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 4

[6] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv:2307.06350*, 2023. 4

[7] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks, 2019. 2

[8] Neehar Kondapaneni, Markus Marks, Manuel Knott, Rogério Guimarães, and Pietro Perona. Text-image alignment for diffusion-based perception. *arXiv preprint arXiv:2310.00031*, 2023. 1, 2

[9] Mykola Lavreniuk, Shariq Farooq Bhat, Matthias Müller, and Peter Wonka. Evp: Enhanced visual perception using inverse multi-attentive feature refinement and regularized image-text alignment, 2023. 1, 2

[10] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1, 3

[11] Yue Ming, Xuyang Meng, Chunxiao Fan, and Hui Yu. Deep learning for monocular depth estimation: A review. *Neurocomputing*, 438:14–33, 2021. 1

[12] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2

[13] Luigi Piccinelli, Christos Sakaridis, and Fisher Yu. idisc: Internal discretization for monocular depth estimation, 2023. 4

[14] 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, 2021. 1

[15] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 1

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1, 2

[17] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 1

[18] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019. 1

[19] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 1

[20] Xiaolong Wang, David Fouhey, and Abhinav Gupta. Designing deep networks for surface normal estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 539–547, 2015. 1

[21] Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling, 2022. 4

[22] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 3

[23] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. *arXiv preprint arXiv:2303.02153*, 2023. 1, 2