On the Robustness of 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

Recent advances in monocular depth estimation have been made 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 methods using language for depth estimation, our findings highlight the opportunities and pitfalls that require careful consideration for effective deployment in real-world settings. [1]

1. Introduction

Computational theories of visual perception have proposed hierarchies of visual understanding, such as in the work of Gestalt psychologists [8], Barrow and Tenenbaum [2], and others. In this hierarchy, *higher-level* vision tasks are aligned with semantics or human-assigned labels; for instance, recognizing scenes, detecting objects and events, answering questions and generating captions about images, or retrieving and generating images from text queries. Breakthroughs in large-scale vision–language pretraining [27, 34, 42] and diffusion-based modeling techniques [35, 37] have been significantly improved the state-of-the-art in such higher-level semantic visual understanding tasks. This success has been driven by the idea that natural language is an effective and free-form way to describe the contents of an image and that leveraging this data for learning shared representations benefits downstream 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 (distance from the camera), surface normals (orientation relative to the camera) of each pixel in an image, or other localization and 3D reconstruction tasks. A physics-based understanding of the scene such as the measurement or estimation of geometric and photometric properties, underlies the solutions for these inverse problems. As such, until now, state of the art techniques [30, 43] for lower-level tasks such as depth estimation have not featured the use of natural language. Surprisingly, recent findings from VPD [52], TADP [24] and EVP [26] have demonstrated that language guidance can help in depth estimation, thus potentially building a bridge between low-level and high-level visual understanding. In this paper, we seek to inspect this bridge by asking a simple question: **what is the impact of the natural language prior on depth estimation?** Our study is positioned to complement early exploration into the role of natural language for training models for low-level tasks, especially given the emerging evidence of state-of-the-art performance on tasks such as depth estimation. We argue that in the age of large language models, it is essential to fully grasp the implications of using language priors for vision tasks and so we examine these methods from multiple perspectives to evaluate their complete potential. This examination is important as depth estimation is used for many real-world and safety-critical applications such as perception in autonomous vehicles, warehouse automation, and robot task and motion planning.

We conduct a systematic evaluation to determine the significance of language-conditioning and the effect of this conditioning under various settings. We also benchmark the generalization capabilities and robustness of these meth-

[1] Code/Data: https://github.com/agneet42/lang_depth

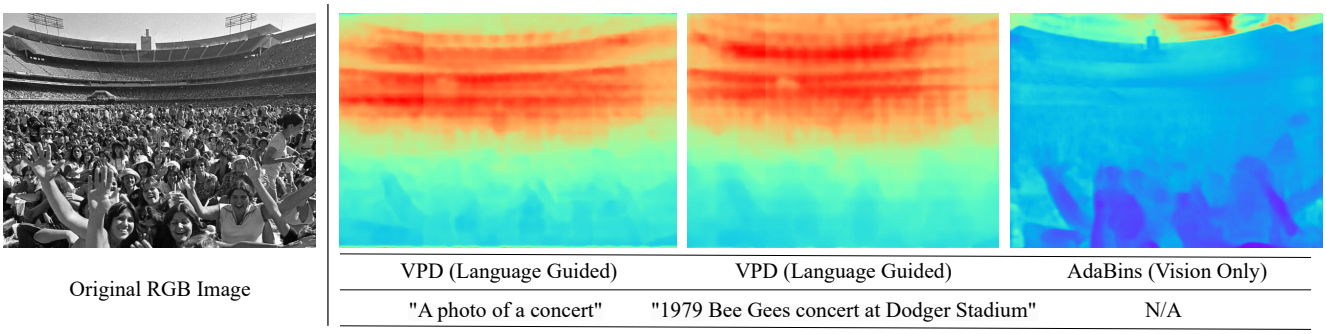

Figure 1. We investigate the efficacy of language guidance for depth estimation by evaluating the robustness, generalization, and spurious biases associated with this approach, comparing it alongside traditional vision-only methods. Shown here is a visual comparison of the depth estimation results between VPD (with additional knowledge) and AdaBins [3] on an out-of-domain outdoor scene.

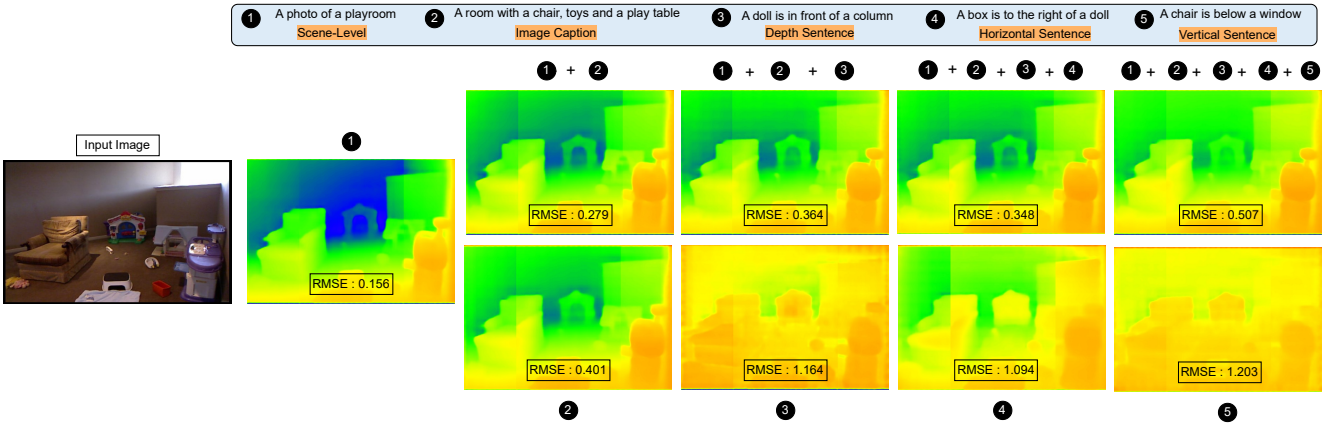

Figure 2. An illustration of depth maps generated by language-guided depth estimation methods such as VPD (**zero-shot**) when prompted with various sentence inputs that we use as part of our study. The first row shows the effect of progressively adding descriptions as input, while the second row shows depth maps generated by single sentence inputs.

ods. Specifically, we create image and language-level transformations to evaluate the true low-level understanding of these models . We construct natural language sentences that encode low-level object-specific spatial relationships, image captions and semantic scene descriptions using pixel-level ground truth annotations. Similarly, we perform image-level adversarial attacks implementing object-level masking, comparing vision-only and language-guided depth estimators on varying degrees of distribution shift.

Through our experiments, we discover that existing language-guided methods work only under the constrained setting of scene-level descriptions such as *"a photo of a bedroom"*, but suffer from performance degradation when the inputs describe relationships between objects such as *"a TV in front of a bed"*, and are unable to adapt to intrinsic image properties. Furthermore, with an increase in domain shift, these methods become less robust in comparison to vision-only methods. We present insights to explain the current challenges of robustness faced by these models and open up avenues for future research. A visual illustration of this performance gap is presented in Figure 1.

Our contributions and findings are summarized below:

- We quantify the guidance provided by language for depth estimation in current methods. 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 models are less robust to distribution shifts than vision-only models.
- We develop a framework to generate natural language sentences that depict low-level 3D spatial relationships in an image by leveraging ground truth pixel-wise and segmentation annotations.

Our findings underline the importance of taking a *deeper* look into the role of language in monocular depth estimation. We quantify the *counter-intuitive* finding that a sen-

tence such as "*The photo of a kitchen*" leads to better depth estimation than "*A knife is in front of a refrigerator*", although the latter explicitly contains depth information. In this paper, we provide trade-offs between accuracy and robustness of these methods, and identify points of failures for future methods to improve upon.

2. Related Work

Monocular Depth Estimation. Monocular depth estimation has a long history of methods ranging from handcrafted feature extraction to deep learning techniques and, recently, language-guided estimation. Saxena et al. [38] designed depth estimators as Markov Random Fields with hierarchical image features, Eigen et al. [9] pioneered convolution networks for this task, and Laina et al. [25] introduced fully convolutional residual networks for depth estimation. [33, 45, 50] aim to jointly learn pixel-level dense prediction tasks such as depth estimation, semantic segmentation, and surface normal estimation in a multi-task learning setting. Unsupervised methods [12, 15] have also been employed, using stereo images and multi-view perspective to guide learning. More recently, VPD [52], TADP [24], EVP [26] and DepthGen [39] have developed depth estimators that use diffusion-based models. Some standard datasets for monocular depth estimation include NYUv2 [31], KITTI [13], Sun RGB-D [40] and Cityscapes [7], which vary based on their settings (indoor/outdoor), annotation type (sparse/dense) and sensor types.

Language Guidance for Lower Level Vision Tasks. Excellent results have been attained by CLIP [34] and ALIGN [21] for zero-shot image recognition and OWL-ViT [29] for open-vocabulary object detection. GroupVIT and OpenSeg [14] perform open vocabulary semantic segmentation, supervised with high-level language, whereas ODISE [47] performs open-vocabulary panoptic segmentation. OWL-ViT [29] performs open-vocabulary object detection by aligning their pretrained, modality specific encoders with lightweight object-detectors and localization heads.

Geometric Guidance for High-Level Vision Tasks. Low-level visual signals have been incorporated to solve high-level vision tasks such as visual question answering (VQA), image captioning, and image generation. [1] make use of depth maps as additional supervision for VQA. For 3D captioning, Scan2Cap [5] proposes a message-passing module via a relational graph, where nodes are objects with edges capturing pairwise spatial relationship, while SpaCap3D [41] constructs an object-centric local 3D coordinate plan, incorporating spatial overlaps and relationships into its supervision. ControlNet [49] conditions text-to-image generation based on low-level feedback from edge maps, segmentation maps, and keypoints. Florence [48] and Prismer [28] leverage depth representations and semantic maps to develop general purpose vision systems.

Robustness Evaluation of Depth Estimation Models. Ranftl et al. [36] study the robustness of depth estimation methods, by developing methods for cross-mixing of diverse datasets and tackle challenges related to scale and shift ambiguity across benchmarks. [6, 51] demonstrate that monocular depth estimation models are susceptible to global and targeted adversarial attacks with severe performance implications. Our work studies the robustness of language-guided depth estimation methods.

3. Language-Guided Depth Estimation

The use of natural language descriptions to facilitate 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 two recent approaches: VPD (visual perception with a pre-trained diffusion model) and TADP (text-alignment for diffusion-based perception) that show state-of-the-art results on standard depth estimation datasets such as NYU-v2 [31]. Our experiments are based on VPD thanks to open-source code.

3.1. Preliminaries

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: $D_I = f(I, S)$. VPD has an encoder-decoder architecture. The encoding block consists of:

(a) a frozen CLIP text encoder which generates text features of S, which are further refined by an MLP based Text Adapter [11] for better alignment, and

(b) a frozen VQGAN [10] encoder which generates features of I in its latent space.

The cross-modal alignment is learnt in the U-Net of the Stable Diffusion model, which generates hierarchical feature maps. Finally, the prediction head, implemented as a Semantic FPN [23], is fed these feature maps for downstream depth prediction, optimizing the Scale-Invariant Loss.

Format of Language Guidance: Language guidance is provided via sentence S, which is a high-level description such as *"an* `[ADJECTIVE]` *of a* `[CLASS]`*"*, where `[ADJECTIVE]` could be (photo, sketch, rendering) and `[CLASS]` is one of the 27 scene labels (bedroom, bathroom, office etc.) that each of the images in NYUv2 belong to. For each scene type, variations of templated descriptions are generated, encoded via CLIP, and averaged to generate its embedding. Finally, each image I based on its scene type, is mapped to the generated embedding, which is considered to be high-level knowledge about I.

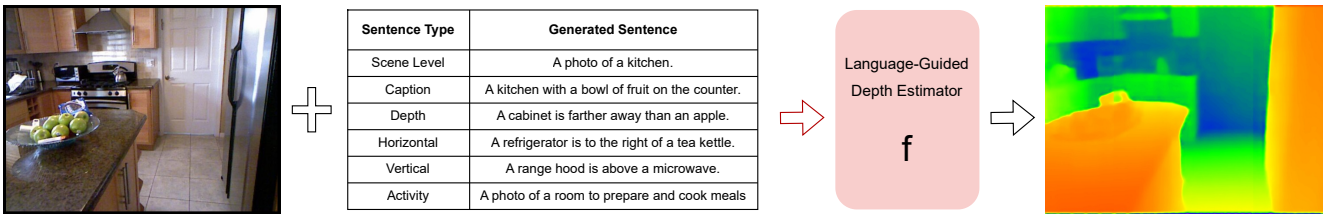

Figure 3. We systematically create additional knowledge for the depth estimator by leveraging intrinsic and low-level image properties. For each image we derive scene addendums, object and spatial level sentences along with semantic, activity based descriptions, and in supervised and zero-shot settings, quantify the effect of these sentences on monocular depth estimation.

3.2. Diverse Sentence Creation

While scene-level description have been successfully used for low-level tasks, the effect of different language types remains under-explored. In this subsection, we outline our framework of creating a diverse set of sentence types in order to evaluate and better understand the influence of language in improving depth estimation. We define S to be the baseline scene-level description (as used in VPD) of image I. Figure 3 displays our workflow of sentence generation.

3.2.1 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 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, twelve relationships can be generated between them as given below:

> *front(A,B), behind(A,B), front(B,A), behind(B,A), above(A,B), below(A,B), above(B,A), below(B,A), right(A,B), left(A,B), right(B,A), left(B,A)*

Relationship Extraction: For an object A, let (X_a, Y_a) be the coordinates of its centroid, R_a be it's maximum radius, (μ_a, σ_a, M_a) be the mean, standard deviation, and maximum object depth. Between A and B, a horizontal relationship is created if : $|(Y_a - Y_b) > \lambda \times (R_a + R_b)|$, where λ controls the amount of overlap allowed between the two objects. Thereafter, A is to the left of B if $(Y_a < Y_b)$, and otherwise. Similarly, a vertical relationship is created if : $|(X_a - X_b) > \lambda \times (R_a + R_b)|$ and A is above B if $(X_a < X_b)$. Finally, a depth relationship is created if : $|(\mu_a - \mu_b) > ((M_a - \mu_a) + (M_b - \mu_b))|$. Thus, A is closer than B if $(\mu_a + \sigma_a < \mu_b + \sigma_b)$, else A is farther than B. Having found R, we map the relationships into one of the templated sentences, as shown below:

> *A is in front of B, A is closer than B, A is nearer than B, A is behind B, A is farther away than B, A is more distant than B, A is above B, A is below B, A is to the right of B , A is to the left of B*

Similar templates have also been recently leveraged in evaluation of T2I Models [16, 20] and for language grounding of 3D spatial relationships [17].

Thus, given an image I, we test the model's performance by generating sentences that explicitly encode depth information and relative spatial locations of objects in the image. Intuitively, depth estimation could benefit from information about objects and their spatial relationships as compared to the scene-level descriptions.

3.2.2 Image Captions and Activity Descriptions

Image captions. We generate captions corresponding to each image, which can be characterized as providing information in addition to scene level description. The rationale is to evaluate the performance of the model, when furnished a holistic scene level interpretation, more verbose in comparison to the baseline sentence S. Captions may still capture object-level information but will not furnish enough low-level information for exact localization.

Activity descriptions. To test the semantic understanding of a scene, we modify the scene name in S, replacing it

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	<u>0.844</u>	<u>0.969</u>	<u>0.994</u>	<u>0.424</u>	<u>0.135</u>	<u>0.055</u>

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.

with a commonplace activity level description of a scene. For example, "A picture of a "*kitchen*" is replaced with "A picture of a "*room to prepare and cook meals*". Full list of transformations are presented in the Appendix. We curate these descriptions via ChatGPT.

To summarize, for an image I, we now possess sentences that delineate various aspects of it, encompassing object-specific, scene-specific, and semantic details. These `<image, text>` pairs are used in the next sections to quantify the impact of language.

4. Measuring the Effect of Language Guidance

In the following subsections, we quantify the importance of language conditioning in supervised and zero-shot settings. Following standard metrics, we report results on the Root Mean Square Error (RMSE), Absolute Mean Relative Error (Abs. REL), Absolute Error in log-scale (Log_{10}), and the percentage of inlier pixels δ^i with a threshold of 1.25^i (i=1,2,3). We use the flip and sliding window techniques during testing. For all subsequent tables, **Bold** and <u>underlined</u> values indicate best and second-best performance, respectively. All our sentences are generated on the 1449 (Train = 795, Test = 654) images from the NYUv2 official dataset, which contain dense annotations. The maximum depth is set to 10m and set $\lambda = 1$.

4.1. 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. We present examples in Figure 4.

4.2. Zero-Shot Findings

All zero-shot experiments are performed on the open-source VPD model. Language embeddings are generated via CLIP

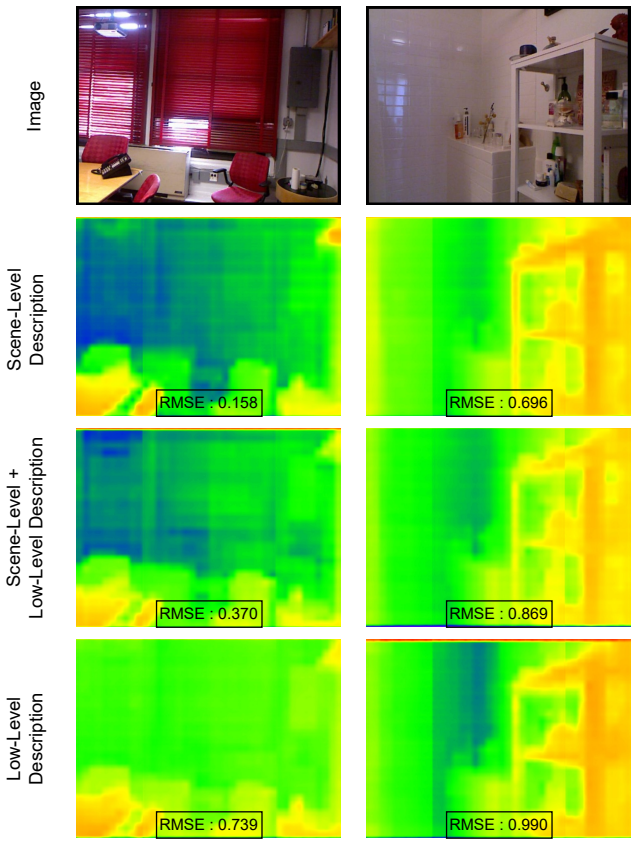

Figure 4. Comparison of depth maps across the three models trained under the supervised setting as described in Table 1. Low-level sentences induce hallucinations in the model; leading to large errors and false positive long-range depth estimates

with an embedding dimension of 768, and image captions are generated using the BLIP-2-OPT-2.7b model [27].

Impact of Sentence Types: We evaluate VPD on our created sentences as shown in Table 2. Sentences are generated for 518 images from the NYUv2 test split, considering only images where at least 1 depth, vertical, and horizontal sentence can be extracted. To avoid ambiguity, we only consider sentences between unique objects in a scene. As mentioned in Section 3, the original method averages out multiple scene-level descriptions, which are created using 80 ImageNet templates [46], 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 key differentiator is that in the former, the weight of the baseline (scene-level) description and the other sentences are equal, while in the latter, the low-level sentences

Sentence Type	$\delta_1(\uparrow)$	$\delta_2(\uparrow)$	$\delta_3(\uparrow)$	RMSE ($\downarrow$)	Abs. REL ($\downarrow$)	Log_{10} ($\downarrow$)
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.

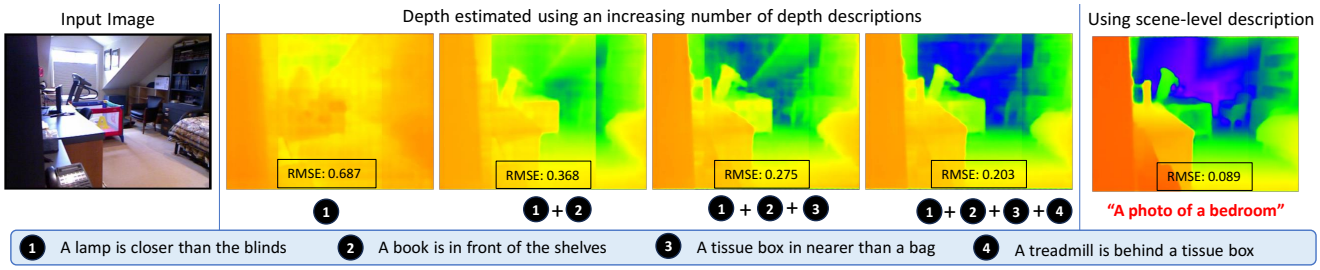

Figure 5. From **left** to **right**, as more bottom-up scene-level information is provided, the model's depth predictions move closer to the baseline predictions made with scene-level sentences. The plot below shows performance improvement across all metrics.

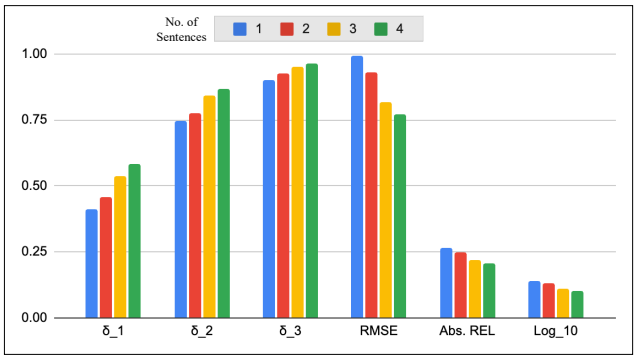

Figure 6. As more depth descriptions are provided, performance improves, signifying scene-level alignment.

have more prominence 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. Counter-intuitively, the performance gradually worsens as additional knowledge (both high and low-level) is provided. Even other forms of high-level language seem to deteriorate performance. Next, 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 2, we present a visual illustration of the scene bias that persists in these methods. The least performance drop is only seen in cases where additional high-level knowledge is provided. Iterative addition of low-level knowledge adversely affects the performance model's spatial comprehension. The model appears to completely apply a smooth depth mask in certain circumstances (such as vertical only), completely disregarding image semantics.

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 Figure 6. When the number of sentences is increased from 1 to 4 we observe a 41% increase and a 30% decrease in δ^1 and RMSE, respectively. We present an illustrative example in Figure 5, and observe that as the number of sentences are

Setup	δ_1 (↑)	δ_2 (↑)	δ_3 (↑)	RMSE (↓)	Abs. REL (↓)	Log_{10} (↓)
Scene Level	**0.963**	**0.994**	**0.998**	**0.254**	**0.069**	**0.029**
Activity Level	0.936	0.991	0.998	0.297	0.085	0.036

Table 3. In contrast to scene level sentences, semantics denoting activity level sentences result in a performance decline.

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.

increased, the depth prediction iteratively aligns with the predictions from the scene-level sentences.

Understanding of Semantics. When we use sentences at the activity level and compare it against the baseline scene-level sentences (Table 3), we see that the RMSE increases by 17%. Despite our transformations being extremely simple, we find mis-alignment between the semantic and scene space. A method exclusively tuned towards scene-level alignment, lacking semantic understanding would lead to unwanted failures, which would be particularly detrimental in a real-world setting.

4.3. Potential Explanations for Failure Modes

Language-guided depth estimation methods that we discussed above have **2** major components, a trainable Stable Diffusion (U-Net) Encoder and a frozen CLIP Encoder. We take a detailed look into each of them to find answers that explain these shortcomings.

The lack of understanding of spatial relationships of Diffusion-based T2I models is well studied by VISOR [16] and T2I-CompBench [20]. Studies [4] 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 where we generate the CLIPScore [18] 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 (~~Table 4~~) 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. (highlighted in red). Similar findings have been recently reported for VQA and image-text matching by Kamath et al. [22] and Hsu et al. [19].

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.

To summarize, while it is tempting to use language guidance in light of the success of LLMs, we show that current methods are prone to multiple modes of failures. In order to facilitate practical deployments, these techniques must be resilient and function well in novel environments and domain shifts, which we study next.

5. Robustness and Distribution Shift

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

Masking : As shown in Figure 7, 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 the following experiments, we compare VPD with AdaBins [3], MIM-Depth [44] and IDisc [32].

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.

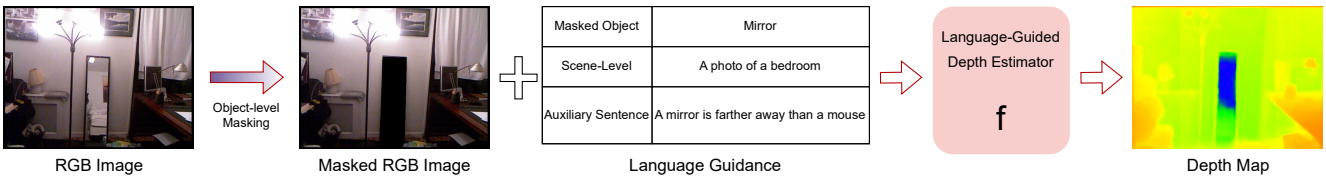

Figure 7. We mask an unique object in the image, using ground-truth segmentation annotations. To compensate for the loss of image signal, we provide additional knowledge about the masked object, followed by performing depth estimation.

Method	Params (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.

Method	δ_1 (↑)	RMSE (↓)	$\Delta_{RMSE(original)}\%$ (↓)	Abs. REL (↓)
AdaBins	0.768	0.476	30.76	0.155
MIM-Depth	**0.857**	**0.367**	27.87	0.132
IDisc	0.838	0.387	**23.64**	**0.128**
VPD	0.786	0.442	74.01	0.143

Table 7. Comparative results between VPD and Vision-only models while zero-shot testing on the Sun RGB-D dataset.

Zero-shot Generalization across Datasets: We perform zero-shot experiments on the models trained on NYUv2 and test its performance on Sun RGB-D [40], without any further fine-tuning. Sun RGB-D contains 5050 testing images across 42 different scene types. We create sentences of the form *"a picture of a* `[SCENE]`*"*.

We find that **language guided depth estimation methods struggle to generalize across image datasets**. VPD has a 20% higher RMSE in comparison to MIM-Depth, the best performing model (Table 7). This occurs even though Sun RGB-D and NYUv2 are both indoor datasets with a 50% overlap of scene type similarity.

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.

We also identify interesting correlations between the impact of low-level knowledge and the intrinsic properties of an image. As shown in Figure 8, the drop in performance of VPD is substantially high where the original RGB image has a large variation in depth.

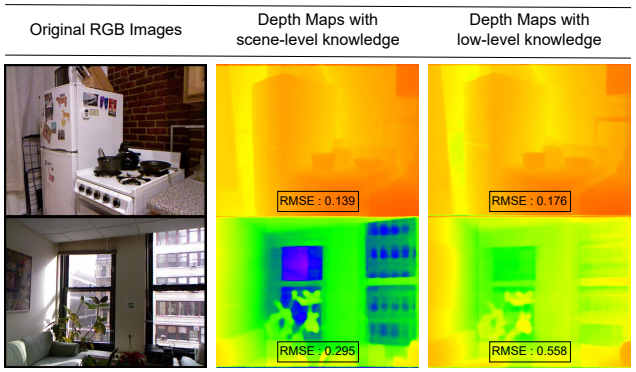

Figure 8. When provided with low-level knowledge such as, "a plant is to the right of a picture" about a scene with a large variation in depth values, VPD does not perform well (**bottom**), as compared to when the scene is more localized (**top**).

6. Conclusion

Applying natural language priors to depth estimation 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, but 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. An examination of the causes of these failures reveals that foundational models are also ineffective in this context. As low-level systems are actively deployed in real-world settings, it is imperative to address these failures and to investigate the role of language *in depth*. The findings from the paper could guide future work into better utilization of language in perception tasks.

Acknowledgements. The authors acknowledge Research Computing at Arizona State University for providing HPC resources and support for this work. This work was supported by NSF RI grants #1750082 and #2132724. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the funding agencies and employers.

References

[1] Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. Weakly supervised relative spatial reasoning for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1908–1918, 2021. 3

[2] H.G. Barrow and J.M. Tenenbaum. Computational vision. *Proceedings of the IEEE*, 69(5):572–595, 1981. 1

[3] 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. 2, 7

[4] Agneet Chatterjee, Yiran Luo, Chitta Baral, and Yezhou Yang. Spade : Training-free improvement of spatial fidelity in text-to-image generation, 2024. 7

[5] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3193–3203, 2021. 3

[6] Zhiyuan Cheng, James Liang, Hongjun Choi, Guanhong Tao, Zhiwen Cao, Dongfang Liu, and Xiangyu Zhang. Physical attack on monocular depth estimation with optimal adversarial patches. In *European Conference on Computer Vision*, pages 514–532. Springer, 2022. 3

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[8] Agné Desolneux, Lionel Moisan, and Jean-Michel Morel. Gestalt theory and computer vision. In *Seeing, thinking and knowing: Meaning and self-organisation in visual cognition and thought*, pages 71–101. Springer, 2004. 1

[9] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network, 2014. 3

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

[11] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters, 2021. 3

[12] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 740–756. Springer, 2016. 3

[13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 3

[14] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 3

[15] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 3

[16] 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, 7

[17] Ankit Goyal, Kaiyu Yang, Dawei Yang, and Jia Deng. Rel3d: A minimally contrastive benchmark for grounding spatial relations in 3d. *Advances in Neural Information Processing Systems*, 33, 2020. 4

[18] 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. 7

[19] Joy Hsu, Jiayuan Mao, Joshua B. Tenenbaum, and Jiajun Wu. What's left? concept grounding with logic-enhanced foundation models, 2023. 7

[20] 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

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 3

[22] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023. 7

[23] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks, 2019. 3

[24] 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, 3

[25] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016. 3

[26] 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, 3

[27] 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, 5

[28] Shikun Liu, Linxi Fan, Edward Johns, Zhiding Yu, Chaowei Xiao, and Anima Anandkumar. Prismer: A vision-language model with an ensemble of experts. *arXiv preprint arXiv:2303.02506*, 2023. 3

[29] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers, 2022. 3

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

[31] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 3

[32] Luigi Piccinelli, Christos Sakaridis, and Fisher Yu. idisc: Internal discretization for monocular depth estimation, 2023. 7

[33] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 283–291, 2018. 3

[34] 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, 3

[35] 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

[36] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 3

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

[38] Ashutosh Saxena, Sung Chung, and Andrew Ng. Learning depth from single monocular images. *Advances in neural information processing systems*, 18, 2005. 3

[39] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J. Fleet. Monocular depth estimation using diffusion models, 2023. 3

[40] 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. 3, 8

[41] Heng Wang, Chaoyi Zhang, Jianhui Yu, and Weidong Cai. Spatiality-guided transformer for 3d dense captioning on point clouds. *arXiv preprint arXiv:2204.10688*, 2022. 3

[42] 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

[43] 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

[44] Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling, 2022. 7

[45] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 675–684, 2018. 3

[46] 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. 5

[47] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 3

[48] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 3

[49] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 3

[50] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 235–251, 2018. 3

[51] Ziqi Zhang, Xinge Zhu, Yingwei Li, Xiangqun Chen, and Yao Guo. Adversarial attacks on monocular depth estimation. *arXiv preprint arXiv:2003.10315*, 2020. 3

[52] 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, 3