FlowDIS: Language-Guided Dichotomous Image Segmentation with Flow Matching
- Supplementary Material -

Andranik Sargsyan Shant Navasardyan
Picsart AI Research (PAIR)

A. Language-Prompt Generation Details

For language-prompt generation, we follow LawDIS [8] while further simplifying and enhancing their pipeline. Fig. S.2 illustrates the overall process. Specifically, we use the multi-modal language model GPT-4V [1] to generate two prompt types from the blacked-out background: a relatively detailed prompt (Prompt_1) and a shorter prompt (Prompt_2). We then employ GPT-4o-mini [4] to produce two additional paraphrased variants of the detailed prompt, denoted as Prompt_3 and Prompt_4. During training, we uniformly sample one of these four prompts to increase linguistic diversity. For quantitative comparisons at test time, we always use Prompt_1 to ensure determinism. Through manual inspection, we found that our language-prompt generation method achieves better alignment between the foreground and the corresponding language description. Therefore, in all quantitative comparisons with LawDIS, we used our prompts to ensure a fair evaluation.

B. Ablation for z^I Conditioning

We condition the velocity prediction model by concatenating z^I to its input, enabling access to the input image at intermediate denoising steps. This design improves fine segmentation details during multi-step inference. To validate its effect, we perform an ablation study on DIS-TE (1–4) with 2-step inference. As shown in Tab. S.1, this conditioning consistently improves all evaluation metrics.

Method	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$
Ours **w/o** additional z^I condition	0.933	0.954	0.017	0.948	0.972
Ours (i.e. **w/** z^I channel-wise concat)	**0.938**	**0.959**	**0.016**	**0.951**	**0.973**

Table S.1. Ablation for z^I conditioning on DIS-TE (1-4).

C. Further Ablation on PAIP

To further evaluate the effect of PAIP, we build a test set from the COCO [2] validation set, excluding stuff categories. We convert the instance annotations into semantic masks, resulting in 4,952 images with 14,246 binary masks, each corresponding to a semantic class. For each mask, we generate a text prompt using the pipeline described in Sec. A. The ablation results on this dataset are shown in Tab. S.2. As shown in the table, PAIP consistently improves all metrics, indicating better language controllability.

Method	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$
FlowDIS **w/o** PAIP	0.327	0.351	0.191	0.561	0.542
FlowDIS **w/** PAIP	**0.511**	**0.545**	**0.075**	**0.700**	**0.719**

Table S.2. Zero-shot ablation of PAIP on COCO-Object.

These quantitative improvements are also reflected in the qualitative examples shown in Fig. S.3. While FlowDIS without PAIP can struggle to follow the text prompt, the version with PAIP produces masks that better match the text description.

D. Comparison with Open-Vocabulary Semantic Segmentation Methods

Using the test set constructed in Sec. C, we compare FlowDIS with several state-of-the-art open-vocabulary semantic segmentation methods. In the zero-shot setting, FlowDIS achieves the best performance among the compared methods (see Tab. S.3).

Method	FlowDIS (Ours)	RF-CLIP [5]	SCLIP [7]	FreeCP [3]
mIoU ↑	**47.7%**	31.8%	28.3%	21.6%

Table S.3. Comparison with open-vocabulary semantic segmentation methods.

E. More Qualitative Comparisons

For a more comprehensive qualitative comparison, we compare our single-step results with other state-of-the-art methods on additional samples from the DIS5K test sets: DIS-TE1 (see Fig. S.4), DIS-TE2 (see Fig. S.5), DIS-TE3 (see Fig. S.6), DIS-TE4 (see Fig. S.7), and DIS-VD (see Fig. S.8). All results are generated with 1-step inference at 1024×1024 px resolution for a fair comparison.

Fig. S.9 shows additional samples demonstrating the language controllability of FlowDIS compared with the state-of-the-art method LawDIS [8].

F. Resolution Scaling

Inference res.	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$
1024×1024 px	0.919	0.946	0.024	0.939	0.964
1280×1280 px	0.925	0.952	0.023	0.944	0.965
1536×1536 px	0.928	0.953	0.022	0.945	0.966
1792×1792 px	0.931	0.955	0.022	0.946	0.966
2048×2048 px	**0.932**	**0.956**	**0.021**	**0.947**	**0.967**

Table S.4. Performance metrics of FlowDIS at different input resolutions on DIS-TE4.

We increase the inference resolution of FlowDIS beyond 1024×1024 px on DIS-TE4, the most challenging subset of DIS5K [6], which contains numerous objects with highly detailed structures. As shown in Tab. S.4, although FlowDIS was trained only at 1024×1024 px, its performance improves consistently with higher-resolution inference. Fig. S.1 shows qualitative results obtained with 2048×2048 px inference on samples with very high levels of detail.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018. 1

[3] Qi Chen, Lingxiao Yang, Yun Chen, Nailong Zhao, Jianhuang Lai, Jie Shao, and Xiaohua Xie. Training-free class purification for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23124–23134, 2025. 1

[4] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1

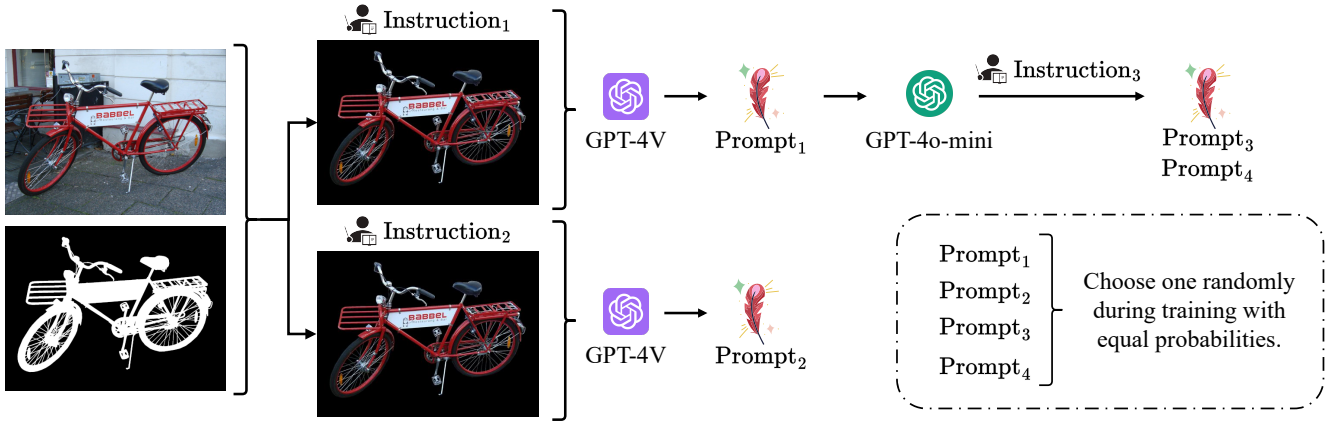
[5] Jiahao Li, Yang Lu, Yachao Zhang, Yong Xie, Fangyong Wang, Yuan Xie, and Yanyun Qu. Target refocusing via attention redistribution for open-vocabulary semantic segmentation: An explainability perspective. *arXiv preprint arXiv:2511.16170*, 2025. 1

[6] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *European Conference on Computer Vision*, 2022. 2

[7] Feng Wang, Jieru Mei, and Alan Yuille. SCLIP: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision*, pages 315–332. Springer, 2024. 1

[8] Xinyu Yan, Meijun Sun, Ge-Peng Ji, Fahad Shahbaz Khan, Salman Khan, and Deng-Ping Fan. LawDIS: Language-window-based controllable dichotomous image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23902–23911, 2025. 1, 2, 10

Figure S.1. FlowDIS results at 2048×2048 px resolution on highly detailed samples from DIS-TE4.



Instruction$_1$	Describe all visible objects in the image, excluding the black background. Keep the description concise, within 75 characters. If the objects are too complex or too numerous for one phrase, use two short sentences, but ensure the total length stays within 75 characters. For example: 'a red floral sculpture with green grass' or 'A white truck with aluminum ramps extending outward in a V-shape'.
Instruction$_2$	Generate a comma separated list of main objects in this image. Keep the description short. For example: 'red chair, wooden table', 'white car', 'tree', 'laptop, headphones', etc. Ignore the black background.
Instruction$_3$	You are provided with a prompt: '[Prompt$_1$]'. Please perform the following tasks: 1. Generate a synonymous prompt based on the given one as Prompt$_3$. 2. Simplify the given prompt while retaining all nouns, and generate it as Prompt$_4$.

Figure S.2. **Illustration of our language-prompt generation pipeline.** GPT-4V generates two prompts from the blacked-out background: a detailed prompt (Prompt$_1$) and a shorter prompt (Prompt$_2$). GPT-4o-mini then produces two paraphrased variants of the detailed prompt (Prompt$_3$ and Prompt$_4$). During training, one of the four prompts is uniformly sampled; at test time, Prompt$_1$ is used for determinism.

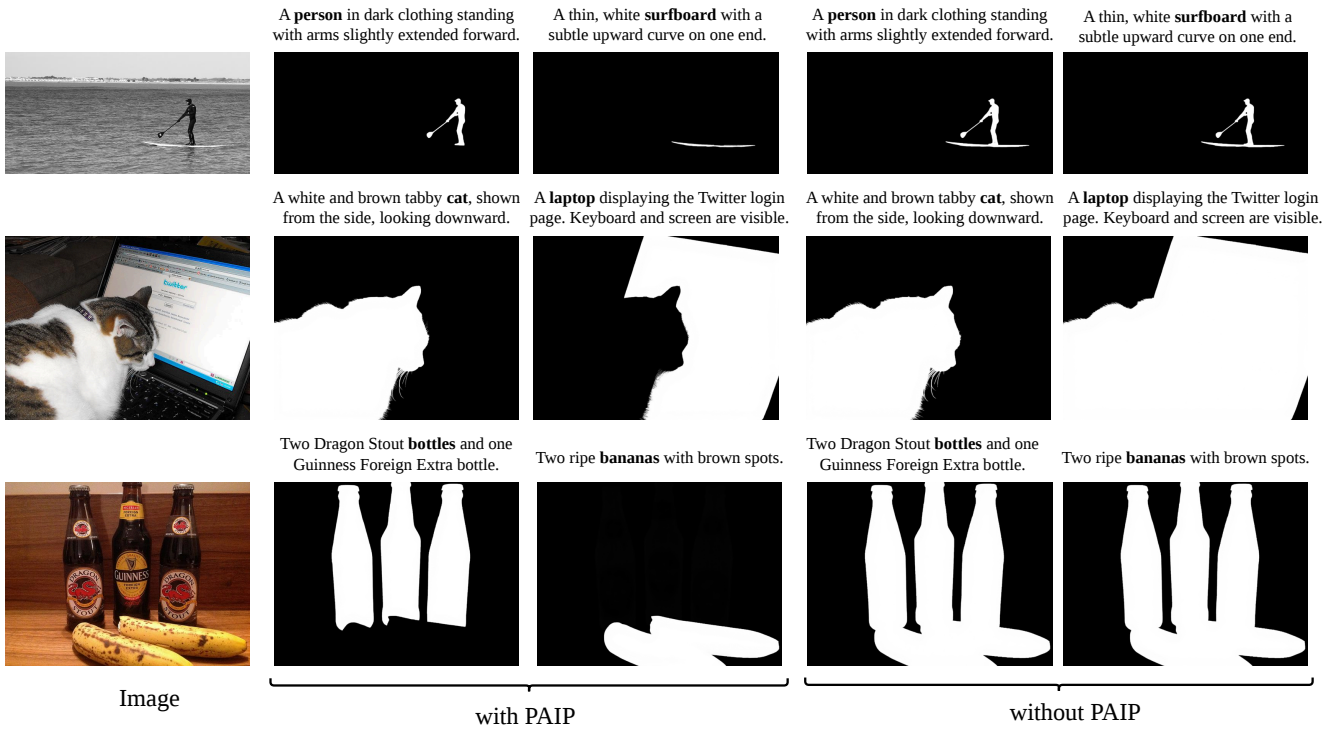


Figure S.3. Qualitative ablation study of PAIP on samples from the COCO dataset.

Figure S.4. Qualitative comparison with state-of-the-art DIS methods on DIS-TE1. Please zoom in to compare finer details.

Figure S.5. Qualitative comparison with state-of-the-art DIS methods on DIS-TE2. Please zoom in to compare finer details.

Figure S.6. Qualitative comparison with state-of-the-art DIS methods on DIS-TE3. Please zoom in to compare finer details.

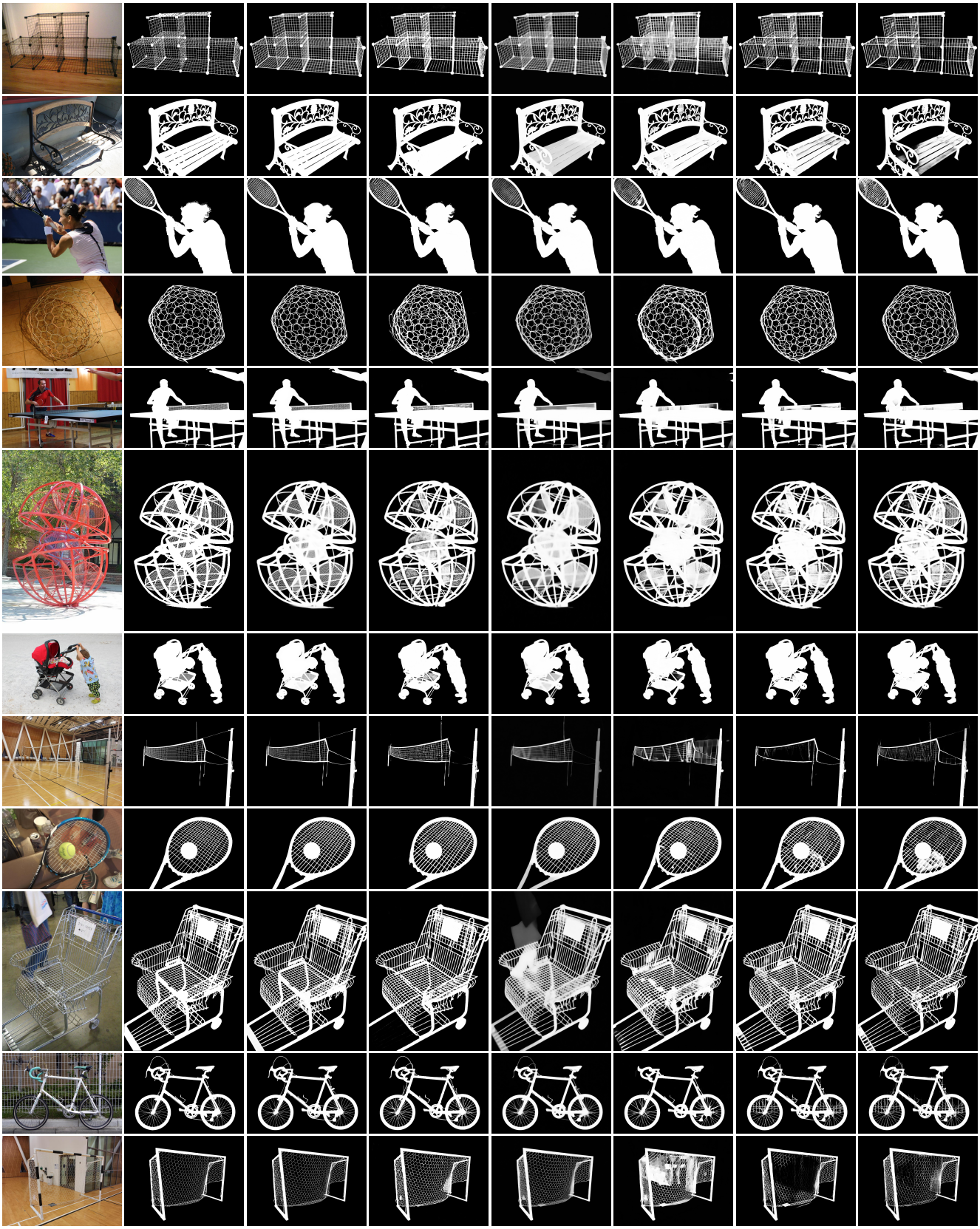


Figure S.7. Qualitative comparison with state-of-the-art DIS methods on DIS-TE4. Please zoom in to compare finer details.

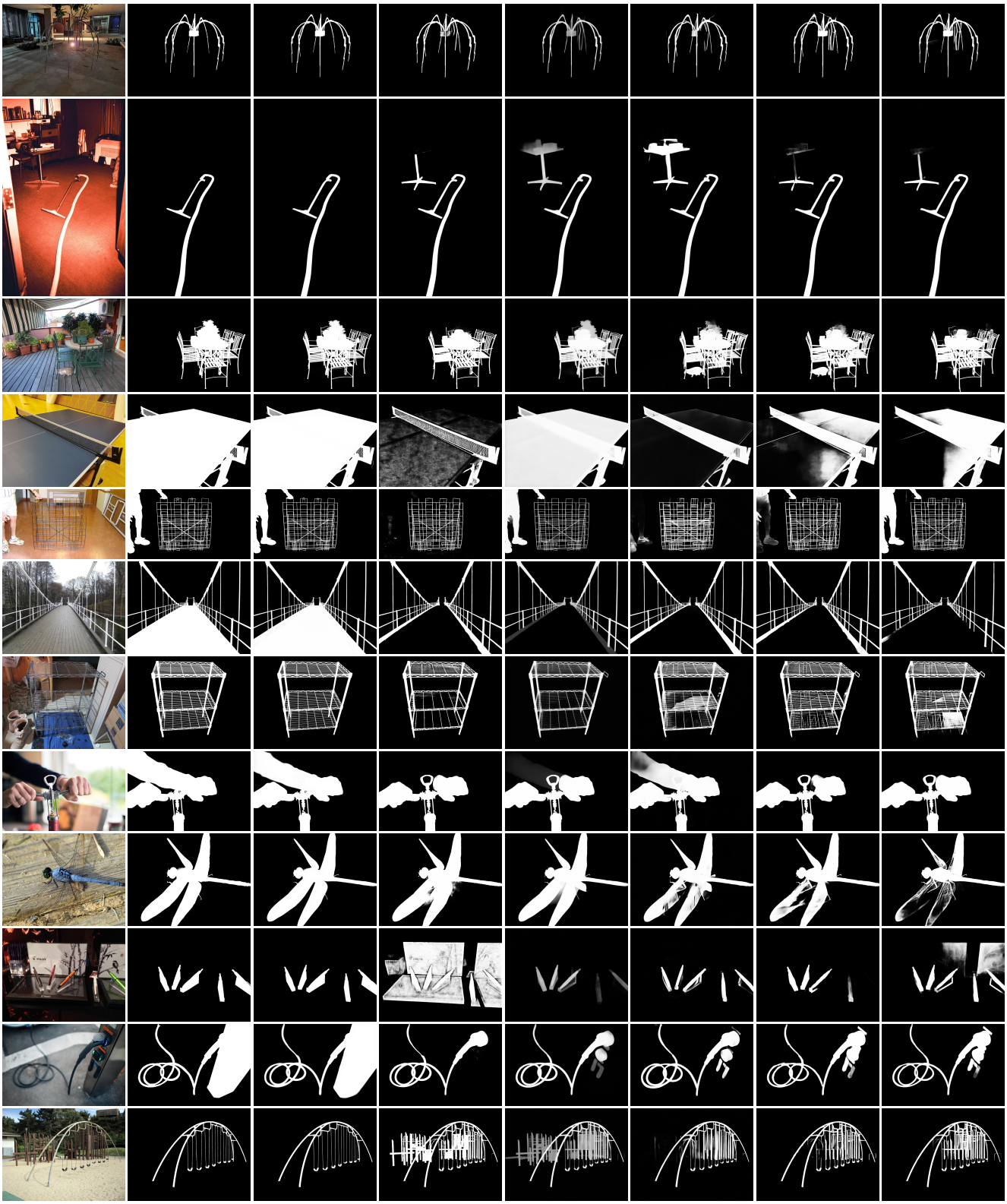


Figure S.8. Qualitative comparison with state-of-the-art DIS methods on DIS-VD. Please zoom in to compare finer details.

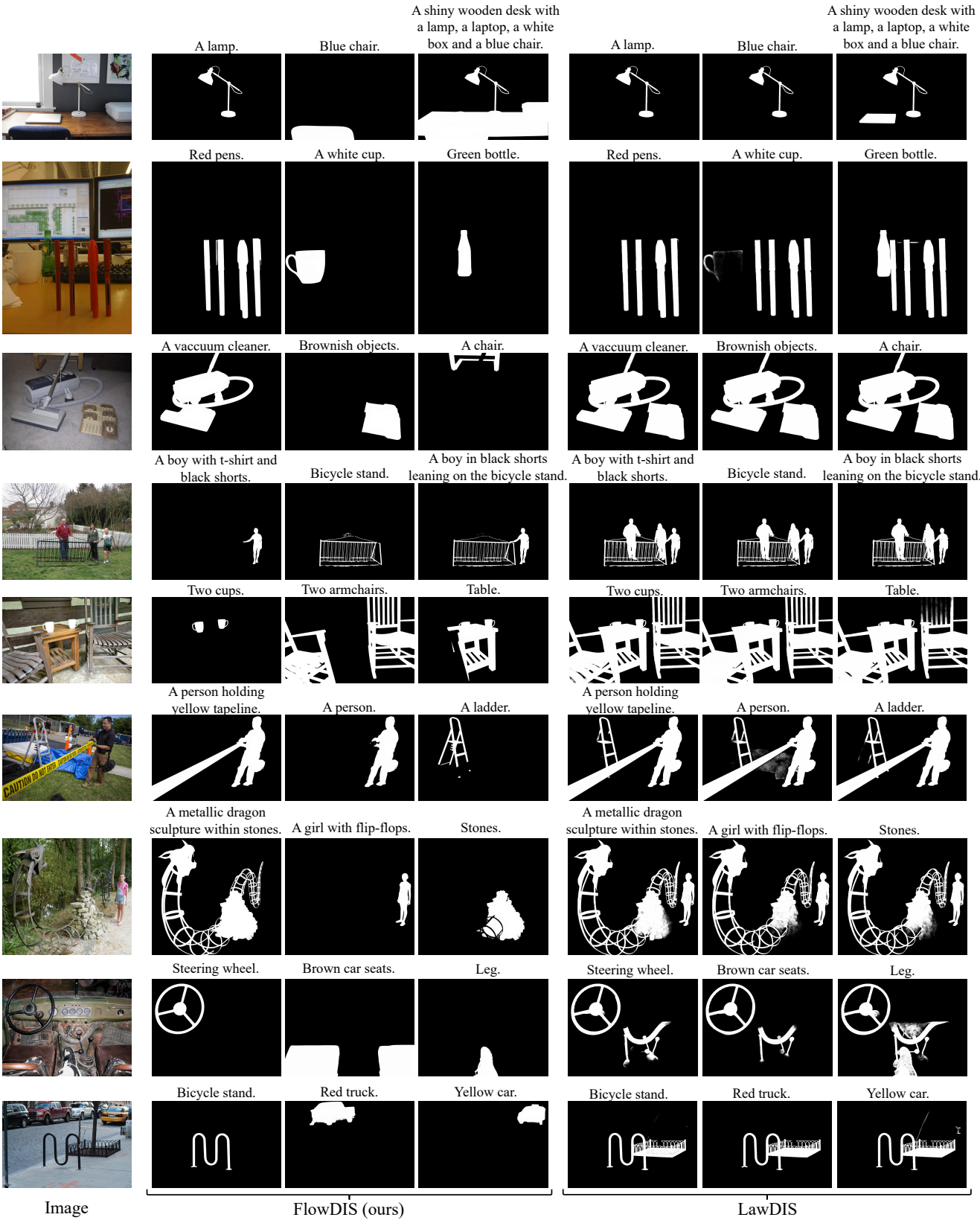


Figure S.9. Comparison of language controllability between our FlowDIS and LawDIS [8]. Each output is generated using the corresponding text prompt shown above.