Blur-Aware Depth Estimation: Enhance Perception in Motion

Boyang Li
University of Toronto
frankli@cs.toronto.edu

Kecen Yao
University of Toronto
kecenyao@cs.toronto.edu

Jinyang Zhao
University of Toronto
jerryzhao@cs.toronto.edu

Abstract—This project aims to improve monocular depth estimation specifically in the presence of motion blur by utilizing Stable Diffusion XL (SDXL) and T2I adapters to augment existing datasets. By generating realistic synthetic depth maps and corresponding RGB images that incorporate both adverse environment conditions (e.g. low light, rain) and motion blur effect, we aim to enhance the training data's ability to simulate challenging real-world dynamics, such as fast-moving vehicles and blurred environments. Fine-tuning state-of-the-art models like Depth Anything V2 [1] with these augmented datasets will enable the model to generate depth maps that realistically reflect motion blur effects. Our approach will be evaluated using benchmark KITTI, aiming to improve the accuracy and consistency of depth estimation in scenarios involving significant motion.

Index Terms—Computational Photography, Depth Estimation, Depth Anything V2, Stable Diffusion XL(SDXL), T2I-Adapters, Midas Depth Estimation, Motion Blur Depth Enhancement

1 Introduction

Depth estimation is an essential task in computer vision with applications in autonomous driving, robotics, and augmented reality. It allows systems to understand the three-dimensional structure of a scene, which is vital for navigation, obstacle avoidance, and interaction with the environment. However, achieving accurate depth estimation is challenging, especially in dynamic scenes where motion blur occurs. Motion blur distorts visual features, making it difficult for models to interpret object depth accurately, particularly in fast-moving objects like cars, bicycles, or drones. These challenges highlight the need for improved depth estimation methods that can handle motion blur effectively.

Current state-of-the-art models, such as Depth Anything V2, perform well in most scenarios but struggle with motion blur. The primary issue is that Depth Anything V2 generates static depth maps for motion-blurred images, treating blurred objects as unmove translucent objects. This leads to unrealistic depth representations, particularly in dynamic scenes with high-speed motion. Such limitations reduce the model's reliability and accuracy in real-world applications where motion blur is common.

This project addresses the challenges of depth estimation under motion blur by enhancing Depth Anything V2's ability to handle such scenarios. Specifically, we aim to incorporate motion blur effects into the depth estimation process to improve model accuracy and realism.

To achieve this, we utilize the Stable Diffusion model with the Depth-MiDaS adapter to augment existing depth estimation datasets. By building on the approach of Tosi et al. [2] , which uses diffusion models to generate images under adverse conditions, we extend the method to include motion blur. Then we use prompts to guide the model to generate realistic RGB images with motion blur effects.

These augmented datasets are then used to fine-tune Depth Anything V2, enabling it to better capture and represent motion blur in depth maps. This fine-tuning process ensures the model produces depth estimates that are consistent with real-world dynamics, improving its robustness and accuracy in challenging, motion-heavy scenarios.

2 Related Work

2.1 Overview of Existing Models

The Depth Anything model (V1) [3] is a deep learning-based method for monocular depth estimation that has shown promising results in standard scenarios. Depth Anything V2 is an advanced version of this model that refines the depth estimation process, enhancing its performance in various real-world applications. It leverages convolutional neural networks (CNNs) trained on large datasets to predict depth from single images. Depth Anything V2 continues the work of its predecessor by providing better generalization across diverse and complex environments. However, it still faces challenges in extremely adverse conditions, such as motion blur.

2.2 Diffusion-Based Approaches for Depth Estimation

Tosi et al.'s paper "Diffusion Models for Monocular Depth Estimation: Overcoming Challenging Conditions" [2] introduces an innovative method to address these challenges using text-to-image diffusion models. Firstly, they use Depth-MiDaS to estimate depth in simple, clear scenes and provide reliable depth information when conditions are not too difficult.

Then they used the stable Diffusion Model to generate images under challenging weather conditions such as rain, snow, fog, low-light environments (night), and reflective or transparent surfaces. This approach improves the performance of depth estimation networks by generating synthetic but realistic data that accounts for these challenging conditions.

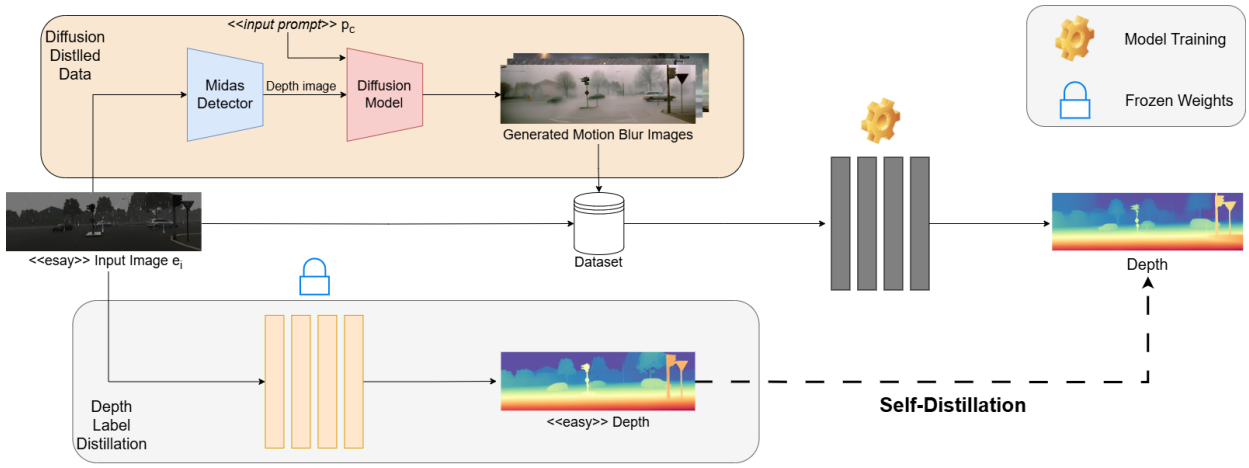

Fig. 1. Flow Chart of Our Method

Finally, they use those newly generated images and depth information to fine-tune the Depth Everything V1 model, improving its ability to handle complex, real-world conditions.

Their use of diffusion models helps produce diverse scenes while preserving underlying depth information, which is crucial for improving the robustness of depth estimation models in real-world scenarios.

2.3 Comparison

Traditional monocular depth models [4] [3] struggle in challenging conditions such as bad weather or motion blur. These models are trained on simpler data and fail to generalize well to complex scenarios.

Tosi et al.'s diffusion-based [2] approach improves on this by generating challenging environments through text-to-image models like ControlNet [5], ensuring that depth estimation models are trained on more diverse and difficult conditions. They also use a self-distillation protocol, which fine-tunes models with synthetic data, improving robustness to environmental changes.

However, Tosi's method [2] doesn't fully address motion blur caused by fast-moving objects, which is a key challenge in real-world dynamic environments.

Our method builds on Tosi's work [2] by incorporating motion blur into the training process, further enhancing depth estimation for fast-moving objects and improving realism in dynamic scenarios.

3 Theory

3.1 Problem Formulation

We observe the primary challenge is that existing monocular depth estimation models, such as Depth Anything V2 [1], generate unrealistic depth maps for motion-blurred scenes, where the model treats the blurred object as static and draws a clear object outline. This leads to inaccurate depth representations inconsistent with the real-world dynamics of moving objects. The motion blur effect occurs when there is relative movement between the camera and the objects in the scene or camera using a long exposure time when looking at moving objects. That kind of effect causes a blending of features that depth estimation models find difficult to interpret, which may confuse the computer and lead to an incorrect operation for things like autonomous driving and robotics.

To address the challenge mentioned above, we reconsider the depth estimation problem by adding the motion blur effect as a factor affecting the output. That is, instead of treating the motion blur effect as noise, we incorporate it into the training process so that the model learns to interpret blurred features and add it as part of the scene's depth structure. To accomplish that, we decide to augment the training data by adding realistic motion-blurred images and then use that augmented dataset to fine-tune the Depth Anything V2 model to enhance its performance and accuracy on the motion blur effect.

3.2 Flowchart Architecture

However, we are not able to find a dataset that provides realistic motion-blurred images with a corresponding depth estimation that explicitly considers the presence of the motion blur effect. Instead of relating to a dataset that provides training data to us, we use the diffusion model to generate our training set by adding different types of adverse environmental conditions and the motion blur effect into a clean image. The given figure 1 shows a flowchart of how we accomplished our goal above, and the flowchart can be divided into two main steps including data augmentation with motion blur and fine-tuning the model. The description of each step will be shown next.

3.2.1 Data Augmentation with Motion Blur

First of all, we randomly pick 2,000 images from KITTIv2 train set as the easy input images. As for each input image, we use ControlNet Midas depth estimation model [5] to generate the corresponding depth image. After that, by using SDXL along with T2I adapters [6] and the above depth image, we can generate challenge RGB images with the same depth structure as the easy input image and with additional adverse environmental conditions and motion blur effect. To encourage SDXL to add required challenge conditions, we employ prompts that include motion blur

and adverse environment descriptions (e.g. "Hazardous smog blanketing city streets with motion blur"), combined with negative prompts (e.g. "Static cars, sharp car details, ...") to avoid generating static or sharp details that are not characteristic of realistic motion blur. Examples are shown by the figure 2 as the following,

Fig. 2. Examples for raw image, depth image and corresponding generated images

Where RGB image is the raw image, and depth image is the depth structure generated by the Midas depth estimation model. All generated images use prompts, which are shown below each image, and the same negative prompt. By using this strategy, we generate 10 images with different adverse environmental and motion-blurred conditions for each input image. After that, we incorporate all input and generated images together to form the training dataset of size 22,000. Combining both static scenes and motion blur images helps the model learn to distinguish between actual object boundaries and blurred features, leading to more accurate depth estimation.

3.2.2 Fine-Tuning the Model

In this stage, the dataset generated above will then be used to fine-tune the model. However, the significant challenge here is that since all motion-blurred images are generated using the Diffusion model, then there is no ground truth depth image for them. And the fine-tuning requires ground truth such that it can compute the model loss which will be used to update model parameters. To accomplish that, we employ the self-distillation protocol. That is, we utilize a robust pre-trained monocular depth network, Depth Anything V2 base model, which will be leveraged to generate depth images for all original, unchallenging scenes in the 2,000 KITTIv2 train set. After that, these generated depth images subsequently serve as pseudo ground truth labels for the corresponding generated images in the training set.

For fine-tuning the model, we employ the pre-trained Depth Anything V2 base model as our start point and train it for 120 epochs using the Adam optimization algorithm, utilizing the augmented dataset above. As for each epoch, the dataset will be shuffled such that the model will observe different sequences of training data which will encourage the model to generalize. Additionally, the model contains 97.5M parameters which are divided into two distinct parts including the pre-trained parameters, which are used to extract useful features from the input images; and the depth head parameters, which are responsible for generating depth images from the extracted features. The learning rate lr is initialized as 5×10^{-6} and will be gradually reduced to 5×10^{-7} during the training process. The learning rate for the pre-trained parameters is initially set to the same value as lr, while the learning rate for the depth head parameters is set to 10 times the value of lr. This is because the pre-trained parameters require minimal adjustments and so less change, whereas the depth head parameters, which are trained from scratch, necessitate a higher learning rate to facilitate faster convergence.

4 Analysis and Evaluation

In this section, we are going to introduce the way used to analyze and evaluate our fine-tuned Depth Anything V2 model. In short, we illustrate how we split the augmented training set above during the fine-tuning process such that one part is used for training the model and another part is for validating the model performance for each epoch. To evaluate the accuracy, we present all metrics we have used. Finally, we demonstrate how we are going to compare our fine-tuned model with the original Depth Anything V2 model qualitatively. More detailed results will be presented in section 5.

4.1 Datasets

We use the augmented dataset of size 22,000 illustrated above for both generating the train set and validating in each epoch. To make sure the model will not be trained and validated with the same image, we split that dataset so that 21,260 images will be used to train the model, whereas the remaining 740 images will be used to validate the current model performance. Since the dataset is separated randomly, the evaluation set contains both generated motion-blurred images and the original easy image in order to evaluate the model performance on distinguishing between motion-blurred and static objects.

Besides the generated dataset based on KITTIv2, we also evaluate using additional images taken by our mobile devices. To capture photos containing both motion-blurred and static objects, we find a place on University Avenue, where there are both vehicles parked on the side of the street and some moving cars. The photos are taken using ISO 50 and shutter speed of 0.1 seconds such that the motion blur effect will appear on all fast-moving objects. Also, the tripod is used to avoid the device movement during the camera exposure. Examples are the RGB images shown in the right part of the figure 3.

4.2 Metrics

To evaluate both model's accuracy and robustness, during the fine-tuning process we choose to use the metrics including Absolute Relative Error (AbsRel), Squared Relative Error (SqRel), Root Mean Squared Error (RMSE), Root Mean Squared Logarithmic Error (RMSLE), Mean Logarithmic Base-10 Error (Log10), and Scale-Invariant Logarithmic

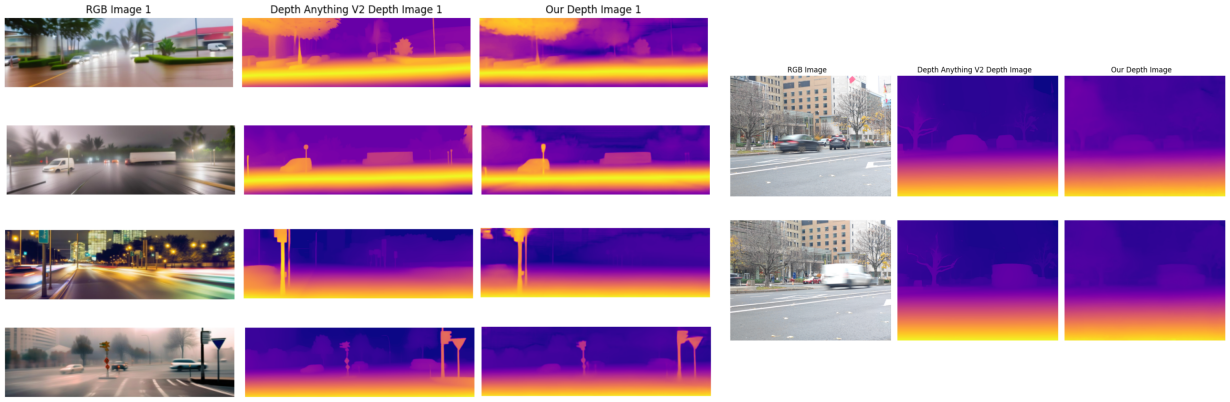

Fig. 3. Visualization result of our model.

Error (SILog). We also compute the percentage of pixels for which the ratio between the predicted and target depth is within a threshold of 1.25, 1.25^2, and 1.25^3. Throughout the training process, we are willing to observe that all the computed metrics preserve an overall downward trend with initial oscillations and eventual stabilization. Inversely, all percentages of pixels with depth prediction accuracy within a threshold should have an overall upward trend. The trend chart of AbsRel and RMSE is shown in figure 4.

4.3 Model Comparisons

We compare our fine-tuned Depth Anything V2 model only with the original base model. Our focus is on assessing performance under motion blur conditions, specifically evaluating whether our fine-tuned model will generate depth images accurately reflecting the blurred dynamics of the scene. At the same time, we are willing to observe that the original Depth Anything V2 model struggles to accurately estimate depth in motion-blurred scenes, and then it generates clear outlines for blurred objects instead.

5 Results

In this section, we present a comprehensive evaluation of our fine-tuned Depth Anything V2 model's performance on motion-blurred images. Our analysis encompasses both qualitative assessments of depth map generation and quantitative measurements of model performance through various metrics. The results demonstrate significant improvements in depth estimation accuracy for motion-blurred scenes compared to the baseline model, with particular emphasis on maintaining structural coherence in areas affected by motion blur. Through extensive experimentation and rigorous evaluation, we validate our approach's effectiveness in addressing the challenges posed by motion blur in depth estimation tasks.

5.1 Qualitative Analysis

Figure 3 visually compares the depth maps produced by the original Depth Anything V2 model and the fine-tuned version on motion-blurred images. Figure 3 (left) demonstrates depth estimation performance under motion blur using RGB images from the KITTIv2 test set, where depth maps from the standard Depth Anything V2 model (middle column) are contrasted with those from our fine-tuned version (right column), optimized for motion blur. The original model tends to generate unrealistic sharp edges for blurred objects, whereas our fine-tuned model produces depth maps that more accurately reflect the blurred dynamics. Figure 3 (right) extends this comparison to depth estimation on motion-blurred images captured by a mobile phone, further validating our approach.

Qualitative results showing that the fine-tuned model is better able to capture the depth of fast-moving vehicles and pedestrians. Our fine-tuned model demonstrates the ability to incorporate motion blur into the depth field, thereby producing depth maps that faithfully represent both the static and dynamic elements of the scene. In contrast, the standard Depth Anything V2 model is limited to generating depth maps as if all objects were still, effectively ignoring the motion blur. The depth maps generated by the fine-tuned model are more consistent with the visual appearance of the scene, providing a more realistic representation of depth under motion blur conditions.

5.2 Training Stability and Quantitative Results

The training process was monitored using both training and validation metrics. Initial fluctuations were observed due to the complexity of motion blur, but these metrics stabilized over time, indicating effective learning of the blur-aware features. The validation performance consistently improved, demonstrating the robustness of our approach.

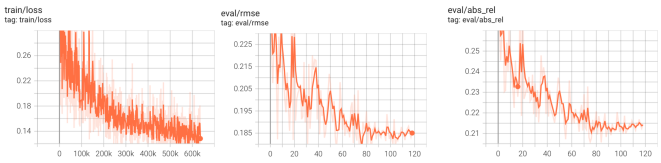

Fig. 4. Detailed result during training.

Figure 4 presents three key metrics during model fine-tuning: training loss (left), evaluation RMSE (center), and evaluation absolute relative error (right). The training loss exhibits a characteristic decreasing trend from approximately 0.26 to 0.14 over 600,000 iterations, demonstrating

consistent optimization convergence. The evaluation metrics - RMSE and absolute relative error - show similar convergent behavior over approximately 120 epochs, stabilizing at approximately 0.185 and 0.215 respectively.

The training dynamics reveal several noteworthy patterns. Initially, there are pronounced oscillations in all metrics during the early stages (first 100,000 iterations for training loss and first 20 epochs for validation metrics), which can be attributed to the model's adaptation to the complex characteristics of motion blur in depth estimation. These fluctuations gradually diminish as training progresses, indicating the model's improving stability in handling motion-blurred inputs.

Of particular interest is the convergence behavior in the validation metrics (RMSE and absolute relative error), which demonstrate asymptotic stabilization without signs of overfitting. The RMSE settles around 0.185 with minimal variance, while the absolute relative error maintains stability near 0.215, suggesting robust generalization to unseen motion-blurred scenarios. This parallel optimization of multiple metrics provides strong evidence for the model's successful adaptation to the motion blur domain while maintaining depth estimation accuracy.

The synchronous improvement across all metrics, coupled with their eventual stabilization, substantiates the effectiveness of the fine-tuning strategy in incorporating motion blur handling capabilities into the depth estimation framework. The absence of divergent behavior in the validation metrics further confirms the model's robustness and practical applicability to real-world scenarios involving motion blur.

These empirical results provide quantitative support for the qualitative improvements observed in the model's depth estimation performance on motion-blurred images, establishing a strong foundation for its deployment in dynamic real-world environments where motion blur is prevalent.

6 Discussion and Conclusion

This research investigates the integration of motion blur considerations into monocular depth estimation models, specifically focusing on the enhancement of the Depth Anything V2 architecture. Our study addresses a critical gap in current depth estimation approaches, which typically assume static scene conditions and struggle with motion-induced blur effects. Through systematic experimentation and analysis, we demonstrate significant improvements in depth map generation for dynamic scenes while identifying key limitations and future research directions. The findings presented herein contribute to the broader understanding of motion blur handling in computer vision tasks and have practical implications for real-world applications such as autonomous systems and robotics. In this section, We will address our key findings, practical implications, current limitations, and proposed future research directions to advance the field of motion-aware depth estimation.

6.1 Key Findings

As mentioned in Section 5, our experiments demonstrate that incorporating motion blur into the training process of monocular depth estimation models significantly improves their ability to generate realistic depth maps for dynamic scenes. The fine-tuned Depth Anything V2 model was able to represent moving objects more accurately, providing depth estimates that were consistent with real-world motion blur.

Fig. 5. Examples of successful and failed cases for motion blur images.

6.2 Implications

This improvement has important implications for applications such as autonomous driving and robotics, where understanding the depth of fast-moving objects is critical. By generating depth maps that correctly reflect motion blur, our model provides more reliable information for decision-making in these applications. For example, in autonomous driving, accurately estimating the depth of a rapidly approaching vehicle can help improve the safety and reliability of the system.

6.3 Limitations

Our approach has several limitations that need to be addressed in future work:

SDXL Limitations: Stable Diffusion XL (SDXL) does not consistently produce realistic motion blur effects in the generated images, as in Figure 5. In some cases, the generated images still contain static cars without realistic motion blur, which limits the effectiveness of the augmented training data.

Lack of Motion Blur Evaluation Datasets: Currently, there are no publicly available datasets that specifically provide motion-blurred images along with the corresponding ground truth depth images. This makes it difficult to evaluate the model's performance on real motion blur scenarios and limits the diversity of training data.

Incorrect Ground Truth for Motion Blur: The ground truth depth maps for the generated motion blur images are derived from the original, clean images. These depth maps are static and contain clear car edges, which means they do not accurately represent the depth of motion-blurred scenes. As a result, the ground truth is incorrect for all generated motion blur images, which impacts the accuracy of training and evaluation for motion blur depth estimation.

Incorrect Ground Truth for Motion Blur: The ground truth depth maps for the generated motion blur images are pseudo ground truth obtained using the pre-trained Depth Anything V2 model on original, clean images. These depth maps are static and contain clear car edges, as they are

generated from single-frame raw images without motion blur. This presents a fundamental mismatch since motion blur is inherently a result of temporal integration over an exposure period, while the pseudo-ground truth is derived from a single frame captured within that period. As a result, these pseudo-ground truth depth maps do not accurately represent the depth characteristics of motion-blurred scenes, where the blur effect encompasses object movement throughout the entire exposure duration. This temporal disparity between single-frame pseudo ground truth and motion-blurred images, which represent integrated motion over time, impacts the accuracy of both training and evaluation for motion blur depth estimation.

6.4 Future Work

To address the limitations identified, several future directions are planned:

Improving Motion Blur Generation: Since SDXL struggles to produce realistic motion blur, we will explore other generative models, such as GANs or VAEs, which may provide better results for generating motion-blurred images. Specifically, we will aim to generate images that more accurately reflect dynamic motion, particularly for vehicles, to enhance the quality of the augmented training data.

Developing a Motion Blur Dataset: Given the lack of a suitable dataset containing both motion-blurred images and corresponding ground truth depth maps, we plan to either create a new dataset or collaborate with others to collect such data. This dataset will be critical for evaluating the model's performance under realistic motion blur conditions and for providing more diverse training data.

Generating Correct Ground Truth Depth: The current approach uses ground truth depth maps from clean images, which do not accurately represent the blurred dynamics. In the future, we plan to develop techniques to generate more appropriate ground truth depth maps for motion-blurred images. This may involve using physics-based models to simulate realistic depth for blurred objects or employing multi-view stereo techniques to create depth maps that reflect the true motion in the scene.

Multi-View Stereo for Motion Blur: We will investigate the use of multi-view stereo techniques to improve depth estimation in motion-blurred scenes. By combining information from multiple viewpoints, we aim to generate more accurate depth estimates that account for the dynamic nature of the scene.

REFERENCES

[1] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *arXiv:2406.09414*, 2024.

[2] F. Tosi, P. Zama Ramirez, and M. Poggi, "Diffusion models for monocular depth estimation: Overcoming challenging conditions," in *European Conference on Computer Vision (ECCV)*, 2024.

[3] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *CVPR*, 2024.

[4] G. et al., "Defeatnet: General monocular depth via simultaneous unsupervised representation learning," in *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[5] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023.

[6] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," *arXiv preprint arXiv:2302.08453*, 2023.