

Supplementary material: Key-point Guided Deformable Image Manipulation Using Diffusion Model

Seok-Hwan Oh^{*1}

joseph9337@kaist.ac.kr

Guil Jung^{*1}

jgl97123@kaist.ac.kr

Myeong-Gee Kim²

suhakprince@kaist.ac.kr

Sang-Yun Kim¹

kmjmkysy@kaist.ac.kr

Young-Min Kim¹

youngmin2007@kaist.ac.kr

Hyeonjik Lee¹

dlguswlr0811@kaist.ac.kr

Hyuk-Sool Kwon³

jinuking3g@snuh.org

Hyeon-Min Bae¹

hmbae@kaist.ac.kr

¹ Electrical Engineering Department

KAIST

Daejeon, South Korea

² Barreleye Inc.

Seoul, South Korea

³ Department of Emergency Medicine

SNUBH

Seong-nam, South Korea

The supplementary materials demonstrate the evaluation of the optical flow estimation (Section 1), effectiveness of the proposed continuous image generation (Section 2), examination of the impact of initialization with linear deformation (Section 3), include additional results on drag-based image editing (Section 4), implementation details (Section 5), and limitations (Section 6). The code for the KDM framework can be found on GitHub at <https://github.com/joseph9337/KDM-Net>.

1 Optical Flow Estimation

The following sections include assessments and discussions of the optical flow estimation schemes employed in the KDM framework. Section 1.1 demonstrates the accuracy of the optical flow estimation network (OF-net). Section 1.2 presents the effectiveness of the key point guided flow synthesis network (K2F-net) in generating the optical flow estimates.

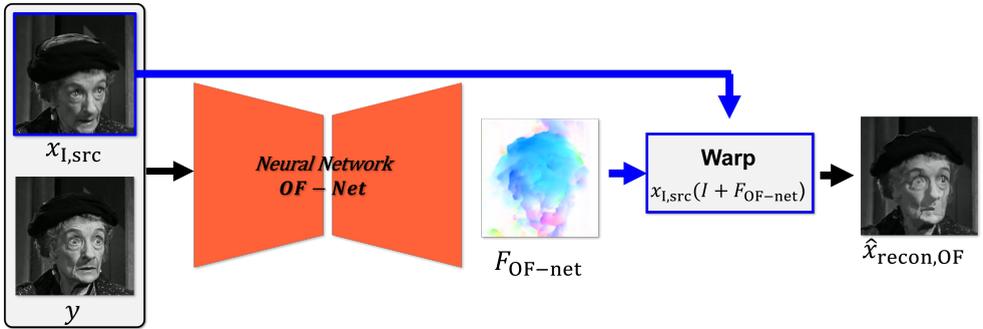
Figure 1: Illustration of the $\hat{x}_{\text{recon,OF}}$ generation procedure.

Table 1: Quantitative assessment of the OF-net and K2F-net.

Model		MSE_{img}	$LPIPS_{alx}$	$LPIPS_{vgg}$
Facial image	K2F-net	0.0222	0.2333	0.3161
	OF-net	0.0080	0.1083	0.1709
Human pose	K2F-net	0.0013	0.1068	0.1494
	OF-net	0.0002	0.0527	0.0831
Echocardiography	K2F-net	0.0121	0.1980	0.3028
	OF-net	0.0014	0.1128	0.2231

1.1 Assessment of the OF-net.

The OF-net generates the optical flow field $F_{\text{OF-net}}$, employing the source $x_{I,\text{src}}$ and target y images as inputs. The warped target image $x_{\text{recon,OF}}$ is generated through linear deformation of the $x_{I,\text{src}}$ with $F_{\text{OF-net}}$ (described in Figure 1). The quantitative assessment is performed by comparing the warped image $\hat{x}_{\text{recon,OF}}$ with the ground-truth target image y . The Mean Squared Error (MSE_{img}) and $LPIPS$ metrics, described in the main script, are employed as evaluation metrics. The OF-net demonstrates 0.0080 and 0.0002 MSE_{img} for the facial image and human pose datasets, respectively. Figure 2, 3, and 4 present a qualitative assessment of the $\hat{x}_{\text{recon,OF}}$ for the facial image, human pose, and echocardiography generation, respectively. The global features of the $\hat{x}_{\text{recon,OF}}$, such as the shape of the mouth, are accurately generated.

1.2 Assessment of the K2F-net.

Table 2: Quantitative assessment of the K2F-net in optical flow estimation.

Model		MAE	$RMSE$
K2F-net	Facial image	6.936	9.720
	Human pose	3.535	7.032
	Echocardiography	2.231	3.471

The K2F-net generates the optical flow $F_{\text{K2F-net}}$, employing the key-point x_K and $x_{I,\text{src}}$. The performance of the K2F-net is evaluated by comparing the accuracy of the $F_{\text{K2F-net}}$

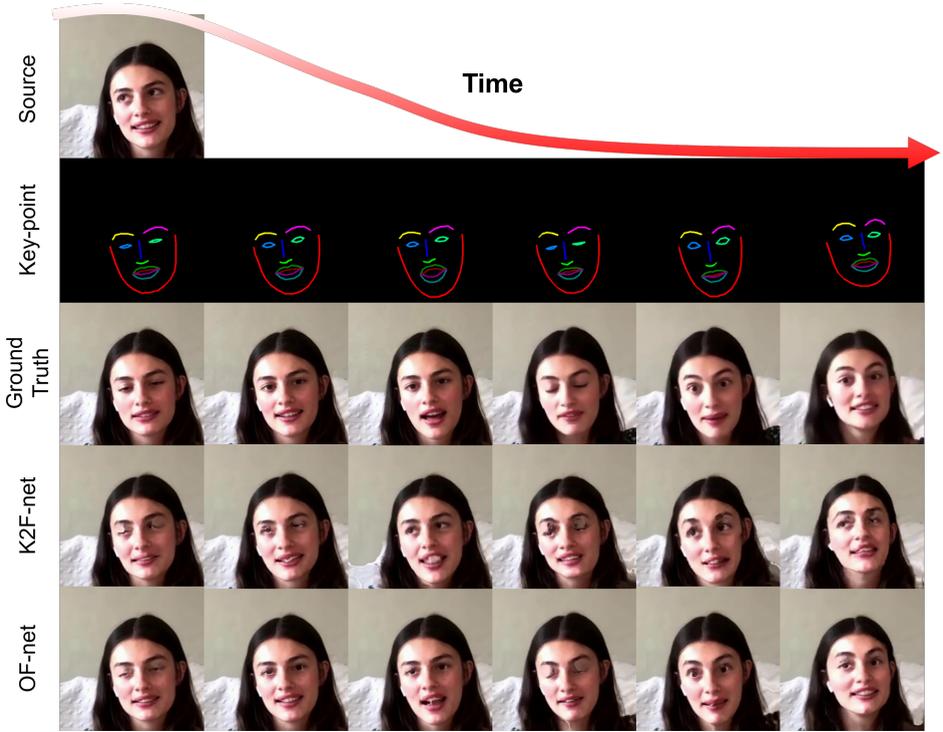


Figure 2: Qualitative assessment of the K2F-net and OF-net for facial image generation.

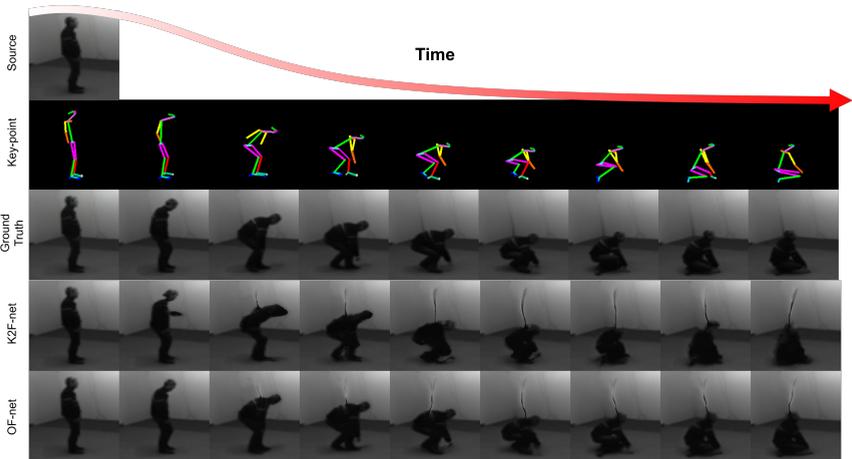


Figure 3: Qualitative assessment of the K2F-net and OF-net for human pose generation.

with the F_{OF-net} . The Root Mean Squared Error (*RMSE*) and Mean Absolute Error (*MAE*) are employed as the evaluation metrics. Table 2 demonstrates quantitative assessments of the K2F-net. The K2F-net demonstrates 9.720 *RMSE* for facial optical flow estimation. Figure 2-4 3rd rows present a warped image $x_{I,src}(I + F_{K2F-net})$ using $F_{K2F-net}$. The evaluation demonstrates that the K2F-net achieves precise estimation of the optical flow using the key-point condition.

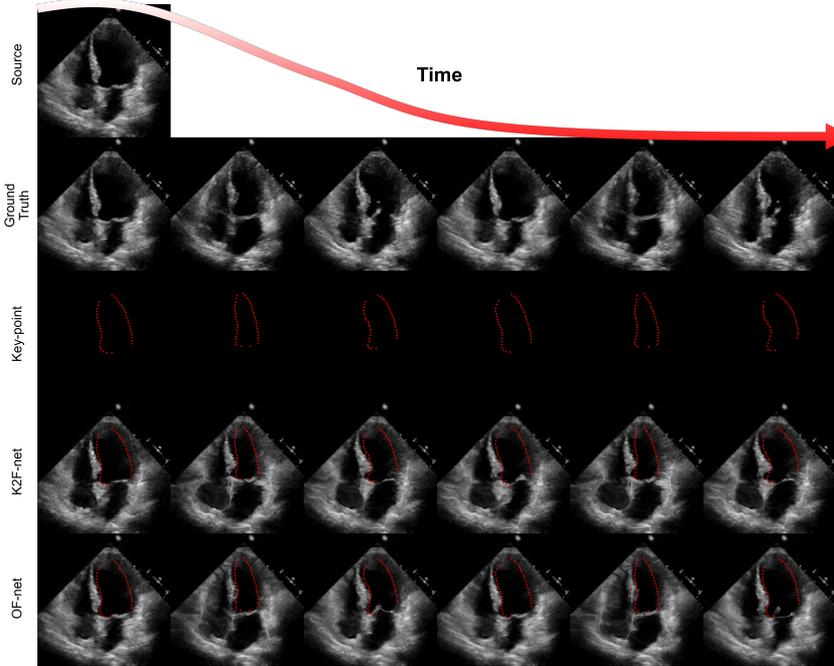


Figure 4: Qualitative assessment of the K2F-net and OF-net for echocardiography image generation.

2 Continuous Image Generation

In this section, additional experiments on continuous image generation are provided. The continuous sequential images are synthesized by the following procedure. Firstly, an optical flow matrix $F_{K2F-net}$ is generated through the denoising process of the K2F-net. Subsequently, the magnitude of $F_{K2F-net}$ is adjusted to $\gamma \cdot F_{K2F-net}$, where $0 \leq \gamma \leq 1$. The $\gamma \cdot F_{K2F-net}$ is employed as an optical flow of the intermediate frame image. Finally, the F2I-net generates intermediate frame images $\phi_{F2I}(x_{I,src}, x_K, \gamma \cdot F, t)$, utilizing the intermediate optical flow $\gamma \cdot F_{K2F-net}$ and the source image $x_{I,src}$. The proposed continuous image generation scheme enhances the efficiency of the KDM framework by reducing the need to infer optical flow for every intermediate frame image. Figure 5 presents continuous image generation of the human face, pose, and echocardiography images, where $\gamma = (\frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \dots, \frac{6}{6})$. The KDM framework succeeds in generating sequentially consistent and authentic images. The continuous generation of KDM can be applied to a range of applications such as video generation and image frame interpolation.

3 Ablation Study: Initialization with Linear Deformation

In this section, we conduct ablation studies on image initialization, which is described in Section 3.4 of the main paper. Figure 6 presents a qualitative assessment of facial image manipulation with varying k parameters. When $k = 0$, the F2I-net is not applied for the x_{recon} generation, and the x_{recon} is identical to $x_{I,src}(I + F_{F2I-net})$. Consequently, the network fails to illustrate detailed facial expressions. For $k = 200$, the F2I-net initiates denoising from

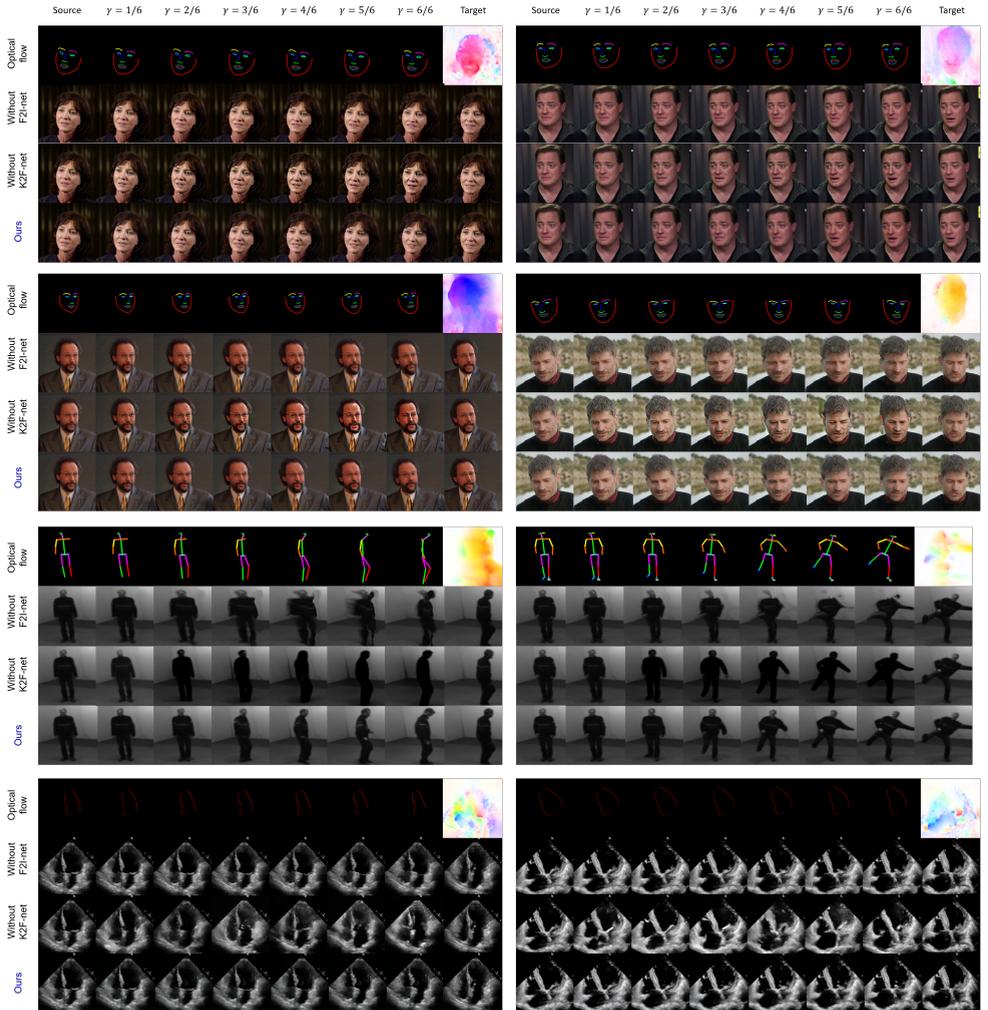


Figure 5: Qualitative assessment of continuous image generation of facial image synthesis, human pose generation and echocardiography video prediction.

$N(0, 1)$, which results in limited sequential consistency of the image. To enhance sequential consistency while achieving precise image synthesis, we set k to 100.

4 Drag-based Image Editing

The KDM framework can be employed for drag-based image manipulation. In drag-based image editing, a user defines a motion vector \mathbf{v} through mouse dragging [10]. Then, the system provides a desired image that corresponds to the user's drag objective. The KDM framework implements the drag-based image editing by generating the target key-point $\mathbf{x}_{K,tgt} = \mathbf{x}_{K,src} + \mathbf{v}$, which is derived through applying motion vector \mathbf{v} to the source key-point $\mathbf{x}_{K,src}$. Figure 7 presents the drag-based human pose editing of the DragGan [2], DragDif-

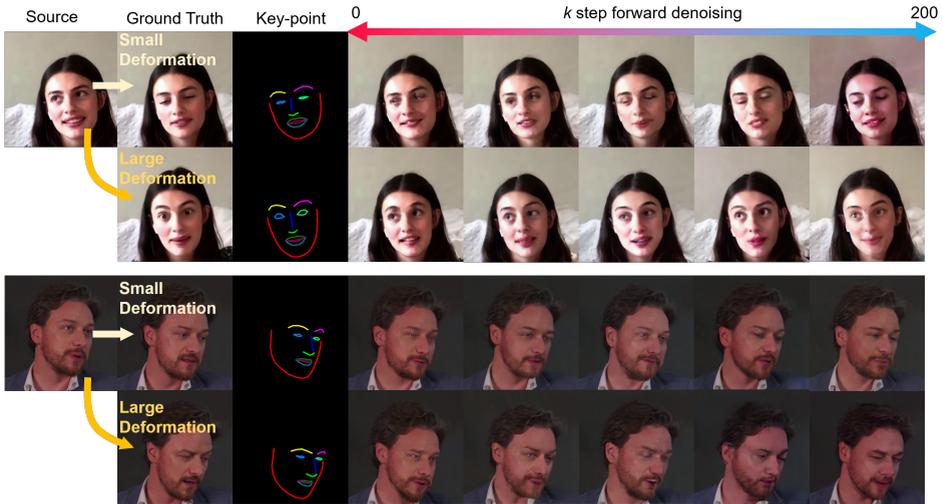


Figure 6: Qualitative assessment of the image initialization with varying k variable is provided.

fusion [9], and KDM framework. The KDM demonstrates versatile image generation under diverse \mathbf{v} conditions. In Figure 7 first row, a drag-based image editing of lifting the left leg is introduced. While DragGan and DragDiffusion show difficulties in accurately representing the modified left leg pose, the KDM achieves precise generation of every body part of the human.

5 Implementation details

AdamW optimizer is employed for the training of the KDM. In order to enhance computational efficiency during the sampling procedure, the images are generated under DDIM process with $T = 200$ denoising steps.

For the comparative study, the single-point editing process of UserControllableLT is extended to multiple key-point manipulations by formulating flow vectors of each key-point. DragGan enables drag-based multi-point manipulation through iterative optimization of the latent space of a pre-trained generative adversarial network. In UserControllableLT and DragGan, the PTI inversion is implemented to embed an image into the latent space. DragDiffusion achieves drag-based diffusion probabilistic image synthesis by employing the LoRA scheme. For validation of drag-based editing models, we configure the number and value of the editing vectors to that of key-points. PG2 proposes to generate person images using target key-points and input image employing a dual-generator model. C2GAN enhances robustness of the key-point conditioned image synthesis through end-to-end training of three cycles generative models. PIDM achieves key-point guided image generation by employing DDPM with cross-attentional condit

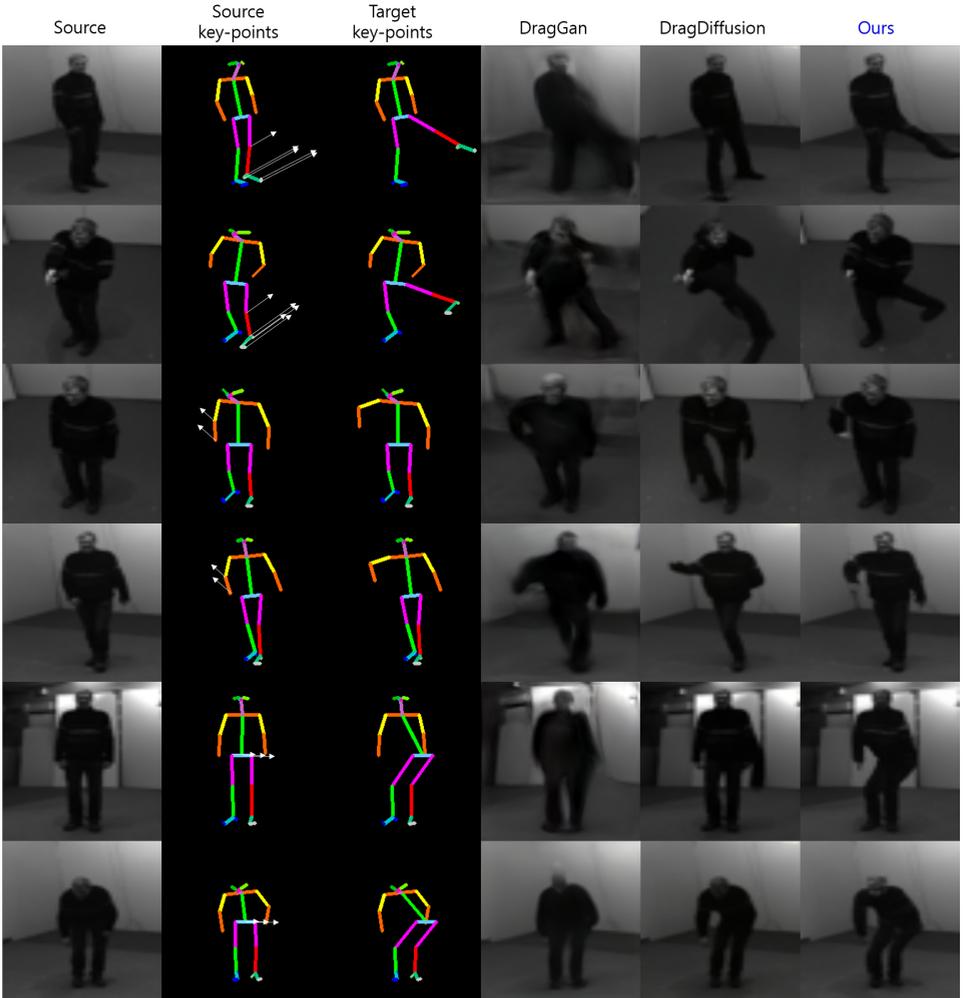


Figure 7: Qualitative assessment of the drag-based image editing. White arrow in source key-points denotes drag vector \mathbf{v} .

6 Limitation

In this section, we present the limitations of the KDM framework. Within the KDM framework, image manipulation is restricted to objects that are semantically related to the key-points. Consequently, the versatility and effectiveness of image manipulation are significantly dependent on the configuration of the key-points. For example, in facial image manipulation, the key-point that indicates the position of the human pupil is not included among the 68 facial landmarks. As a result, the position of the pupil remains unchanged, leading to unnatural image generation, as depicted in Figure 8. We anticipate configuring appropriate key-points tailored to the application of the KDM framework, thereby expanding its applicability and facilitating precise image manipulation aligned with specific objectives.

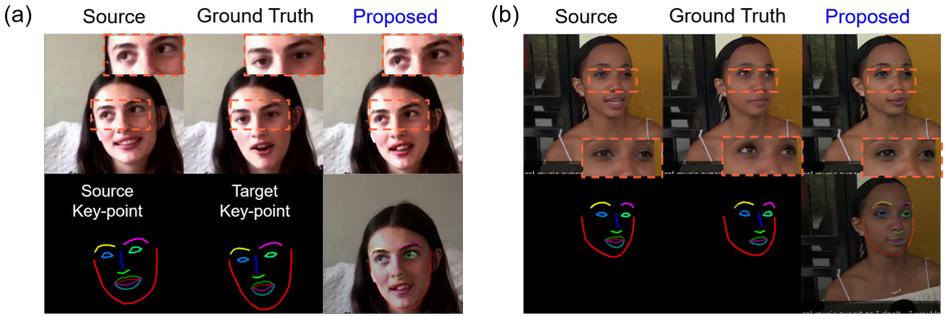


Figure 8: Unnatural image examples of the KDM framework.

References

- [1] Yuki Endo. User-controllable latent transformer for stylegan image layout editing. In *Computer Graphics Forum*, volume 41, pages 395–406. Wiley Online Library, 2022.
- [2] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.
- [3] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *arXiv preprint arXiv:2306.14435*, 2023.