

A Implementation Details

In this section, we provide more details about the training process. For the main training phase, we set all loss function weights (λ_1 to λ_4) to 1. The AdamW optimizer was used with a learning rate of 0.0003 and no weight decay. Training was conducted for 400,000 iterations, with LLaVA-7B [27] employed as the Multimodal Large Language Model (\mathcal{F}) in our architecture. The main training utilized 2 Nvidia A100 80GB GPUs for approximately three days, with a batch size of 16.

In the additional training phase, the surrogate module was trained using the AdamW optimizer with a learning rate of 0.0003. This phase involved 1,000 iterations of training.

Following the completion of the surrogate module training, we resumed fine-tuning the MLLM, Token Broadcaster, and Token Decoder components by setting the weight of the MSE loss term (λ_5) to 10. During this phase, we employed the AdamW optimizer with a decreased learning rate of 0.0001 and conducted training for an additional 400,000 iterations.

B Metrics Details

We utilize a diverse set of metrics including L1/L2, CLIP-I, DINO, CLIP-T, CLIP-dir, and PickScore. L1 and L2 are used to calculate the mean absolute pixel-wise discrepancy between the generated and ground truth images. Beyond pixel-level evaluation, we use CLIP-I [35] and DINO [33], both of which assess image quality via cosine similarity between embeddings of the edited and ground truth images. CLIP-T evaluates the alignment between the generated image and the goal image global description. Additionally, we employ CLIP-dir, which measures the agreement between changes in images and changes in captions. Lastly, PickScore [22] serves as a proxy for human preference by quantifying how likely a generated image is to be favored by humans.

C Dataset Details

C.1 Dataset Design

The MagicBrush dataset [47] is specifically designed to support single-instruction image editing tasks, encompassing both single-turn and multi-turn scenarios. Single-turn tasks involve a single instruction for an image edit, while multi-turn tasks consist of sequential single-instruction edits, such as two-turn and three-turn edits. Two-turn edits are composed of two sequential single-instruction edits, and three-turn edits consist of three sequential single-instruction edits. Using the multi-turn data from the MagicBrush dataset, specifically the two-turn and three-turn edits, we restructured them into multi-instruction tasks. By combining the individual instructions into cohesive multi-instruction prompts, we created more complex editing scenarios that simulate real-world use cases. This restructuring involved generating connecting phrases to logically link the instructions, enabling seamless transitions between steps in the multi-instruction tasks. In addition to MagicBrush dataset, we incorporate the EMU dataset [40], which is a single-turn image editing dataset.

For Context-Aware Instruction Image Editing task, we utilized ChatGPT-4V [32] to generate non-applicable instructions. For each input image, we obtained five randomly generated non-applicable instructions from the model. During training, one of these instructions was randomly selected after verifying its non-applicability to the given image using the same GPT model and incorporated into the multi-instruction tasks, enhancing the model’s robustness to distinguish between applicable and non-applicable instructions. This two-stage filtering ensured that the synthetic instructions were both semantically grounded and diverse, avoiding trivial or unrealistic cases. During testing, one of the five non-applicable instructions was fixed and consistently used for evaluation, ensuring a standardized assessment of the model’s performance. The number of test dataset examples for each task is shown in Table 6.

C.2 Dataset Biases

To mitigate potential biases introduced by using ChatGPT-4V [32] for generating non-applicable instructions, we carefully designed the prompting process to align with the linguistic structure of the MagicBrush dataset. Specifically, we instructed the model to generate instructions that begin with

Table 6: **Details on the number of image editing samples for each task in MagicBrush dataset.**

Sessions with	Single-turn Inst.	Multi-turn Inst.	Multi Inst.	Context-Aware Inst.
One Edit	216	216	-	-
Two Edits	240	120	120	1053
Three Edits	597	199	597	1571
Total	1053	535	717	2624

common syntactic patterns found in MagicBrush dataset (e.g., "Put something", "Remove something", "Replace something", "Let there be", "Make something to something").

To validate the representational fidelity of the generated instructions, we categorized all instructions into four major types: *Add object*, *Replace object*, *Remove object*, and *Change something (action, color, etc.)*. The distribution of instruction categories in our generated non-applicable dataset (3,677 instances) was compared with the original MagicBrush’s instructions. As shown in Table 7, although there is a slight variation in category proportions, the overall instruction distribution was designed to reflect the original dataset’s structure and characteristics.

Table 7: **Instruction distribution comparison between MagicBrush and our generated dataset.**

Instruction Category	MagicBrush	Generated Non-applicable Instructions
Add object	39.0%	34.3%
Replace object	17.9%	20.5%
Remove object	7.0%	21.1%
Change something (action, color, etc.)	36.1%	24.1%

D Additional Ablation Study

D.1 Token Decoder Variations

A straightforward approach might be to use pretrained segmentation models [21, 19, 36] to replace the Token Decoder for generating masks. Thus, we replaced our Token Decoder with SAM [21], and evaluated its performance within our framework, as shown in Table 8. However, SAM does not take edit instructions as input, lacking information about which parts of the image actually need editing. In contrast, our 2-layer transformer-based model utilizes edit instructions to produce targeted masks, resulting in superior CLIP and DINO scores, showing we can better preserve the intended meaning of edits.

D.2 Inference Time Comparison

We measure the inference time of each method and report the results in Table 9. All experiments are conducted using an NVIDIA A100 80GB GPU. We separately measure the time for MLLM-related modules and diffusion-based modules. CAMILA achieves the fastest inference time for the MLLM module while maintaining comparable total latency. The slightly higher diffusion time in FoI and CAMILA is attributed to attention modulation applied to the IP2P backbone. Unlike other baselines, CAMILA supports single-shot, multi-instruction editing without requiring preprocessing steps such as keyword extraction.

E Additional Results

E.1 Preference-Based Evaluation Results

We additionally report PickScore [22], a metric trained to reflect human preferences in image generation tasks. As shown in Table 10, our method achieves the highest PickScore under both multi-instruction and context-aware settings, outperforming IP2P, MGIE, SmartEdit, and FoI. These

Table 8: **Comparison of results using different segmentation models.** Our Token Decoder demonstrates superior alignment in CLIP and DINO metrics compared to SAM segmentation models.

Task		Segmentation Model	L1↓	L2↓	CLIP-I↑	DINO↑	CLIP-T↑
Single Inst.	Single -Turn	SAM	0.0585	0.0184	0.9355	0.9056	0.2974
		Ours	0.0602	0.0194	0.9367	0.9067	0.3020
	Multi -Turn	SAM	0.0903	0.0321	0.8932	0.8298	0.2917
		Ours	0.0931	0.0339	0.8969	0.8357	0.3011
Multi Inst.	Multi	SAM	0.0934	0.0359	0.8946	0.8303	0.2895
		Ours	0.0957	0.0372	0.8961	0.8329	0.2975
	Context -Aware	SAM	0.0647	0.0215	0.9273	0.8905	0.2952
		Ours	0.0673	0.0228	0.9284	0.8910	0.3002

Table 9: **Inference time comparison of MLLM part and diffusion model part.**

Method	IP2P	MGIE	SmartEdit	FoI	CAMILA (Ours)
MLLM part	—	2.1s	1.5s	—	0.7s
Diffusion part	7.2s	3.4s	4.0s	9.1s	8.5s
Total Inference Time	7.2s	5.5s	5.5s	9.1s	9.2s

results suggest that our model generates outputs that are more aligned with human preferences compared to the baselines.

E.2 Comparison with Multi-instruction-based Method

As illustrated in Figure 6, we compare our method with FoI [11], one of the state-of-the-art multi-instruction image editing approaches. FoI employs a pretrained GPT model [31, 32] to extract keywords, which are subsequently used as masks based on the cross-attention maps of the U-Net. While the attention maps in (a) and (b) appear to reflect the objects in the image, they lead to suboptimal results due to a lack of context-awareness. In (a), the attention map focuses on the elephant’s face instead of the head, which the instruction specifies, resulting in an inaccurate edit. Similarly, in (b), while the instruction specifies “cupcake with frosting,” the attention map erroneously attends to all three cupcakes in the image instead of isolating the one on the far left, leading to an incorrect edit. In contrast, the decoded [MASK] generated by our method leverages the contextual understanding of the instruction to accurately identify the specific cupcake requiring modification. In case of (c), it highlights the impact of an incorrect attention map, where the edit fails to meet the instruction. In (d) and (e), the extracted keywords, such as “background” or “top”, are vague. Even if keywords like “building” or “cherry,” which represent the objects intended to be added, were selected, they are not present in the given image, meaning the attention maps would still fail to provide meaningful guidance in such cases. In (f), while the attention map partially focuses on the bird, areas with weaker attention intensity remain unedited. By contrast, our proposed method, CAMILA, generates binary masks using the context-aware capabilities of the MLLM to precisely determine which regions need modification. This approach enables more accurate and context-aware image editing compared to existing methods, as demonstrated in the provided examples.

E.3 Failure Cases

We address several cases where our method shows limitations in achieving optimal editing, primarily due to the pretrained diffusion model used in our framework. Figure 7 presents examples of failure cases of CAMILA. While the binary mask identifies the intended editing regions with high accuracy, it occasionally fails to account for their relative sizes. For instance, in cases (a)–(c), although the binary mask captures the correct regions for editing, the resulting additions are smaller than expected relative to the surrounding objects. Similarly, in case (d), the generated binary mask focuses on a less relevant area, despite the presence of more suitable editing regions, such as the grass in the top-left or bottom-right corners. This results in the addition of relatively small objects. Segmentation inaccuracies for small or thin structures often arise from the coarse granularity of ground-truth masks in the MagicBrush dataset and the internal downsampling of Stable Diffusion during attention

Table 10: **PickScore comparison across baseline methods.**

Method	IP2P	MGIE	SmartEdit	FoI	CAMILA (Ours)
Multi-instruction	0.1598	0.1404	0.1844	0.2363	0.2790
Context-Aware	0.1666	0.1350	0.1865	0.2285	0.2834

modulation to a low spatial resolution. These limitations restrict the model’s ability to localize fine details or very small regions. Incorporating more advanced diffusion models or fine-tuning the diffusion backbone to better integrate with mask-based editing, as well as adopting higher-resolution conditioning and fine-grained supervision, could further improve the precision of our method and will be explored in future work.

E.4 Further Qualitative Results

We present additional qualitative results in Figures 8 to 10, highlighting comparisons with other baselines [4, 10, 16, 11] across single-instruction, multi-instruction, and context-aware instruction image editing task, respectively. Additionally, we include both the MLLM token outputs and the corresponding decoded binary mask results for further analysis.

F Broader Impacts

The integration of a mechanism to filter non-applicable instructions in CAMILA has important social implications. By preventing the execution of irrelevant or misleading inputs, the system reduces user frustration and promotes a more intuitive editing experience. This contributes to broader accessibility, enabling individuals with limited technical expertise or physical impairments to achieve their editing goals through natural language alone. Moreover, the ability to reject non-executable instructions helps safeguard against misuse, fostering more responsible and trustworthy deployment of generative image editing technologies.

G License of Assets

In our study, we utilize several instruction-based image editing models [4, 10, 16, 11] as baselines. Specifically, InstructPix2Pix [4] is released under the Creative ML OpenRAIL-M license, as it is built upon Stable Diffusion [38]. MGIE [10] is distributed under the Apple Custom License. SmartEdit [16] is available under the MIT License. The licensing information for FoI [11] was not specified in the available resources.

For training and evaluation, we employ the MagicBrush [47] dataset, which is released under the Creative Commons Attribution 4.0 License. This license permits users to share and adapt the dataset for any purpose, including commercial use, provided appropriate credit is given and any changes made are indicated. Building on this dataset, we construct a new context-aware image editing dataset that enables the evaluation of models under various scenarios.

H Limitations

In our work, we employ the LLaVA-7B backbone [27] for its MLLM component and the Stable Diffusion v1.5 backbone [38] to maintain consistency and fair comparison with prior instruction-guided image editing methods that adopt the same setting. We have not yet explored larger MLLM variants or alternative diffusion architectures, which may further enhance image editing quality. Investigating the impact of scaling the MLLM and replacing the diffusion backbone with more recent architectures will be an important direction for future work.

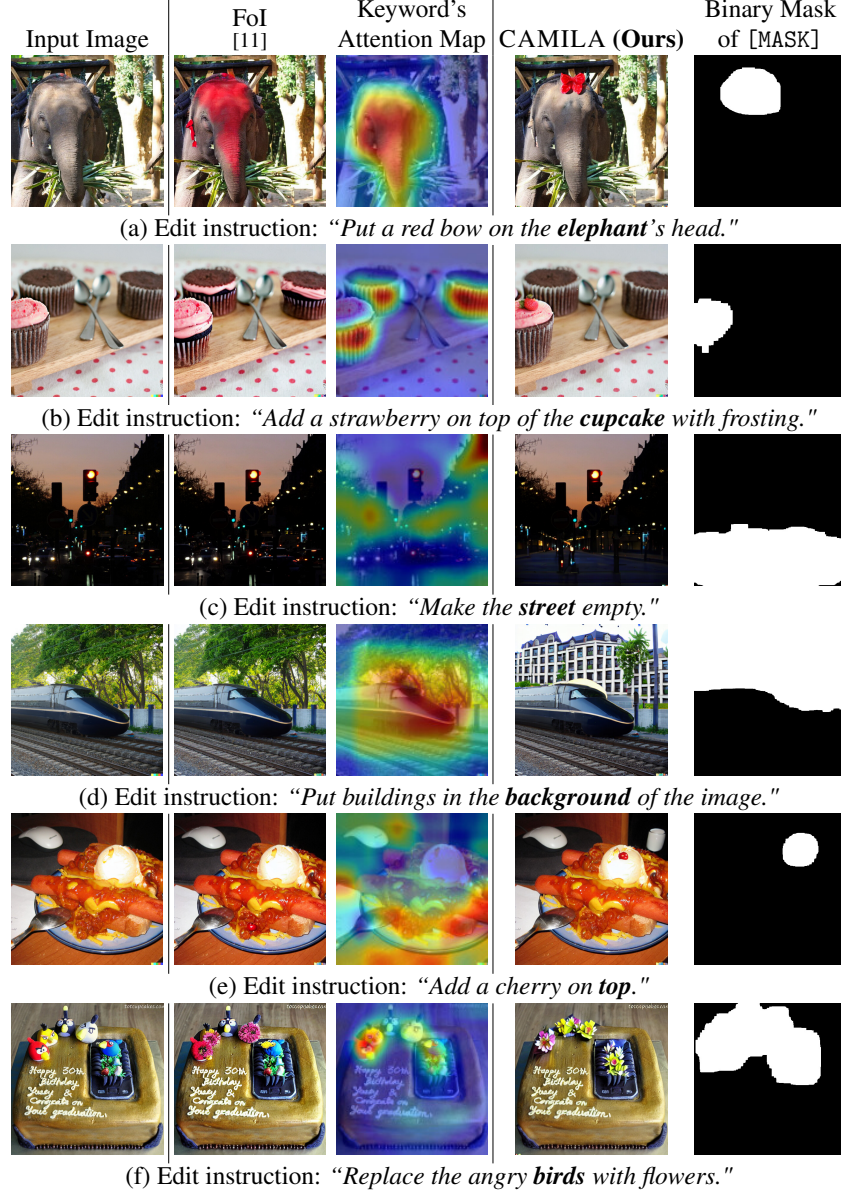


Figure 6: **Qualitative comparisons between CAMILA and FoI.** FoI relies on attention maps derived from extracted keywords (highlighted in **bold**), which often lack context-awareness, leading to inaccurate edits. The red regions in the attention maps represent areas with higher attention intensity, indicating where the model focuses during image editing. In contrast, our proposed method, CAMILA, utilizes context-aware [MASK] decoding to generate precise binary masks, enabling more accurate and instruction-compliant image modifications.

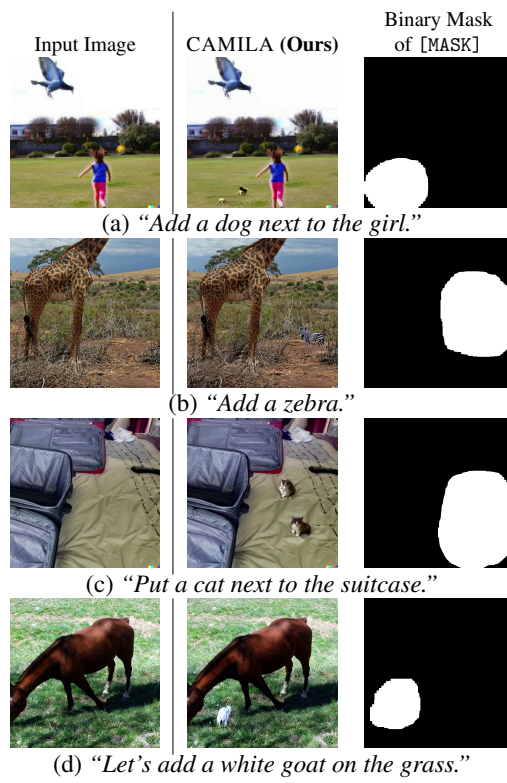


Figure 7: **Failure cases of our method CAMILA.**

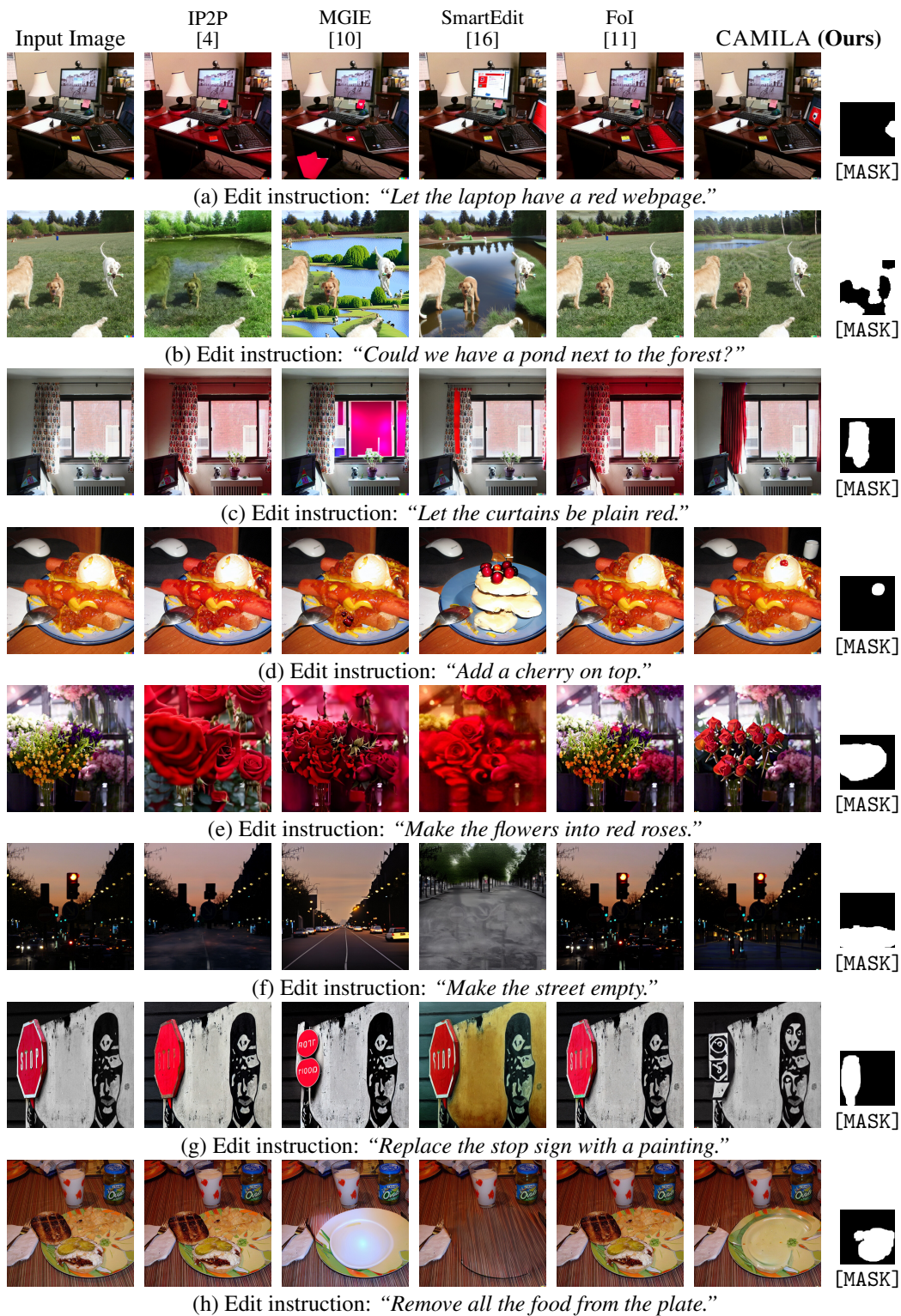


Figure 8: **Qualitative comparisons for single instruction task**

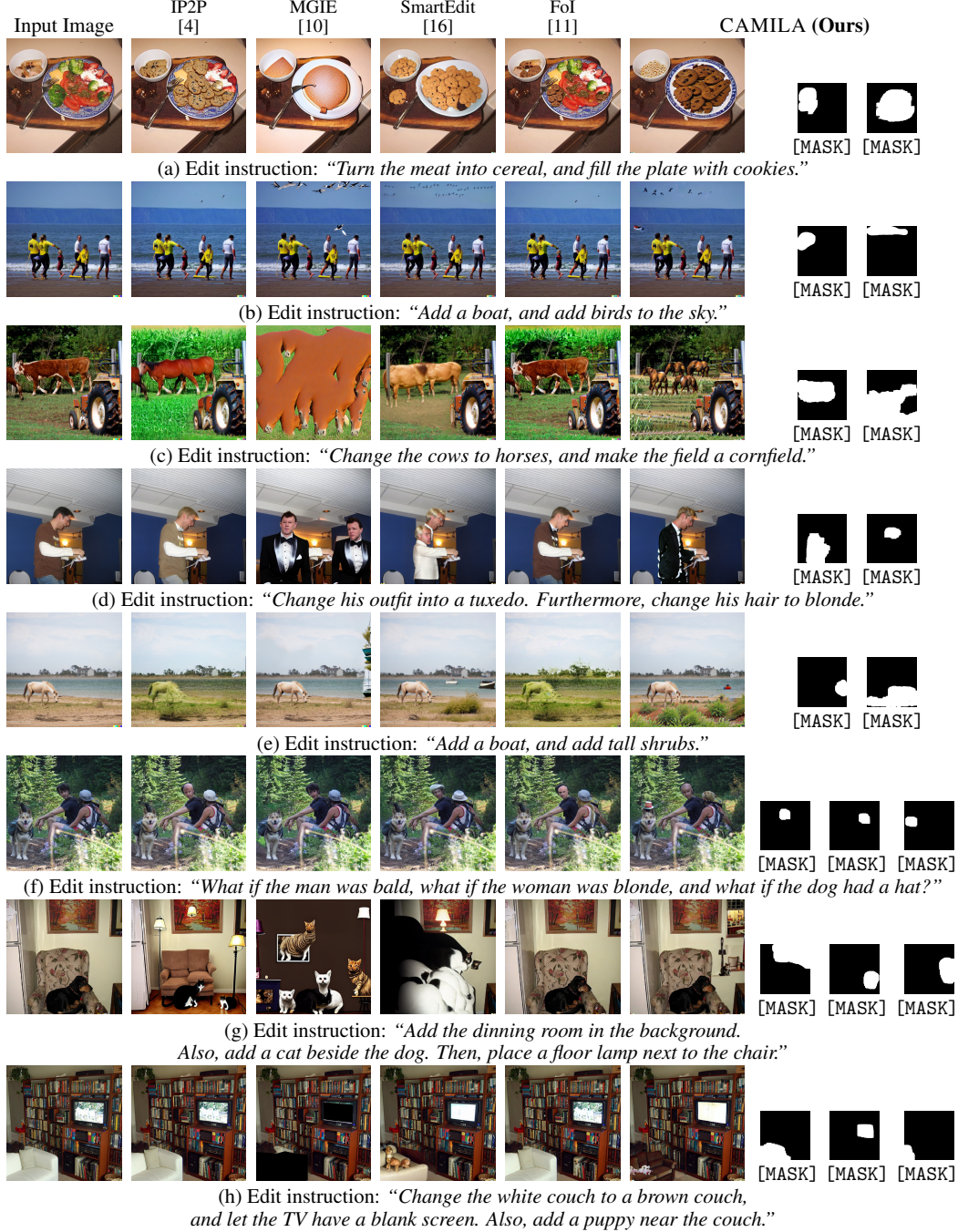


Figure 9: **Qualitative comparisons for multi-instruction task**

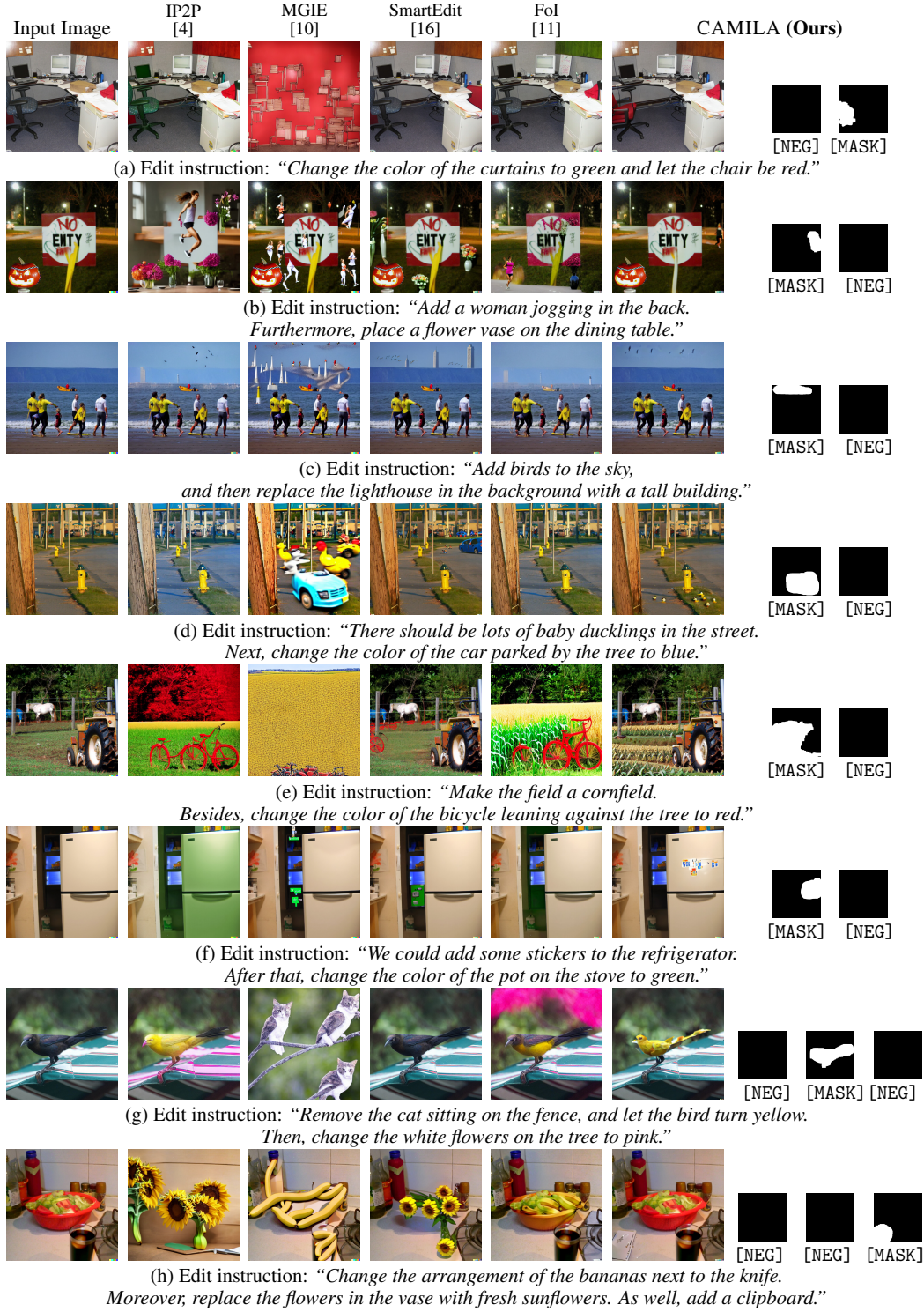


Figure 10: **Qualitative comparisons for context-aware instruction task**

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: Yes. The abstract outlines the high-level scope. Also, the introduction section clearly states the main contributions, accurately reflecting the paper's content and focus.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The limitations of our method, especially in failure scenarios, are further discussed in Section E.3. We further analyze the computational efficiency of our model in Section D.2

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not present theoretical assumptions or provide formal proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide all the training details to reproduce the main experimental results in various sections including Section 4, Section 5.2, and Section A.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We describe how we generated the new dataset in Section C.1. The code and dataset will be made publicly available after the paper is accepted for publication.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide all the training details to reproduce the main experimental results in various sections including Section 4, Section 5.2, and Section A.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We do not report error bars as our evaluation relies on deterministic metrics that yield consistent values across runs. Since the outputs are fixed given the same inputs, statistical variance is not applicable in this setting.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The inference time for image generation is reported in Section D.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Our work conforms with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We address this in Section F for further details.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: **[No]**

Justification: Although we fine-tune a pretrained LLM, our method does not introduce additional misuse risks beyond those associated with the base model. No safeguards were added beyond those provided by the original pretrained model.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: **[Yes]**

Justification: We address this in Section G for further details.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not release any new assets in this paper.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our work does not involve human subjects, and therefore does not require the details mentioned in this point.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our work does not involve human subjects and does not require an IRB approval or equivalent.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method of CAMILA was developed without the involvement of LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.