TEXTBOKEH: LANGUAGE-GUIDED DEPTH-OF-FIELD EDITING WITH TARGET AND STRENGTH CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Depth of field controls which subject appears sharp and how strongly the rest of a scene is defocused. Existing post-capture interfaces commonly require a click, mask, focal distance, or numerical blur setting, and many are designed only for all-in-focus inputs. We present TEXTBOKEH, a language-guided depth-of-field editor conditioned on a focal-object description and a relative defocus level. We train one model to synthesize bokeh from all-in-focus images and refocus already-defocused images at a requested defocus level. We construct its supervision by choosing the Chebyshev center of the target's estimated disparity interval as the focal plane, then holding this plane fixed while solving for the range of f-numbers that keeps the target sharp, produces visible off-target defocus, limits excessive blur, and respects the source lens range. Sampling this interval yields an ordered, scene-adaptive stack of training targets. Across the retained aperture levels, TEXTBOKEH achieves substantially higher full-reference and perceptual fidelity than level-agnostic general-purpose editors on both AIF-to-Bokeh synthesis and cross-target refocusing. We will publicly release the trained LoRA adapter weights, the complete data synthesis pipeline, and the evaluation resources upon acceptance.

1 INTRODUCTION

A photograph does more than record a scene; it directs attention. By keeping one subject sharp and defocusing other depth layers, depth of field establishes visual hierarchy and separates the subject from its surroundings. Choosing a depth-of-field effect therefore decides not only how much blur a photograph contains, but what the viewer should look at.

Existing bokeh methods typically control depth of field through clicks, masks, focal-plane settings, or blur strength (Wang et al., 2018; Busam et al., 2019; Peng et al., 2022a; 2024; Chen et al., 2024). Recent photographic control methods built on text-conditioned generative models primarily target camera parameters or defocus strength (Fang et al., 2024; Fortes et al., 2025; Qin et al., 2026; Yang et al., 2025), while BokehFlow provides textual control over discrete focus regions, such as the foreground and background (Huang et al., 2026). Users, however, often wish to directly describe the specific object that should remain sharp—for example, *the child in red on the left*—and thereby determine the photograph's visual focus. We study depth-of-field editing centered on such object descriptions: the user specifies a focal target in natural language and selects a relative defocus level, and the model produces the corresponding depth-of-field effect.

Realizing this interaction requires addressing a key challenge: an object description identifies what should remain sharp but does not directly specify a suitable focal plane or aperture. An object may span multiple depths, so choosing a representative target depth as the focal plane does not by itself ensure that the valid target region remains sharp; the aperture also determines how strongly the surrounding scene can be defocused. We therefore construct training supervision subject to target-sharpness constraints, first determining the focal plane and then solving for a feasible aperture range. This construction provides ordered, scene-adaptive defocus levels for each language-specified target.

We first solve a subject-limited target-preservation problem: maximize the blur strength while keeping the valid target interior sufficiently sharp. Its closed-form solution places the focal plane at the midpoint of the target's estimated disparity range and determines the corresponding subject-limited

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

aperture. We then hold this minimax focal plane fixed and intersect the subject-sharpness bound with off-target-defocus and lens constraints. This construction gives a feasible f-number interval whose smallest value produces the strongest feasible bokeh at the fixed focal plane. Progressively stopping down the aperture then produces images at different defocus levels.

The resulting feasible interval anchors the relative strength scale to the scene and target geometry. Together with the fixed sampling rule, it assigns a rendered reference to each retained target–level condition. We use these target-specific aperture states to supervise TEXTBOKEH. As illustrated in Figure 1, one model generates ordered bokeh levels from an AIF image and edits an already-defocused image according to the requested focal target and defocus level.

Our main contributions are:

- We formulate target-preserving bokeh supervision: a closed-form minimax focal plane followed by a feasible aperture interval preserves the language-specified target while controlling off-target defocus.
- We construct a language-guided aperture-state dataset containing approximately 95K source photographs, 306K focal targets, 829K rendered states, and 1.83M training pairs.
- We train a unified editor for bokeh synthesis and refocusing, conditioned on a focal-target description and a relative aperture level.

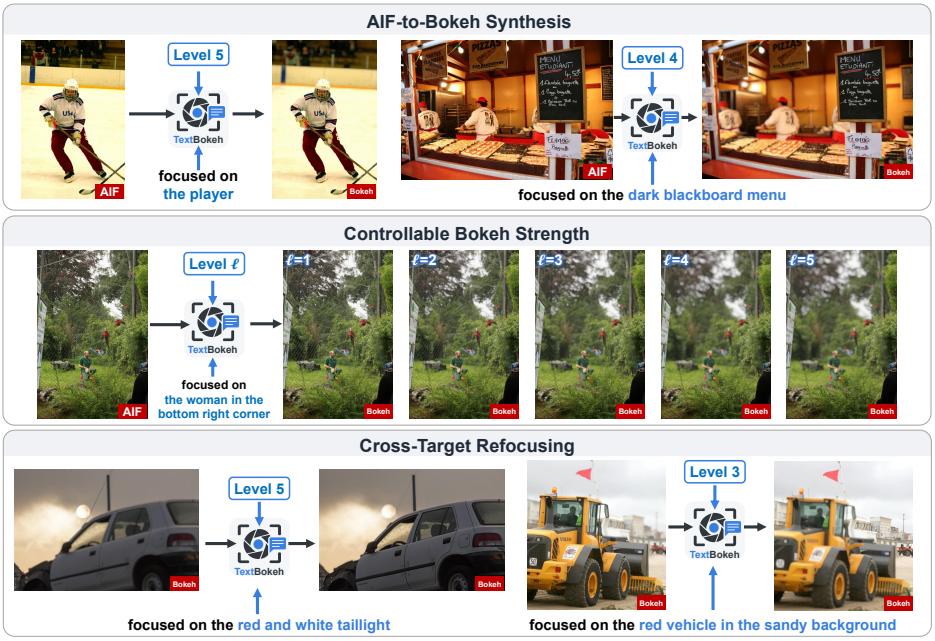

Figure 1: **Unified language-guided depth-of-field editing.** Given a focal-target description and a relative level ℓ, TEXTBOKEH synthesizes bokeh from an all-in-focus image and produces an ordered range of defocus strengths. From an already-defocused input, the same model shifts focus to a newly specified target at the requested defocus level.

2 RELATED WORK

Single-image depth-of-field rendering. Early methods combine monocular depth with spatially varying filtering or scattering (Yang et al., 2016; Wang et al., 2018; Busam et al., 2019). Neural renderers improve occlusion handling and reduce artifacts at depth boundaries (Peng et al., 2022a; 2024; 2022b; Sheng et al., 2024). Diffusion and flow-based approaches further improve perceptual quality and may avoid explicit depth at inference (Zhu et al., 2025; Huang et al., 2026). We use an existing renderer to generate training images; our contribution lies in choosing the focal plane and the admissible range of blur strengths.

Controllable focus and language-guided editing. Recent methods condition image generators on camera parameters, clicks, masks, or text (Qin et al., 2026; Fortes et al., 2025; Chen et al., 2024; Fang et al., 2024; Yang et al., 2025; Deng et al., 2024; Huang et al., 2026). BokehFlow (Huang et al., 2026) uses text to control focus regions and blur intensity for AIF inputs, while CamEdit (Qin et al., 2026) supports continuous camera-parameter editing. We focus on constructing object-specific supervision that accounts for the target's depth extent and provides an admissible range of defocus strengths, then learning transitions between the resulting states. Vision-language and image-generation models provide semantic localization and editing priors (Bai et al., 2023; Wang et al., 2024; Liu et al., 2023; Wu et al., 2025). Our condition combines an object-level text description with an ordered defocus level, and our training data vary both factors within each scene.

Refocusing from defocused inputs. Refocusing a defocused image is harder than adding bokeh to an all-in-focus image because detail at the new target may already be missing. Hardware-assisted methods use dual-pixel or multi-frame observations, whereas single-image methods depend on learned priors or staged reconstruction (Pan et al., 2021; Alzayer et al., 2023; Wang et al., 2023; Mu et al., 2025; Seizinger et al., 2025). GenFocus (Mu et al., 2025) decomposes refocusing into deblurring and bokeh synthesis, with the synthesis stage conditioned on a defocus map derived from depth and explicit focus and blur-strength settings. We study direct, single-image editing conditioned on a semantic target and a relative level, without requiring these geometric controls at inference.

3 Focus-State Formulation and Aperture Supervision

Let I_s be an all-in-focus (AIF) or defocused source image. A desired output state is specified by a semantic focal target q_t and an ordinal defocus level $\ell_t \in \{1, \ldots, 5\}$, and the editor predicts

$$I_t = G(I_s, q_t, \ell_t). \tag{1}$$

The target description and level specify the desired output, while the source focus state is inferred from the image.

Given an AIF image and a language-specified target q, let M_0 denote the target region and $M \subseteq M_0$ its valid interior, where we impose the sharpness constraint. In Figure 2, the target is the cat and M_0 covers the full animal. We determine a target-preserving focal plane and a feasible aperture interval, then sample reference states that vary the surrounding defocus.

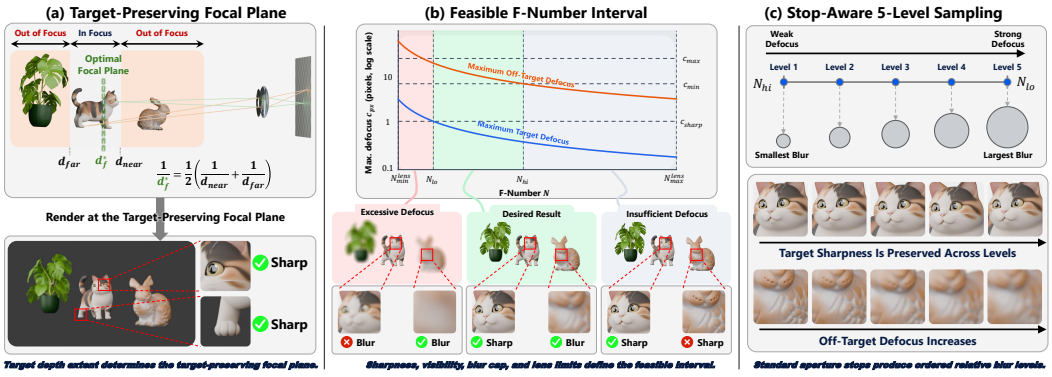

Figure 2: **Construction of target-specific aperture supervision.** (a) The subject-limited optimization places the focal plane at the midpoint of the valid target region's disparity range and gives the strongest blur that preserves this region. (b) At this fixed focal plane, target-sharpness, off-target-defocus, and lens constraints define the feasible interval $[N_{\mathrm{lo}}, N_{\mathrm{hi}}]$; N_{lo} gives the strongest feasible bokeh. (c) Stopping down the aperture produces weaker defocus states while preserving the same focal target.

3.1 SUBJECT-LIMITED CLOSED-FORM OPTIMUM

Let $d(x)$ be the estimated metric depth at pixel x, $\delta(x) = 1/d(x)$ its disparity, and δ_f the focal disparity. The valid target region M has disparity endpoints

$$\Delta_{\min} = \min_{x \in M} \delta(x), \qquad \Delta_{\max} = \max_{x \in M} \delta(x). \tag{2}$$

Let f_0 be the physical focal length, N the f-number, and p_0 the effective pixel pitch at the rendering resolution. Following the blur-radius convention of BokehMe (Peng et al., 2022a), the pixel-space defocus magnitude is

$$c_{\text{px}}(x; N, \delta_f) = \frac{f_0^2}{N p_0} |\delta(x) - \delta_f|. \tag{3}$$

Define the physical blur-strength factor $K = f_0^2/N$; increasing K, or equivalently decreasing N, strengthens defocus. The strongest blur that keeps every valid target-interior pixel below the sharpness threshold c_{sharp} is the joint optimization

$$\max_{K>0,\, \delta_f} K \quad \text{subject to} \quad \frac{K}{p_0} |\delta(x) - \delta_f| \le c_{\text{sharp}}, \quad \forall x \in M. \tag{4}$$

For a non-degenerate target disparity interval with $\Delta_{\max} > \Delta_{\min}$, this is a one-dimensional Chebyshev-center problem with the closed-form solution

$$\delta_f^* = \frac{\Delta_{\min} + \Delta_{\max}}{2}, \tag{5}$$

$$K_{\text{subj}}^* = \frac{2 p_0 c_{\text{sharp}}}{\Delta_{\max} - \Delta_{\min}}, \tag{6}$$

$$N_{\text{subj}} = \frac{f_0^2}{K_{\text{subj}}^*}. \tag{7}$$

As illustrated in Figure 2(a), the midpoint minimizes the maximum disparity deviation within the target, allowing the strongest blur under the target-sharpness constraint; this strength is inversely proportional to the target's disparity range.

3.2 PRACTICAL FEASIBLE APERTURE INTERVAL

At the fixed focal disparity δ_f^*, define the maximum off-target disparity deviation as $\delta_{\text{off}} = \max_{x \notin M_0} |\delta(x) - \delta_f^*|$. Requiring its defocus magnitude to lie between $c_{\min}$ and $c_{\max}$, while respecting target sharpness and the lens range $[N_{\min}^{\text{lens}}, N_{\max}^{\text{lens}}]$, gives the feasible interval in Figure 2(b):

$$N_{\text{lo}} = \max\left(N_{\min}^{\text{lens}}, \frac{f_0^2 \delta_{\text{off}}}{p_0 c_{\max}}, N_{\text{subj}} \right), \tag{8}$$

$$N_{\text{hi}} = \min\left(N_{\max}^{\text{lens}}, \frac{f_0^2 \delta_{\text{off}}}{p_0 c_{\min}} \right). \tag{9}$$

The interval is nonempty when $N_{\text{lo}} \le N_{\text{hi}}$. Since defocus increases as N decreases, N_{lo} gives the strongest feasible setting at the fixed minimax focal plane.

3.3 TARGET-SPECIFIC APERTURE LEVELS

With the target and focal plane fixed, we sample five candidate states uniformly in $\log N$ over the feasible interval and assign levels $\ell = 1, \ldots, 5$ from weak to strong defocus, as illustrated in Figure 2(c). Each level is ordinal within the scene- and target-specific stack, rather than a common f-number across photographs. Appendix A provides the full optimality proof, parameter configuration, and state discretization details.

4 DATASET CONSTRUCTION

Figure 3 summarizes our data construction pipeline: we screen AIF photographs, obtain depth and language-described focal targets, render target-specific aperture states using the construction in Section 3, and assemble them into editing pairs.

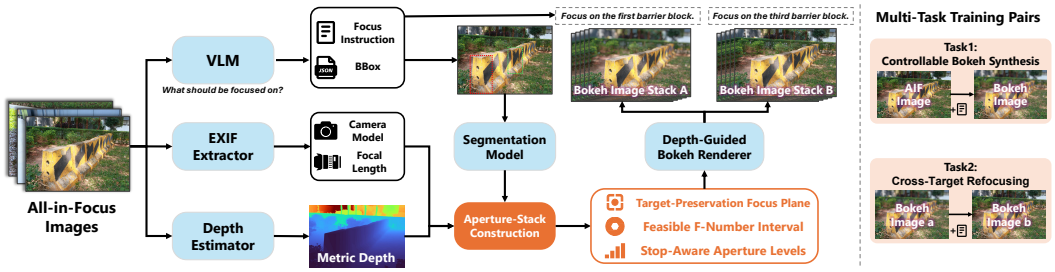

Figure 3: **Training-data construction.** Screened AIF photographs provide camera metadata, estimated depth, and language-described focal targets with masks. These inputs determine the target-specific aperture stacks, which yield AIF-to-Bokeh and cross-target Bokeh-to-Bokeh training pairs.

4.1 From AIF sources to aperture stacks

We start from a non-public image collection obtained through commercial procurement and assemble an initial pool of five million candidate photographs. Defocus screening and camera-parameter filtering yield 95,307 AIF source photographs with the required camera metadata. Details of the defocus screening procedure are provided in Appendix B.

For each source image, MoGe-2 (Wang et al., 2026) estimates metric-scale depth, Qwen3-VL[1] proposes focal objects and bounding boxes, and Segment Anything (SAM) (Kirillov et al., 2023) converts the boxes to masks. Gemini rewrites accepted proposals as self-contained focus instructions. We use the depth, masks, and camera parameters to construct aperture states following Section 3, and render the resulting stacks for valid targets.

4.2 From aperture stacks to editing pairs

From the rendered states $\{I(q_i, \ell_j)\}$ of each image, we construct two types of directed training pairs, shown on the right of Figure 3:

1. AIF images to rendered bokeh states;
2. bokeh states with different focal targets, at the same or different relative levels.

Each pair is conditioned on the output target description and level. The resulting dataset contains 1,829,353 AIF-to-Bokeh and cross-target Bokeh-to-Bokeh editing pairs. Appendices A–C provide filtering rules, model configurations, and dataset and transition statistics.

5 Unified Depth-of-Field Editor

5.1 Conditioning and backbone

We model the state transitions constructed in Section 4 as a unified conditional image-editing task. We combine the focal-target description q_t and the relative defocus level ℓ_t into a text prompt c_t. The editor receives the source image I_s and this prompt, and infers the source focus state directly from the image. We adapt Qwen-Image-Edit-2511 (Wu et al., 2025) using LoRA (Hu et al., 2022). In the training prompts, the ordinal defocus level is expressed as an aperture level. For example:

> *The camera is focused on the chrome front wheel of the pink vehicle. The photo is captured at aperture level 2 of 5.*

[1]We use Qwen3-VL-235B-A22B-Instruct: `https://huggingface.co/Qwen/Qwen3-VL-235B-A22B-Instruct`.

5.2 MULTI-STATE TRAINING

AIF-to-Bokeh pairs supervise bokeh synthesis at the requested level, and cross-target Bokeh-to-Bokeh pairs supervise refocusing at the requested output level. We jointly train on both pair types with a shared flow-matching objective.

Let z_0 be the latent encoding of target image I_t, let $\epsilon \sim \mathcal{N}(0, I)$, and sample a noise level σ. We form

$$z_\sigma = (1 - \sigma)z_0 + \sigma\epsilon, \tag{10}$$

and optimize the weighted velocity-prediction loss

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\left[w(\sigma)\left\|v_\theta(z_\sigma, I_s, c_t, \sigma) - (\epsilon - z_0)\right\|_2^2\right], \tag{11}$$

where c_t encodes the requested target and level, and $w(\sigma)$ is the loss weight assigned by the training scheduler. Training configurations are provided in Appendix C.

5.3 INFERENCE

At inference, users specify the focal target in natural language and select a relative defocus level from 1 to 5. The same editor takes an AIF or already-defocused image and predicts the requested output state. Users may request a single level or generate the five ordered candidates and select the preferred result.

6 EXPERIMENTS

Our main evaluations focus on AIF-to-Bokeh synthesis and cross-target refocusing, using separately constructed test sets with Open Images photographs (Kuznetsova et al., 2020) and BokehMe-rendered references built with the training-data pipeline. We also evaluate 96 pairs of GenFocus-generated AIF input proxies (Mu et al., 2025) and the corresponding original photographs with real lens blur.

6.1 EVALUATION SETUP

The AIF-to-Bokeh and cross-target refocusing sets contain 457 and 370 scene-target cases, with 1,822 and 1,331 valid level-specific reference pairs, respectively. All methods are evaluated at 1024-pixel resolution using PSNR and SSIM (Wang et al., 2004) (higher is better), and LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2020) (lower is better).

We compare with general-purpose editors that directly accept a source image and a focal-target instruction. Qwen, FLUX, HY, GPT, and Gemini denote Qwen-Image-Edit-2511 (Wu et al., 2025), FLUX.1 (Black Forest Labs et al., 2025), Hunyuan-Image 3.0 (Tencent Hunyuan Foundation Model Team, 2025), GPT-Image 2.0, and Gemini 2.5 Flash Image, respectively. Baseline scope and reproducibility are detailed in Appendix D, and task-enhanced comparison prompts in Appendix E.

TEXTBOKEH produces an output at the requested aperture level and is evaluated against the reference at that level. Exact-level Avg., our primary result, averages all valid requested-level pairs. Best Level selects, for each scene-target case, the prediction-reference pair with the lowest LPIPS among the corresponding levels. Each baseline produces one output per focal-target instruction without a level condition; Best-GT matches that output to the available reference with the lowest LPIPS. Best Level and Best-GT are reference-based oracle results; all metrics are computed on the same LPIPS-selected pair. Full protocol details are provided in Appendix D.

6.2 EVALUATION RESULTS

AIF-to-Bokeh synthesis. Table 1 shows that TEXTBOKEH achieves higher reference fidelity than the general-purpose baselines. Across 1,822 requested-level prediction-reference pairs, its Exact-level Avg. reaches 31.49 dB PSNR, 0.915 SSIM, 0.071 LPIPS, and 0.033 DISTS. Under Best-GT matching, Hunyuan, the strongest baseline on these metrics, obtains 23.22 dB PSNR and 0.182 LPIPS. Per-level results are reported in Table 7.

Table 1: AIF-to-Bokeh synthesis on Open Images. Exact-level Avg. is the primary requested-level result over 1,822 valid pairs from 457 cases. Best Level and baseline Best-GT use LPIPS oracle selection; all metrics are evaluated on the same selected pair.

Method	Protocol	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
TEXTBOKEH	**Exact-level Avg. (Primary)**	31.49	0.915	0.071	0.033
TEXTBOKEH	Best Level	32.39	0.928	0.050	0.024
Qwen	Best-GT	15.96	0.542	0.490	0.195
FLUX	Best-GT	11.63	0.363	0.604	0.167
HY	Best-GT	23.22	0.822	0.182	0.073
Gemini	Best-GT	17.26	0.558	0.361	0.126
GPT	Best-GT	21.63	0.735	0.254	0.101

Figure 4 demonstrates selection of specific objects through appearance and spatial descriptions: the distant blue-glass tower among several buildings, and the small TVR badge on the green car. In both examples, TEXTBOKEH keeps the requested target sharp while producing stronger off-target defocus at L5 than at L1. The badge example also shows that the nearby headlight remains sharp at a similar depth, although it is outside the named target. This illustrates depth-consistent focus beyond the selected object; GPT instead blurs the headlight beside the sharp badge. The target crops and full images show that TEXTBOKEH retains the tower's appearance and the car's layout during these edits. Figure 15 further demonstrates selecting different focal objects from the same AIF input by changing the target description at a shared relative level.

The five-level sequence in Figure 1 shows outputs generated from the same AIF input, with the requested target kept sharp as surrounding defocus increases from L1 to L5. Figures 11 and 12 provide further multi-level examples with different targets and scene geometries. These examples illustrate the relative strength control learned from the target-specific reference stacks.



Figure 4: **Qualitative AIF-to-Bokeh comparison at Levels 1 and 5.** The instructions specify the distant blue-glass tower on the left and the TVR logo badge on the green car. Target and off-target crops show the response to the requested level; the rightmost crops compare scene preservation and the depth consistency of defocus.

Table 2: Cross-target refocusing on Open Images. Exact-level Avg. is the primary requested-level result over 1,331 valid pairs from 370 cases. Best Level and baseline Best-GT are LPIPS oracle diagnostics, following Table 1.

Method	Protocol	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
TEXTBOKEH	**Exact-level Avg. (Primary)**	30.80	0.904	0.076	0.036
TEXTBOKEH	Best Level	31.48	0.914	0.059	0.029
Qwen	Best-GT	14.78	0.488	0.519	0.198
FLUX	Best-GT	12.07	0.357	0.605	0.165
HY	Best-GT	24.05	0.808	0.213	0.080
Gemini	Best-GT	18.39	0.581	0.316	0.110
GPT	Best-GT	24.04	0.793	0.192	0.076

Cross-target refocusing. Table 2 evaluates refocusing from already-defocused inputs. Across 1,331 requested-level pairs, TEXTBOKEH obtains 30.80 dB PSNR, 0.904 SSIM, 0.076 LPIPS, and 0.036 DISTS under exact-level evaluation. Its LPIPS is lower than the best baseline value of 0.192, obtained by GPT under Best-GT matching. Per-level results are provided in Table 8.

Figure 5 shows how object descriptions redirect focus within an existing scene. In the museum example, the instruction identifies the background pedestrian by his light-colored checked shirt and bag. TEXTBOKEH brings this initially blurred target into focus while defocusing the previously sharp car, with stronger blur on the car at L5 than at L1. In the kayak example, the color and spatial description identifies the stack against the brick wall rather than the large orange kayak in the foreground. The output preserves the scene layout, sharpens the requested background region, and increases foreground defocus at the stronger level. The crops show the complementary changes at the new target and the original focal region. By comparison, Qwen and FLUX alter people or scene composition in these examples, while several other baseline outputs retain much of the original focus state. Further target descriptions and intermediate output levels are shown in Figure 13. Additional qualitative results in Figure 14 demonstrate language-guided refocusing on real camera-captured photographs, including background targets, object parts, and individual subjects among similar objects.

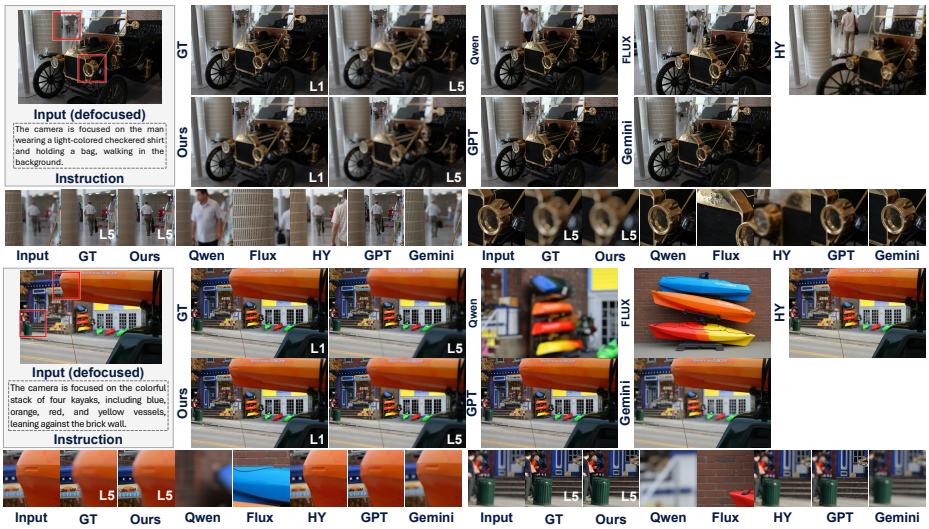

Figure 5: **Cross-target refocusing at Levels 1 and 5.** The instructions identify a background pedestrian and a stack of kayaks. TEXTBOKEH sharpens the requested target and varies defocus in the original focal region; crops compare both regions with the input and level-specific references.

Table 3: Evaluation on 96 real-defocus reference pairs. Best-Pred selects among the five TEXTBOKEH outputs by LPIPS after inference and is an oracle diagnostic; each baseline produces one output.

Method	Output	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
TEXTBOKEH	Best-Pred	27.45	0.838	0.149	0.054
Qwen	Single	14.10	0.396	0.582	0.209
FLUX	Single	12.49	0.279	0.583	0.173
HY	Single	23.96	0.771	0.223	0.088
Gemini	Single	16.87	0.403	0.353	0.113
GPT	Single	22.13	0.691	0.241	0.086

Evaluation with real-defocus references. Table 3 evaluates 96 pairs of AIF input proxies and original photographs with real lens blur. With Best-Pred selecting the lowest-LPIPS result among five outputs, TEXTBOKEH obtains 27.45 dB PSNR, 0.838 SSIM, 0.149 LPIPS, and 0.054 DISTS; each baseline produces one output. Figure 6 shows that the selected outputs reproduce key spatial features of the real blur: the foreground people are defocused while the specified woman across the table remains sharp, and the mixer in front of the watch-wearing hand is blurred while the hand remains in focus. These results support the ability of the generated candidate set to approximate real photographic defocus beyond the synthetic references. Additional examples and protocol details are provided in Appendix D and Figure 10.



Figure 6: **Qualitative comparison with real photographic defocus.** Each input is an AIF proxy paired with the corresponding real-defocus photograph, which serves as GT. Given the focal-target instruction, TEXTBOKEH recovers the intended focal subject and the main spatial distribution of defocus.

7 CONCLUSION

We formulate text-guided depth-of-field editing in terms of a focal target and a relative defocus level. Our data construction chooses a target-preserving focal plane and samples only f-numbers that satisfy target-sharpness, off-target-defocus, and lens constraints. On the Open Images test sets, the resulting editor obtains 31.49 dB PSNR and 0.071 LPIPS for AIF-to-Bokeh synthesis, and 30.80 dB PSNR and 0.076 LPIPS for cross-target refocusing. Its candidate focus states also approximate real photographic defocus references.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing and to support the retrieval and discovery of relevant literature. Qwen3-VL and Gemini assisted synthetic dataset construction through focal-target proposal and focus-instruction generation, as detailed in Section 4 and Appendix B. The authors take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Upon acceptance, we will publicly release the trained LoRA adapter weights for TEXTBOKEH, the complete data synthesis pipeline, and the evaluation resources used in this work. The pipeline release will include code, prompts, configuration files, and instructions covering AIF screening, camera-parameter filtering, depth estimation, focal-target proposal and segmentation, focus-instruction generation, target-specific aperture-state construction and rendering, and editing-pair generation. We will also release the trained defocus estimator used for AIF screening and document the third-party model dependencies and their configurations. This will enable researchers to construct training data from their own source photographs. The original commercially procured training image collection will remain non-public. For evaluation, we will provide source-image URLs, annotations, prompts, configuration files, and code to reconstruct the ground-truth references and reproduce the evaluation.

REFERENCES

Hadi Alzayer, Abdullah Abuolaim, Leung Chun Chan, Yang Yang, Ying Chen Lou, Jia-Bin Huang, and Abhishek Kar. DC 2: Dual-camera defocus control by learning to refocus. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21488–21497. IEEE, 2023.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Benjamin Busam, Matthieu Hog, Steven McDonagh, and Gregory Slabaugh. Sterefo: Efficient image refocusing with stereo vision. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3295–3304. IEEE, 2019.

Kang Chen, Shijun Yan, Aiwen Jiang, Han Li, and Zhifeng Wang. Variable aperture bokeh rendering via customized focal plane guidance. *arXiv preprint arXiv:2410.14400*, 2024.

Jieren Deng, Xin Zhou, Hao Tian, Zhihong Pan, and Derek Aguiar. GBSD: Generative bokeh with stage diffusion. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7070–7074. IEEE, 2024.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

I-Sheng Fang, Yue-Hua Han, and Jun-Cheng Chen. Camera settings as tokens: Modeling photography on latent diffusion models. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Armando Fortes, Tianyi Wei, Shangchen Zhou, and Xingang Pan. Bokeh Diffusion: Defocus blur control in text-to-image diffusion models. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yachuan Huang, Xianrui Luo, Qiwen Wang, Liao Shen, Jiaqi Li, Huiqiang Sun, Zihao Huang, Wei Jiang, and Zhiguo Cao. BokehFlow: Depth-free controllable bokeh rendering via flow matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5167–5175, 2026.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment Anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Chun-Wei Tuan Mu, Cheng-De Fan, Jia-Bin Huang, and Yu-Lun Liu. Generative refocusing: Flexible defocus control from a single image. *arXiv preprint arXiv:2512.16923*, 2025.

Liyuan Pan, Shah Chowdhury, Richard Hartley, Miaomiao Liu, Hongguang Zhang, and Hongdong Li. Dual pixel exploration: Simultaneous depth estimation and image restoration. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4338–4347. IEEE, 2021.

Juewen Peng, Zhiguo Cao, Xianrui Luo, Hao Lu, Ke Xian, and Jianming Zhang. BokehMe: When neural rendering meets classical rendering. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16262–16271. IEEE, 2022a.

Juewen Peng, Jianming Zhang, Xianrui Luo, Hao Lu, Ke Xian, and Zhiguo Cao. MPIB: An mpi-based bokeh rendering framework for realistic partial occlusion effects. In *European Conference on Computer Vision*, pp. 590–607. Springer, 2022b.

Juewen Peng, Zhiguo Cao, Xianrui Luo, Ke Xian, Wenfeng Tang, Jianming Zhang, and Guosheng Lin. BokehMe++: Harmonious fusion of classical and neural rendering for versatile bokeh creation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1530–1547, 2024.

Xinran Qin, Zhixin Wang, Fan Li, Haoyu Chen, Renjing Pei, Wenbo Li, and Xiaochun Cao. CamEdit: Continuous camera parameter control for photorealistic image editing. *Advances in Neural Information Processing Systems*, 38:114152–114171, 2026.

Qwen Team. Qwen3 technical report, 2025. URL `https://arxiv.org/abs/2505.09388`.

Tim Seizinger, Florin-Alexandru Vasluianu, Marcos V Conde, Zongwei Wu, and Radu Timofte. Bokehlicious: Photorealistic bokeh rendering with controllable apertures. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8908–8917. IEEE, 2025.

Yichen Sheng, Zixun Yu, Lu Ling, Zhiwen Cao, Xuaner Zhang, Xin Lu, Ke Xian, Haiting Lin, and Bedrich Benes. Dr. Bokeh: Differentiable occlusion-aware bokeh rendering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4515–4525. IEEE, 2024.

Tencent Hunyuan Foundation Model Team. HunyuanImage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

Chao Wang, Ana Serrano, Xingang Pan, Krzysztof Wolski, Bin Chen, Hans-Peter Seidel, Christian Theobalt, Karol Myszkowski, and Thomas Leimkühler. A neural implicit representation for the image stack: Depth, all in focus, and high dynamic range. *ACM Transactions on Graphics*, 42(6), 2023.

Lijun Wang, Xiaohui Shen, Jianming Zhang, Oliver Wang, Zhe Lin, Chih-Yao Hsieh, Sarah Kong, and Huchuan Lu. DeepLens: Shallow depth of field from a single image. *arXiv preprint arXiv:1810.08100*, 2018.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate monocular geometry with metric scale and sharp details. *Advances in Neural Information Processing Systems*, 38:35928–35959, 2026.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Qirui Yang, Yang Yang, Ying Zeng, Xiaobin Hu, Bo Li, Huanjing Yue, Jingyu Yang, and Peng-Tao Jiang. CameraMaster: Unified camera semantic-parameter control for photography retouching. *arXiv preprint arXiv:2511.21024*, 2025.

Yang Yang, Haiting Lin, Zhan Yu, Sylvain Paris, and Jingyi Yu. Virtual dslr: High quality dynamic depth-of-field synthesis on mobile platforms. *Electronic Imaging*, 28:1–9, 2016.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Chengxuan Zhu, Qingnan Fan, Qi Zhang, Jinwei Chen, Huaqi Zhang, Chao Xu, and Boxin Shi. BokehDiff: Neural lens blur with one-step diffusion. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1369–1379. IEEE, 2025.

A Optical Derivation and Configuration

A.1 Defocus model and unit conventions

We measure distances and focal length in millimeters, disparity in mm^{-1}, effective pixel pitch in millimeters per pixel, and the renderer's pixel-space defocus magnitude in pixels. Under the thin-lens approximation, the image-space defocus extent is proportional to the aperture diameter and to the disparity difference from the focal plane. We use the BokehMe convention

$$s(d; N, d_f) = \frac{f_0^2}{N}\left(\frac{1}{d} - \frac{1}{d_f}\right), \tag{12}$$

where the sign indicates on which side of the focal plane a point lies. Dividing its magnitude by the effective pixel pitch p_0 gives the pixel-space defocus magnitude in Equation 3. If the image is resized by a factor r, then $p_0 = p_{\mathrm{native}}/r$.

A.2 Subject-limited optimality proof

Let $m = (\Delta_{\min} + \Delta_{\max})/2$ and $h = (\Delta_{\max} - \Delta_{\min})/2 > 0$. For any focal disparity z, the maximum target disparity deviation is

$$R(z) = \max_{x \in M} |\delta(x) - z| = \max\{|\Delta_{\min} - z|, |\Delta_{\max} - z|\} = h + |z - m|. \tag{13}$$

The largest strength satisfying the target-sharpness constraint in Equation 4 is therefore

$$K_{\mathrm{subj}}(z) = \frac{p_0 c_{\mathrm{sharp}}}{h + |z - m|} \le \frac{p_0 c_{\mathrm{sharp}}}{h} = K^*_{\mathrm{subj}}, \tag{14}$$

with equality only at $z = m$, where the two endpoint deviations are equal. This yields the joint closed-form solution in Equations 5–7. A mean- or median-based focal disparity that differs from m must use a smaller blur strength to preserve the same target extent; using K^*_{subj} instead would violate the sharpness bound at an endpoint. The minimax plane is thus the exact solution to the subject-preservation objective under the adopted defocus model. We hold this plane fixed when imposing the remaining constraints.

A.3 Feasible aperture interval

At δ_f^*, let $\delta_{\text{subj}} = \max_{x \in M} |\delta(x) - \delta_f^*| = (\Delta_{\max} - \Delta_{\min})/2$. The aperture constraints are

$$\frac{f_0^2 \delta_{\text{subj}}}{N p_0} \le c_{\text{sharp}}, \qquad c_{\min} \le \frac{f_0^2 \delta_{\text{off}}}{N p_0} \le c_{\max}, \tag{15}$$

$$N_{\min}^{\text{lens}} \le N, \qquad N \le N_{\max}^{\text{lens}}. \tag{16}$$

Rearranging these inequalities gives the lower and upper bounds in Equations 8 and 9. Values below N_{lo} violate at least one of the target-sharpness, blur-cap, or lens constraints; values above N_{hi} violate the minimum off-target-defocus requirement or the lens range. The interval is nonempty if and only if $N_{\text{lo}} \le N_{\text{hi}}$. Its strongest feasible setting, N_{lo}, equals N_{subj} when the subject-sharpness bound is active; otherwise, a lens or excessive-blur constraint determines it.

A.4 Implementation and aperture sampling

By default, we request five candidates and set $c_{\min} = 0.002 \max(H, W)$, $c_{\max} = \min\{0.05 \max(H, W), 30\}$, $c_{\text{sharp}} = 1$, and $N_{\max}^{\text{lens}} = 22$. For each source photograph, we recover the focal length, maximum aperture, pixel pitch, and field of view required by the depth estimator and renderer. Samples with missing optical parameters are discarded rather than assigned a generic camera configuration. The lower f-number bound comes from the lens maximum-aperture metadata; unless a bound is specified explicitly, samples without this value are discarded.

Before computing the target disparity range, we morphologically erode the original segmentation mask M_0 to obtain the valid target interior $M \subseteq M_0$. This reduces the influence of unreliable depth estimates near object boundaries. Target-sharpness constraints are imposed on M, while off-target disparity is computed outside M_0. We reject invalid masks, targets with negligible maximum off-target disparity deviation δ_{off}, and empty feasible intervals.

We sample five candidate f-numbers uniformly in $\log N$ from weak to strong defocus:

$$\log \widetilde{N}_k = \log N_{\text{hi}} + \frac{k}{4} (\log N_{\text{lo}} - \log N_{\text{hi}}), \quad k = 0, \dots, 4. \tag{17}$$

Uniform spacing in $\log N$ follows the multiplicative spacing of photographic aperture stops. Candidate k receives the fixed level label $\ell = k + 1$. We snap each $\widetilde{N}_k$ to the nearest standard f-number within the feasible interval. If multiple candidates map to the same snapped f-number, the first is retained and later duplicates are removed; the surviving candidates keep their original labels. A target may therefore retain a sparse subset such as Levels 1, 3, and 5. We store the physical f-number in the metadata, but condition the model only on its level on the fixed five-level scale.

B Dataset Construction and Statistics

Source collection. The initial pool contains five million candidate photographs drawn from a non-public image collection obtained through commercial procurement. Defocus screening and camera-parameter filtering retain 95,307 AIF source photographs.

AIF screening. Large image collections lack reliable all-in-focus labels, and pre-existing defocus makes images unsuitable as sources for our rendering pipeline. We manually curate a seed set of 5,000 AIF photographs and use BokehMe (Peng et al., 2022a) to generate defocused variants with dense blur supervision. We train a U-Net-based dense defocus estimator on these data to predict a pixel-wise blur map from an image and its estimated disparity. For large-scale screening, we exclude low-texture pixels, divide the predicted map into a 16×16 grid, and retain an image when the maximum patch-averaged defocus does not exceed 1 pixel. This conservatively selects images with low predicted defocus across textured regions. Camera-parameter requirements and target filtering are detailed in Appendix A.

Annotation and rendering models. The data pipeline uses MoGe-2 (Wang et al., 2026) with the recovered field of view for metric-scale depth estimation, Qwen3-VL-235B-A22B-Instruct to propose focal objects and bounding boxes, Segment Anything with a ViT-H backbone (Kirillov et al., 2023) to

convert boxes into masks, BokehMe (Peng et al., 2022a) for depth-guided rendering, and Gemini 2.5 Flash to rewrite accepted proposals as focus instructions. Qwen3-VL uses a Qwen3 language-model backbone (Qwen Team, 2025).

Focus instructions. Each retained target is represented by one self-contained instruction, for example:

> *The camera is focused on the chrome front wheel of the pink vehicle. The photo is captured at aperture level 2 of 5.*

Each instruction identifies the target through visible attributes and spatial relations. We discard targets that cannot be grounded unambiguously rather than reverting to the original short label. The level denominator is always five during both training and inference, even when snapping and duplicate removal leave fewer distinct rendered states for a target.

Figure 7 summarizes the focal-object vocabulary extracted from these instructions.



Figure 7: **Focal-target vocabulary in the training set.** Frequencies are computed over 306,011 unique focal targets, counting different aperture levels and editing pairs for the same target only once. Word sizes use square-root-scaled frequencies to keep less frequent target types visible.

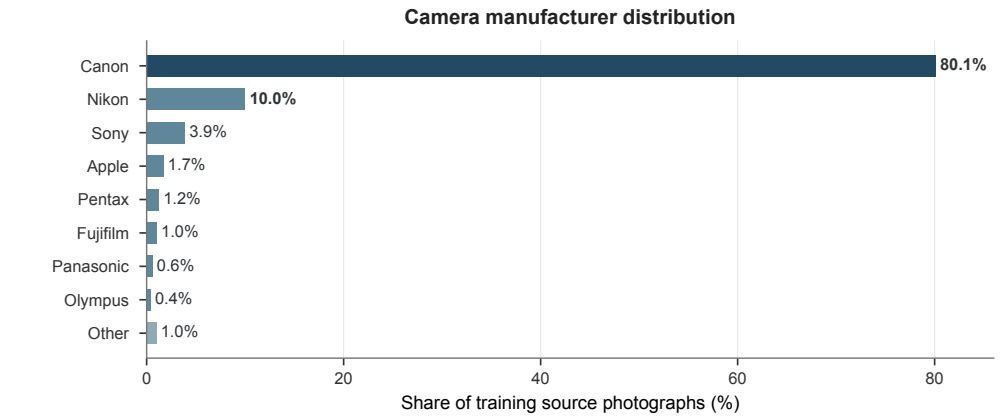

Figure 8: **Camera manufacturers in the training set.** Percentages are computed over the 95,307 unique source photographs in the final training set. We show the eight most frequent manufacturers and combine the remainder as Other. Manufacturer names are inferred from the recorded camera-model metadata.

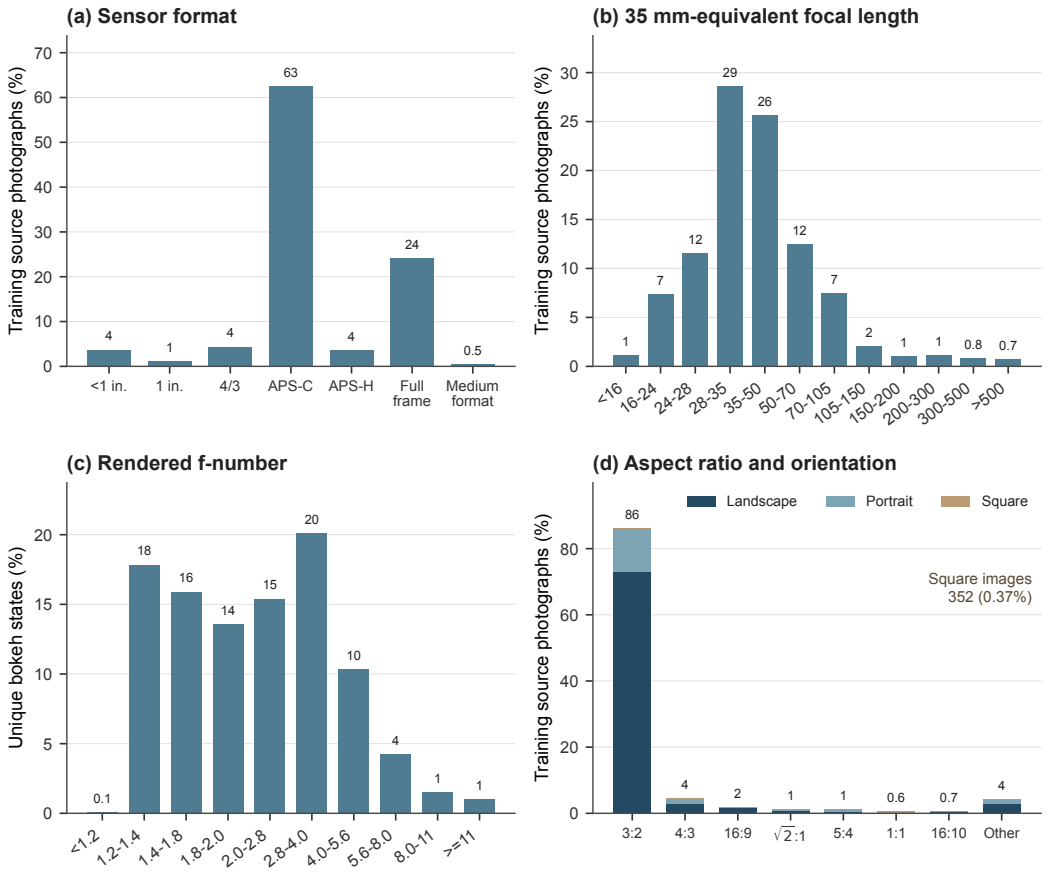

Figure 9: **Optical and image distributions of the training set.** Sensor format, 35 mm-equivalent focal length, and aspect ratio are computed over 95,307 unique source photographs. Rendered f-number is computed over 829,351 unique Bokeh states rather than their repeated occurrences in editing pairs. The aspect-ratio panel also separates landscape, portrait, and square orientations.

These statistics count each of the 95,307 source photographs in the final training set once. Figure 8 shows that the recorded camera metadata are concentrated in Canon devices, followed by Nikon and Sony. Figure 9 reports the corresponding optical and image properties. Most sources use APS-C or full-frame sensors and 24–70 mm equivalent focal lengths. The rendered states span the retained f-number range, while the image geometry is dominated by the 3:2 aspect ratio.

Standard-aperture snapping and duplicate removal can leave fewer than the five requested levels for a target. Table 4 reports the resulting distribution. The 306,011 retained targets contain 829,351 rendered aperture states in total.

Table 4: Distribution of the number of aperture levels retained per focal target. Percentages are rounded to the nearest integer.

Retained levels	Focal targets	Share
1	80,376	26%
2	75,146	25%
3	39,681	13%
4	74,400	24%
5	36,408	12%
Total	306,011	100%

Table 5: Final aperture-sweep dataset statistics.

Quantity	Count / rate
Candidate photographs	5,000,000
Retained AIF source photographs	95,307
Retained focal targets	306,011
Rendered aperture states	829,351
AIF $\rightarrow$ Bokeh pairs	829,351
Cross-target Bokeh $\rightarrow$ Bokeh pairs	1,000,002
Total training pairs	1,829,353

C Training Configuration and Transition Sampling

The training data contain 1,829,353 pairs: 829,351 AIF-to-Bokeh pairs and 1,000,002 cross-target Bokeh-to-Bokeh pairs. The prompt specifies only the output state. All training images are downsampled to 1K resolution while preserving their aspect ratios.

Model adaptation. The Qwen-Image-Edit-2511 base model remains frozen. Rank-64 LoRA adapters with $\alpha = 64$ are applied to the attention, feed-forward, and modulation projections of the diffusion transformer.

Loss weighting. We use the flow-matching SFT loss in DiffSynth-Studio. Training samples a timestep index uniformly from the scheduler's 1,000 training timesteps and multiplies the velocity-prediction mean squared error by the corresponding scheduler weight. For noise levels $\{\sigma_i\}_{i=1}^{1000}$, this weight is

$$a_i = \exp\left[-2(\sigma_i - 1/2)^2\right], \qquad w(\sigma_i) = \frac{1000(a_i - \min_j a_j)}{\sum_j (a_j - \min_k a_k)}. \tag{18}$$

Inference configuration. We use the default inference configuration of Qwen-Image-Edit-2511. The editor receives only the source image and a prompt specifying the output target and level; depth, masks, and camera metadata are used for constructing supervision, not as inference inputs.

Table 6: Configuration of the final fine-tuning run.

Field	Final value
Pretrained editing backbone	Qwen-Image-Edit-2511
Training objective	Flow-matching SFT
Trainable modules	DiT LoRA adapters
LoRA rank and α	64 and 64
Input/output resolution	1K, aspect-ratio preserving
Optimizer	AdamW, lr 2×10^{-4}, wd 0.01
Global batch size	544
Precision and parallelism	bf16, DeepSpeed ZeRO-2
Compute	68 nodes, 544 NVIDIA H20 GPUs
Selected checkpoint	step-10000
Training to selected checkpoint	10,000 optimization steps
Wall time to selected checkpoint	24.31 hours
Compute to selected checkpoint	$\sim$13,225 H20 GPU-hours

D Evaluation Details

The Open Images evaluation sets are constructed separately from the training corpus. We evaluate rendered references and real-defocus references separately: the former measure agreement with the synthesis pipeline, whereas the latter retain the appearance of real lens blur. The AIF-to-Bokeh and cross-target sets contain 457 and 370 scene-target cases, respectively. All methods are evaluated at 1024-pixel resolution.

Baseline scope and reproducibility. We compare with general-purpose image editors that accept a source image and a focal-target instruction without externally supplied depth, masks, or camera parameters. At the time of evaluation, BokehFlow (Huang et al., 2026) and CamEdit (Qin et al., 2026) did not provide publicly available inference code and pretrained weights for reproducing their reported methods.[2] GenFocus (Mu et al., 2025) requires explicit focus and blur-strength settings for its defocus-map condition; evaluating it from our semantic instructions would additionally require a text-to-geometry and strength-calibration procedure. The baseline comparisons therefore assess reference fidelity under direct image-and-language editing.

Evaluation protocols. For AIF-to-Bokeh and cross-target refocusing, TextBokeh is compared with the reference at the requested level. Exact-level Avg. pools all valid requested-level pairs. Best Level selects the level-specific TextBokeh prediction with the lowest LPIPS to its corresponding reference for each scene-target case. A general-purpose baseline produces one output without a level condition; Best-GT matches it to the available reference with the lowest LPIPS. Both oracle protocols use LPIPS as the single selection criterion, and all other metrics are evaluated on the same selected pair rather than optimized independently. We choose LPIPS because the candidate states differ primarily in perceptual blur strength and local texture attenuation. Best Level and Best-GT diagnose candidate-set fidelity and do not represent deployable level selection or level-control performance.

Oracle selection statistics. Best Level selects L1–L5 for 180/23/64/61/129 cases in AIF-to-Bokeh and 128/18/37/69/118 cases in refocusing. The matched-level counts for Qwen/FLUX/HY/Gemini under Best-GT are L1–L5: 83/3/14/46/311, 160/23/76/76/122, 128/24/89/99/117, and 147/13/30/45/222 for AIF-to-Bokeh; the corresponding counts are 139/2/12/29/188, 204/11/42/35/78, 172/17/62/61/58, and 184/8/28/34/116 for refocusing.

Per-level reference sets. Standard-aperture snapping and duplicate removal leave different subsets of the fixed five levels across cases, resulting in 1,822 and 1,331 level-specific reference pairs for AIF-to-Bokeh and cross-target refocusing, respectively. Surviving states retain their original candidate

[2]The public CamEdit repository provides project-page materials: `https://github.com/XinranQin/camedit.github.io`.

labels, so the number of cases need not decrease monotonically with level. Tables 7 and 8 report the per-level counts and results; N denotes the number of cases with a reference at that level.

Table 7: Per-level AIF-to-Bokeh reference fidelity. N is the number of valid cases at each level.

Level	N	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
L1	457	31.63	0.914	0.068	0.029
L2	150	31.51	0.912	0.075	0.034
L3	400	31.29	0.912	0.075	0.034
L4	415	31.46	0.916	0.070	0.033
L5	400	31.54	0.918	0.069	0.034

Table 8: Per-level cross-target refocusing fidelity. N is the number of valid cases at each level.

Level	N	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
L1	370	30.51	0.897	0.080	0.035
L2	96	30.74	0.901	0.078	0.038
L3	279	30.76	0.904	0.078	0.037
L4	294	30.89	0.907	0.074	0.037
L5	292	31.13	0.911	0.073	0.037

Real-defocus reference evaluation. We construct 96 paired examples from photographs with real defocus. GenFocus (Mu et al., 2025) provides the corresponding AIF input proxies, while the original photographs are retained as real-defocus references. This evaluation measures agreement with real photographic blur rather than BokehMe-rendered targets. Each of the five TEXTBOKEH predictions is compared with the same reference, and Best-Pred selects the prediction with the lowest LPIPS for each case. It selects L1–L5 for 7/12/14/6/57 cases, respectively; all metrics are computed on the same selected prediction-reference pair. The general-purpose editors produce a single output per case. Best-Pred uses the reference after inference and is an oracle diagnostic of candidate-set fidelity, not a deployed single-output comparison or a test of prescribed-level accuracy on real photographs. The AIF proxy inputs also distinguish this experiment from cross-target refocusing of arbitrary real-defocused inputs. Figure 10 extends the representative cases in Figure 6 with additional subjects and scene layouts under the same paired-reference protocol.



Figure 10: **Additional qualitative results with real-defocus references.** Each group shows an AIF input proxy, the corresponding real-defocus photograph used as GT, and the output of TEXTBOKEH for the displayed focal-target instruction. The examples extend Figure 6 with additional people, animals, clothing, and scene layouts.

E PROMPTS USED IN QUALITATIVE COMPARISONS

For reproducibility, we specify the prompts used in the qualitative comparisons. Let `FOCUS_INSTRUCTION` denote the complete focal-target sentence displayed in each figure, including its terminal period. For TEXTBOKEH, the prompt at level $\ell \in \{1, \ldots, 5\}$ is constructed as

> `FOCUS_INSTRUCTION` + *"The photo is captured at aperture level ℓ of 5."*

Qwen, FLUX, Gemini, and GPT use the same task-enhanced prompt. Substituting the displayed sentence for `FOCUS_INSTRUCTION` gives the exact external prompt used for each example. For AIF-to-Bokeh synthesis, the prompt is

> *"Turn this all-in-focus photograph into a realistic shallow-depth-of-field image."* + `FOCUS_INSTRUCTION` + *"Keep that subject optically sharp and render smooth large-aperture bokeh in the rest of the scene. Do not change camera pose, framing, lighting, or scene content."*

For cross-target refocusing and real-defocus evaluation, the prompt is

> *"Refocus this photograph by changing the optical focus plane."* + `FOCUS_INSTRUCTION` + *"Keep that subject optically sharp with natural depth-of-field blur elsewhere. Do not change camera pose, framing, lighting, or scene content."*

Hunyuan receives the same external task-enhanced prompt. HunyuanImage-3.0-Instruct-Distil natively supports reasoning-based Prompt Self-Rewrite, described by its official inference interface as think–rewrite–image generation.[3] We retain this native inference behavior and do not manually edit, select, or replace the rewritten prompt. The exact `FOCUS_INSTRUCTION` is printed with each baseline comparison, so the complete external prompt used for each comparison is obtained directly from the corresponding template above.

F ADDITIONAL QUALITATIVE COMPARISONS

Details of the main AIF-to-Bokeh comparisons. In the tower example of Figure 4, Qwen changes the tower's color and rooftop, and FLUX modifies building details. HY and GPT leave locally sharp patches near the foreground silhouette despite blurring nearby regions. In the badge example, Qwen enlarges the badge and changes the framing, while GPT preserves the badge but blurs the nearby headlight. Gemini retains much of the original scene sharpness in the tower example, whereas in the badge example it softens both the badge and nearby car details. These cases illustrate distinct errors in scene preservation and the spatial distribution of defocus.

Additional AIF-to-Bokeh examples. Figures 11 and 12 extend the AIF-to-Bokeh comparisons in Figure 4 with four additional scenes. In these examples, TEXTBOKEH produces progressively stronger off-target defocus while keeping the instructed subject sharp and preserving the scene. The detail crops further show coherent blur across regions at similar depths and natural transitions around object boundaries, consistent with the rendered references.

The baselines vary in their ability to preserve content and follow the focal-target instruction. FLUX introduces substantial structural changes in these examples. It can nevertheless produce selective blur, as the vase example shows, although the composition changes substantially. Gemini leaves the white-car example largely unchanged, while softening the requested person together with nearby content in the library example. Qwen selects the requested targets in the library and blue-car examples, but leaves the white-car example largely unchanged and alters the framing and apparent size of the target in the vase example.

[3] https://github.com/Tencent-Hunyuan/HunyuanImage-3.0

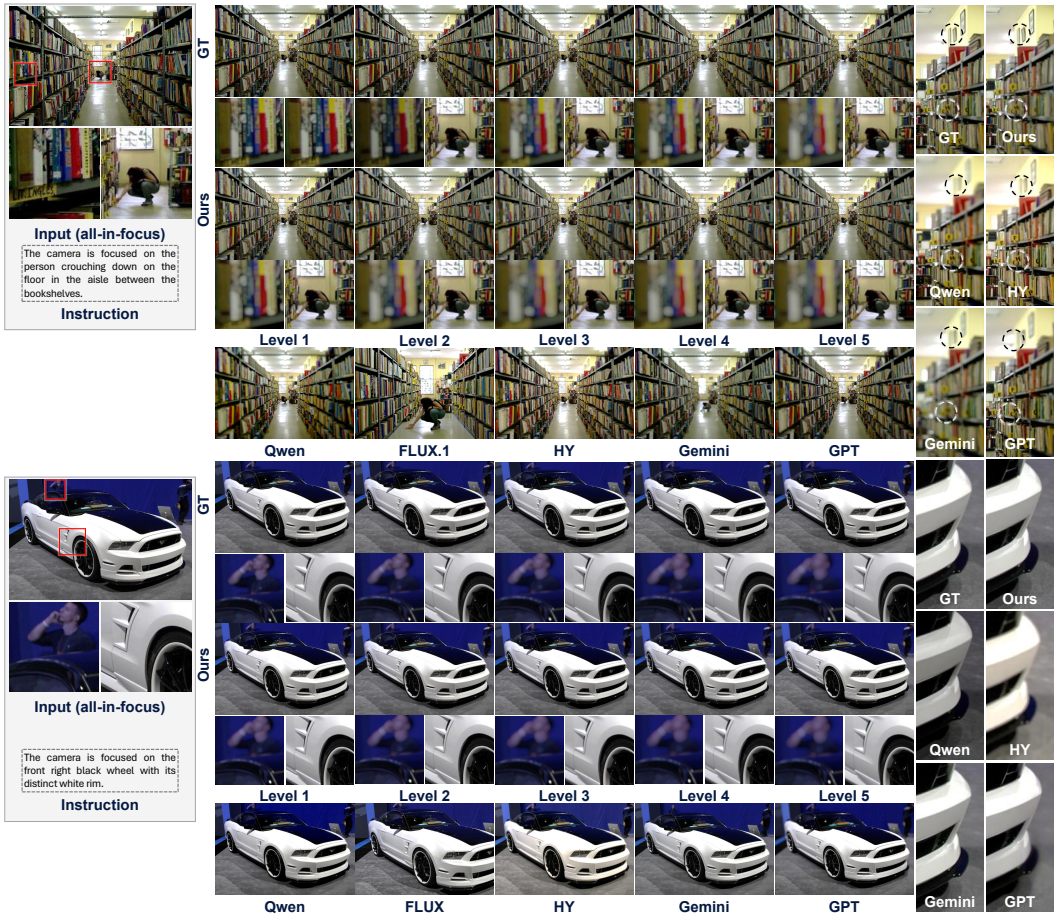

Figure 11: **Additional AIF-to-Bokeh comparisons.** Results at the displayed aperture levels, baseline outputs, and detail crops.

Correct target selection does not always yield depth-consistent defocus. In the library example, the circled regions lie at approximately the same depth, yet Qwen and GPT render them with noticeably different blur strengths. Similarly, in the blue-car example, Qwen, HY, and GPT produce markedly different defocus in the left and right crops despite their similar depth ranges. GPT also introduces boundary artifacts along the white car's contour, illustrating the difficulty of handling depth discontinuities in bokeh rendering (Peng et al., 2022a). In these regions, TEXTBOKEH more closely reproduces the spatial blur distribution and boundary transitions of the references.

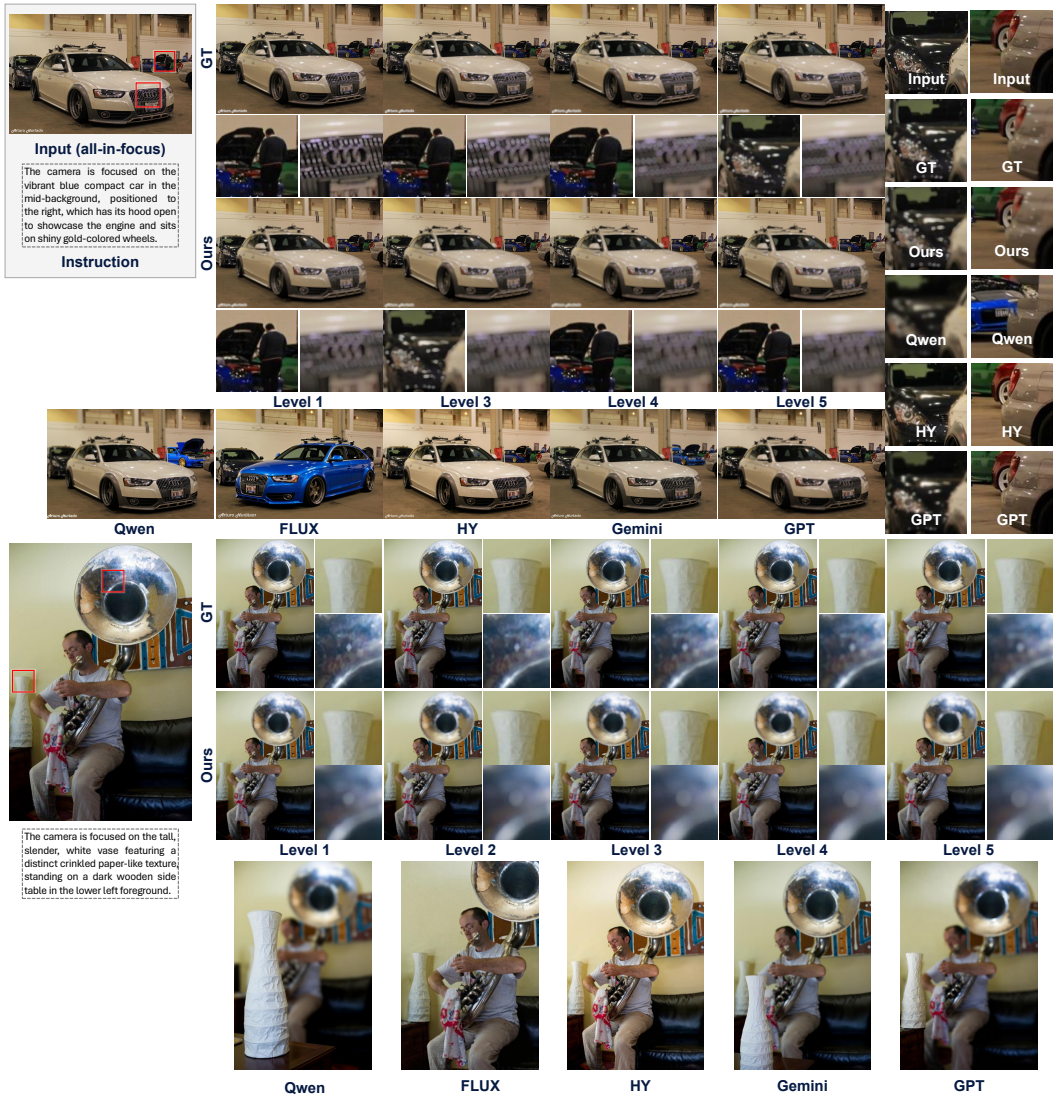

Figure 12: **Additional AIF-to-Bokeh comparisons (continued).** Results at the displayed aperture levels, baseline outputs, and detail crops.

Figure 13 extends the refocusing comparisons with three additional scenes. In these examples, TEXTBOKEH follows the requested focus shift and produces progressive changes in off-target defocus while preserving scene content. The examples distinguish successful target selection from faithful content preservation and depth-consistent blur.

In the first example, focus is shifted from the background to the foreground license plate. All methods preserve the overall composition, but Gemini blurs the scene broadly and FLUX produces a largely all-in-focus result. Qwen, HY, and GPT bring the plate into focus but alter its lettering or appearance. HY also leaves other parts of the car's front blurred despite their proximity in depth to the plate, producing sharpness confined to the selected object.

In the second example, the requested target is the nurse attending to a donor in the background. Qwen, FLUX, Gemini, and GPT change the nurse's appearance or position, with the changes less pronounced in GPT. HY preserves the scene content but fails to transfer focus from the foreground donor to the nurse.

In the third example, GPT successfully refocuses on the sports car's front license plate, but noticeably blurs the white road marking at the lower right despite its similar depth to the target. Gemini again produces broad blur across the image, while FLUX and HY largely retain the original focus state. Qwen shifts focus to the requested plate but introduces substantially stronger off-target blur than the displayed references.

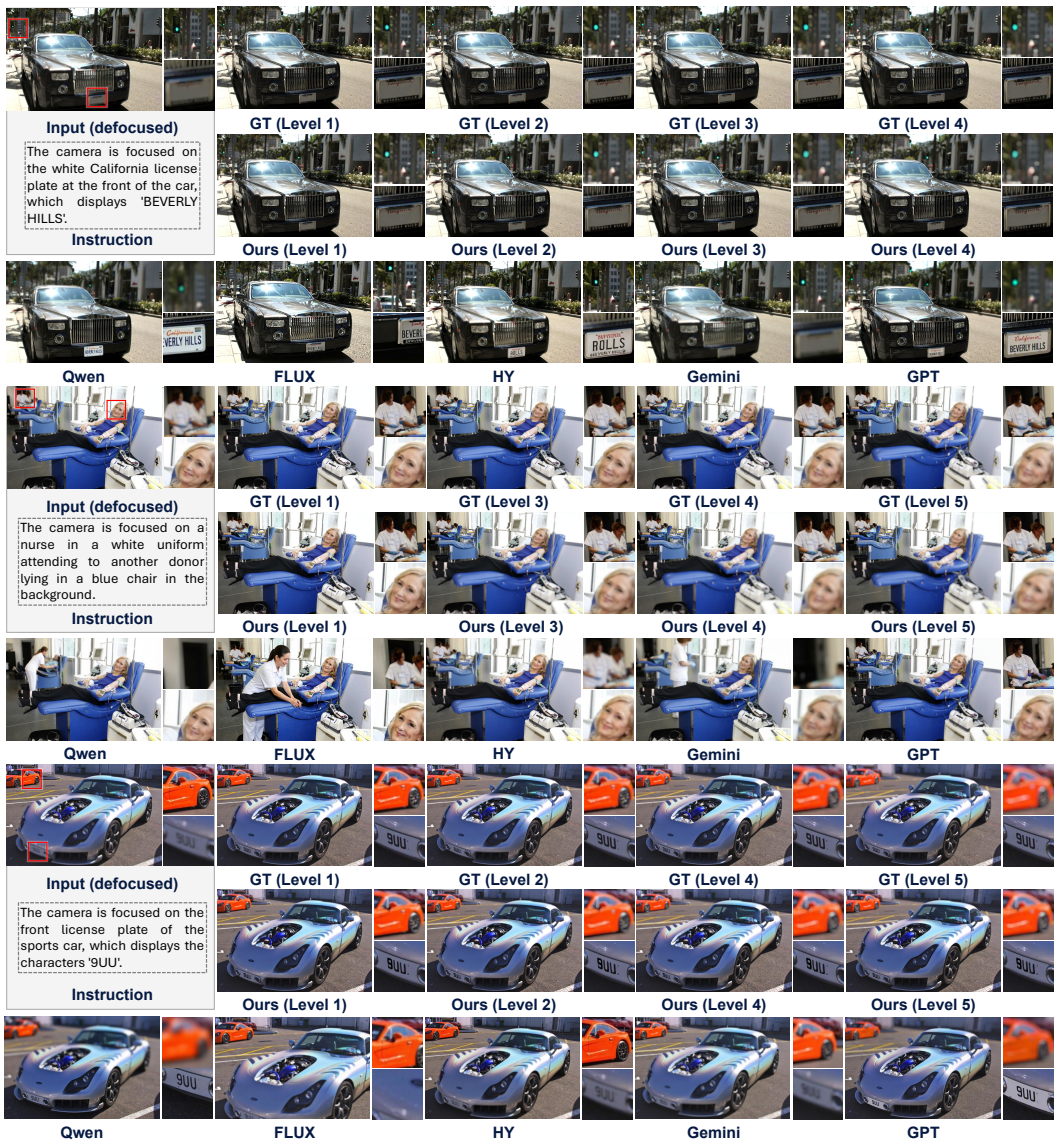

Figure 13: **Additional cross-target refocusing comparisons.** Results at the displayed aperture levels and baseline outputs.

Refocusing from real camera-captured inputs. Figure 14 presents ten qualitative examples of language-guided refocusing on real camera-captured photographs. The target descriptions specify objects through appearance, spatial relationships, and object-part attributes. All outputs in this figure use the requested relative defocus level $\ell = 5$. These examples complement the synthetic-reference evaluations by illustrating focus changes on real photographic inputs.

Several examples redirect focus from a prominent foreground subject to a specified background target. Selecting the flowering cactus sharpens its ribs and spines while defocusing the foreground fire hydrant. Selecting the grain silos makes their surface structure clearer while softening the overturned vehicle. In the swimming example, the blue cap and spatial description distinguish the distant swimmer from the woman in the foreground; the output brings the requested swimmer into clearer focus while blurring the woman. The distant person, exposed tree roots, and canopy tent provide further examples of background selection across different scenes.

Other examples illustrate finer target specification. The Labrador prompt selects the nose, whose surface and nostrils become more clearly defined, while the tennis prompt selects the small court-number placard rather than the visually prominent player. In the chicken example, the description identifies the nearest individual by its position, red face, and orientation; its head becomes clearer while the other birds are softened. The basset-hound example additionally shows facial targeting with a more modest change in sharpness. Across these scenes, the outputs largely retain the original framing and object arrangement while shifting visual emphasis toward the described targets.

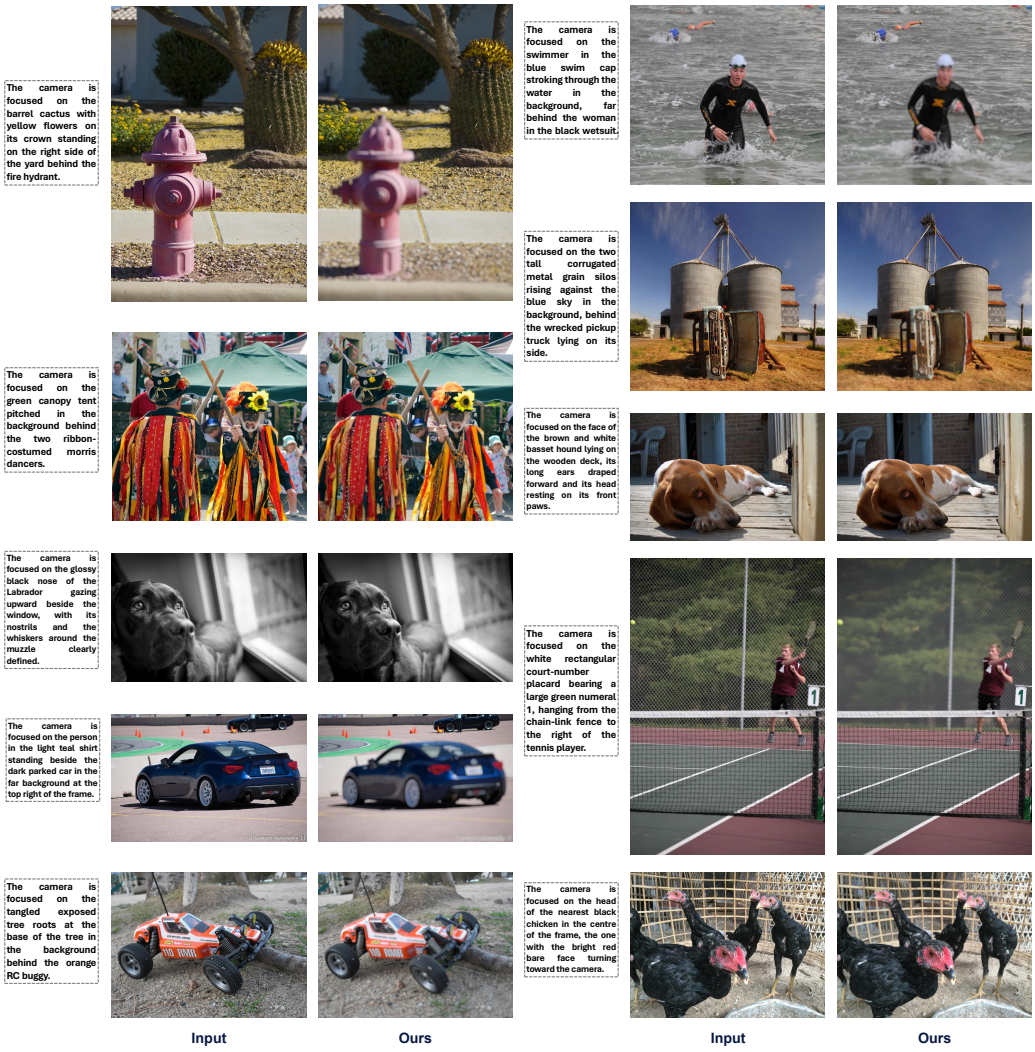

Figure 14: **Language-guided refocusing on real camera-captured photographs.** Each example pairs a real photographic input with the output of TEXTBOKEH under the displayed focal-target instruction. The ten examples cover people, animals, vegetation, structures, and small object details, illustrating changes in target sharpness and surrounding defocus while largely retaining the scene layout.

Different focal targets from the same input. Figure 15 holds the AIF input fixed within each scene and varies the focal-target description. Each target is shown at several relative levels, allowing comparison both across targets at a shared level and across strengths for a fixed target. At Level 3, for example, both scenes show distinct focus choices from the same source photograph.

In the wedding scene, the first instruction identifies the open pizza box on the foreground table, while the second describes the couple cutting the wedding cake through their clothing and shared action. Focusing on the pizza box retains foreground detail and softens the people farther back. Selecting the couple instead brings them into focus while allowing the foreground tableware and flowers to become softer at stronger levels. The two output sequences retain the scene arrangement while changing which subject receives visual emphasis.

In the vehicle scene, the descriptions distinguish the black license plate on the white minivan from the tall light pole with a red banner in the upper-left background. With the plate selected, the front of the minivan remains sharp. With the pole selected, the minivan becomes defocused, with stronger foreground blur at higher displayed levels. Together, these examples illustrate object selection through appearance, location, and activity descriptions, alongside relative strength control for each selected target.



Figure 15: **Different focal targets from the same AIF input.** Each scene shares one source image across two target descriptions. The wedding example selects the foreground pizza box or the couple cutting the cake; the vehicle example selects the minivan's license plate or the background light pole. Outputs show the resulting focus choices and defocus strengths at the displayed relative levels.

Generalization to same-target strength adjustment. Our training pairs do not include Bokeh-to-Bokeh edits between different aperture levels of the same focal target. Figure 16 shows that TEXTBOKEH can nevertheless respond to requests for stronger or weaker defocus while retaining the focal target. These examples provide qualitative evidence of generalization to same-target strength adjustment without training on this type of editing pair.

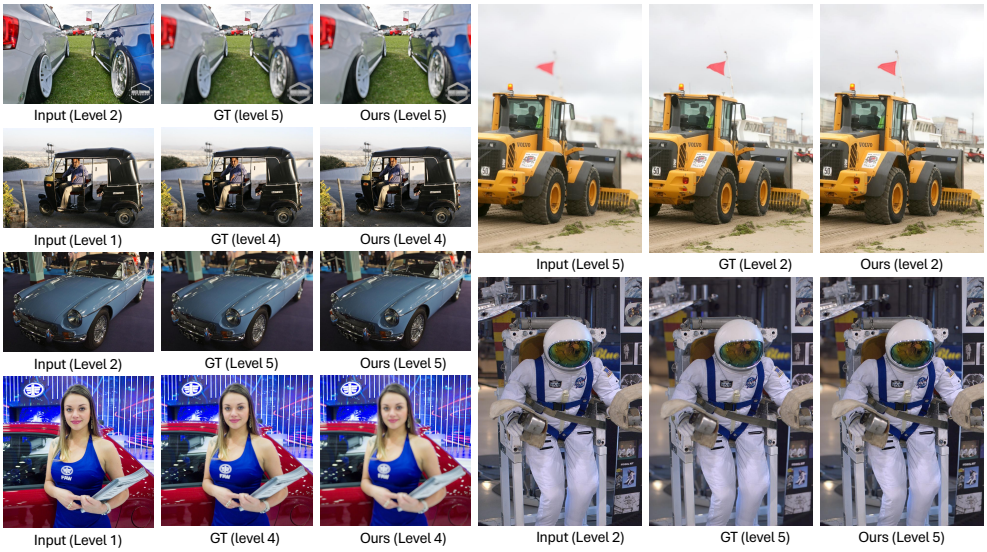

Figure 16: **Generalization to same-target strength adjustment.** Each triplet shows the input at its source level, the reference at the requested level, and the corresponding output of TEXTBOKEH. The examples illustrate stronger and weaker defocus with the focal target retained, although this type of Bokeh-to-Bokeh editing pair is absent from training.

G LIMITATIONS

The editor can misidentify a requested target, introduce boundary artifacts, or alter local details. During supervision construction, errors in estimated depth or target masks can shift the focal plane and feasible aperture interval; these geometric estimates are not inference inputs. Strong source defocus may remove detail that the model cannot recover reliably. The defocus levels are relative within each target stack and are not calibrated across images. Although the candidate outputs approximate real photographic blur, differences remain in defocus strength, depth-boundary transitions, and local texture attenuation. Synthetic supervision captures only part of the variation in real lens blur.