일반논문 (Regular Paper)
방송공학회논문지 제29권 제5호, 2024년 9월 (JBE Vol.29, No.5, September 2024)
https://doi.org/10.5909/JBE.2024.29.5.662
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

재난 경보에서 시각적 콘텐츠 생성을 위한 텍스트-투-3D 시네마그래프: LLM 및 확산 모델을 사용한 생성형 AI 프레임워크

원 루 빈[a], 최 민 지[a], 최 지 훈[b], 배 병 준[a)b)‡]

Text-To-3D Cinemagraphs for Generation of Visual Content in Disaster Alerts: A Generative AI Framework with LLMs and Diffusion Models

Ru-Bin Won[a], Minji Choi[a], Ji Hoon Choi[b], and Byungjun Bae[a)b)‡]

요 약

본 연구는 재난 경보 시스템에서 생성 AI 기술을 사용하는 새로운 프레임워크 Text-To-3D Cinemagraph를 제안한다. 이 프레임워크는 텍스트 및 이미지 생성과 광학 흐름, 3D 카메라 움직임과 같은 애니메이션 기술을 결합하여 동적인 시각 경보를 만든다. 현재의 Text-To-Video 기술이 복잡하고 자원을 많이 소모하는 것과 달리, Text-To-3D Cinemagraph 접근 방식은 더 간단하고 빠르며 재난 시나리오와 같은 특정 도메인에 맞게 설계가 가능하다. 먼저, 실제 이미지에서 추출한 메타데이터를 사용하여 LLM model을 fine-tuning 하여 만든 metadata generator를 통해 이미지 생성을 위한 prompt를 만든 후, Diffusion based Text-To-Image Generative Model을 통해 이미지를 생성한다. 이를 애니메이션화하여 재난을 생생하게 묘사함으로써 취약계층을 위한 재난 경보의 명확성과 접근성을 향상시킨다.

Abstract

The study proposes a novel framework called Text-To-3D Cinemagraph to enhance disaster communication using generative AI technologies. This framework uses a combination of text and image generation along with animation techniques, such as optical flow and 3D camera movements, to create dynamic visual alerts. Unlike current Text-To-Video technologies, which are complex and resource-intensive, the Text-To-3D Cinemagraph approach is simpler, faster, and more adaptable, specifically designed for disaster scenarios. The LLM model is fine-tuned using metadata extracted from real-world images and serves as a metadata generator to create prompts for image generation. The images are generated by a diffusion-based Text-To-Image Generative Model and are then animated to vividly depict disasters, improving the clarity and accessibility of alerts for vulnerable populations.

Keyword : Deep Generative AI, Text-To-3D Cinemagraph, Text-To-Image (T2I), Fine-tuning, Large Language Model (LLM)

I. Introduction

Current disaster alert systems often depend on text messages, which may not effectively reach groups like the elderly, children, and those with disabilities. This study introduces a new method using advanced generative AI[3,16,22] to improve disaster communication. Our approach, called Text-To-3D Cinemagraph, combines text generation, image creation, and animations such as optical flow and 3D camera movements to develop dynamic visual alerts. Despite their impressive capabilities, T2V models face significant technical challenges due to their complexity. Current Text-To-Video (T2V) generation technologies, like Google's Imagen Video[19] and Meta's Make-A-Video[13], use Diffusion Video Modeling[12,20,21], which is complex and resource-intensive. These models utilize millions of parameters which can be computationally intensive and resource-heavy[11]. In contrast, our framework is simpler and more accurate, as we produce a single image highly related to the disaster scenario using the fine-tuned LLM. The so-called meta generator produces the prompt, and we then animate the image using Eulerian Motion Fields[9,17] and camera movements.

Meanwhile, research on Text-To-3D generative models has been rapidly advancing[25,26,27]. For instance, Dual3D[25] introduced a novel dual-mode multi-view latent diffusion model that efficiently generates high-quality 3D assets by toggling between 2D and 3D modes during the denoising process, significantly reducing computational costs without sacrificing quality. Similarly, DreamBooth3D[26] combines the personalization capabilities of the Text-to-Image model DreamBooth[27] with the Text-To-3D generation framework DreamFusion[28], enabling the creation of personalized 3D assets from a few images. However, while these models focus on general 3D object generation rather than specific scenarios, our proposed framework is centered on generating accurate final results that closely reflect the input text, especially in the context of disaster scenarios.

We propose the Text-To-3D Cinemagraph framework to address various disasters in Korea, including droughts, earthquakes, floods, heavy snow, typhoons, and wildfires. In this research, we focused on disasters－specifically floods and wildfires－to demonstrate the vivid results. To train our model, we collected a dataset of disaster images, which the metadata, the textual data describing each image are extracted from the each image using image captioning technology from the BLIP model[4] and then used to fine-tune the large language model, LLaMA-2[14,15], for generating disaster-relevant metadata. This metadata is used as a prompt of Text-To-Image (T2I) model, which is the next step of the process. When natural disasters occur and the disaster alerting system is activated, the system uses keywords from CAP data to generate prompts for the T2I model. These prompts create a single disaster image that is then animated into a 3D cinemagraph, offering a vivid depiction of disaster scenarios. Our animation process builds on our foundational research[1,2,10], incorporating automated mask generation while maintaining 3D camera pose features. This method significantly reduces computational demands compared to more complex T2V models, improving both the clarity and accessibility of disaster information. Additionally, the adaptability and efficiency of the Text-To-3D Cinemagraph framework have potential applications beyond disaster communication, making it ideal for any sector requiring rapid, clear, and accessible visual communication.

a) 과학기술연합대학원대학교(University of Science and Technology(UST))
b) 한국전자통신연구원(Electronics and Telecommunications Research Institute(ETRI))
‡ Corresponding Author : 배병준(Byungjun Bae)
E-mail: 1080i@etri.re.kr
Tel: +82-42-860-3888
ORCID: https://orcid.org/0000-0002-0872-325X
※ This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (2022-0-00083, Development of customized disaster media service platform technology for the vulnerable in disaster information awareness)
· Manuscript July 3, 2024; Revised August 14, 2024; Accepted August 19, 2024.

Our contributions can be summarized as follows:

1. Pioneering 3D Cinemagraph Generation from Text Descriptions: We believe this is the first study to create 3D cinemagraphs directly from text descriptions. By using a fine-tuned LLM[23,24] and a meta generator for high accuracy, the process is fully automated, eliminating the need for human labor.

2. Contextually Relevant and Accurate Image Generation: The metadata produced by the LLM, which was fine-tuned using metadata from a real image dataset, served as an effective prompt for the Text-To-Image generative model, leading to contextually relevant generated images. This ensures that our visual outputs closely mirror real-world scenarios, providing reliable and actionable information during disasters.

3. Elimination of Manual Labeling through Automated Mask Generation: Our framework introduces a fully automated system for creating 3D cinemagraphs, leveraging the LangSAM model to automatically generate masks and segment motion areas within the generated images.

The key objective of this research is to develop an efficient, automated framework for generating 3D cinemagraphs from text descriptions. Unlike previous methods that required manual input[1] or relied on complex resources[19,13], our approach leverages advanced AI to create accurate and contextually relevant disaster visuals, significantly reducing both computational complexity and manual effort.

II. Related Work

1. Domain-Specific Data Generation

Fine-tuning a Language Model (LLM) for domain-specific tasks, such as generating prompts for Text-To-Image (T2I) models, involves adapting a pre-trained LLM to produce text descriptions that align with visual content. This process incorporates image-related metadata, enhancing the LLM's ability to create contextually relevant and detailed prompts for accurate image generation. For example, enriching the LLM with object data within an image improves its ability to produce precise textual descriptions.

The process begins with collecting real-world image data, processed through the BLIP model (Bootstrapped Language Image Pretraining) to generate descriptive captions. This metadata fine-tunes the LLM parameters, tailoring it to generate suitable prompts for visual content

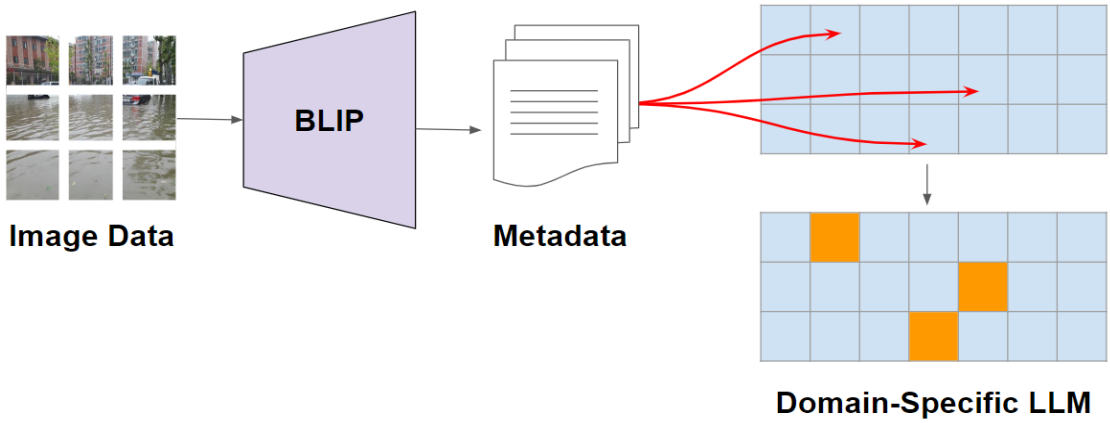

그림 1. 도메인 특화 작업을 위한 LLM의 파인튜닝
Fig. 1. Overview of Fine-Tuning an LLM for Domain-Specific Tasks

related to disaster scenarios. The arrows in Figure 1 indicate how metadata influences LLM parameters, enhancing task performance related to the initial image data.

Our goal is to create highly accurate textual descriptions representing complex disaster scenarios, converting them into image prompts. Domain-specific data generation ensures the generated content is rooted in disaster domain nuances, enhancing the quality of text generation and fostering a comprehensive understanding of the domain for richer, more accurate content.

2. BLIP Model for Image Captioning

Image captioning is crucial in computer vision and natural language processing, aiming to generate textual descriptions of images. Our goal is to fine-tune the LLM model to create high-quality prompts for T2I generation using metadata from real images and image captioning techniques. The BLIP (Bootstrapped Language Image Pretraining) model excels in this task due to its integrated understanding of visual and textual data.

BLIP's architecture processes images and text together, generating relevant and contextually appropriate captions. It identifies objects and describes their attributes, actions, and interactions, providing detailed and nuanced captions. BLIP's standout feature is incorporating context, considering both immediate visual details and broader thematic elements, enhancing the captions' relevance and informativeness, especially in complex scenes like social events.

3. LangSAM for Automated Mask Generation

The LangSAM (Language Segment-Anything Model)[5] enhances image segmentation by integrating natural language processing with text descriptions. Unlike traditional models relying solely on visual data, LangSAM uses text cues for precise object and area identification in images. Combining instance segmentation with text prompts,

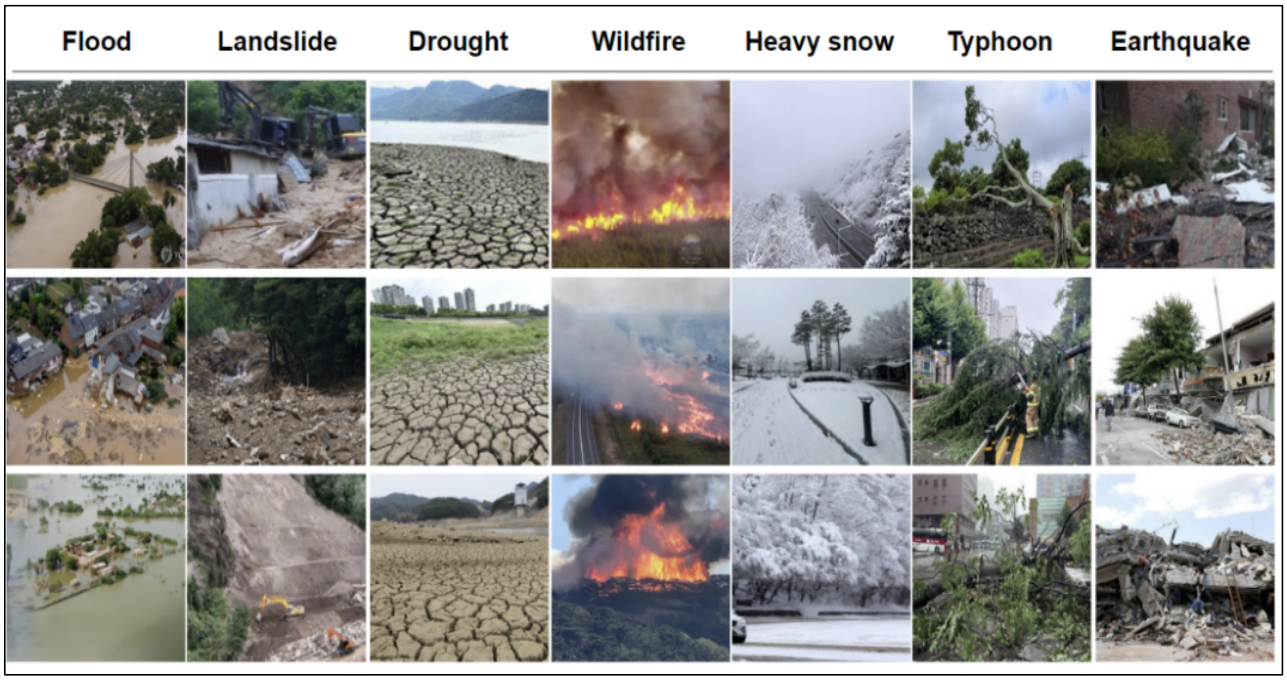

그림 2. 재난 카테고리별 이미지 데이터셋
Fig. 2. Image Data Set Across the Disaster Categories

LangSAM accurately creates masks for specific objects. Building on Meta's segment-anything model and the GroundingDINO detection model, it uses text prompts to highlight objects like "fire" or "river". This makes LangSAM a powerful tool for tasks merging textual and visual data.

In cinemagraph creation, LangSAM generates masks from text descriptions, identifying which image parts should move or stay still. This streamlines the process, especially when combined with Stable Diffusion[6] to create initial images from text. LangSAM's precision and understanding of text and visuals mark a significant advancement in automating cinemagraph production, improving accuracy and comprehension of content.

III. Proposed Method

1. Dataset Preparation

We assembled a dataset of 200 images for each disaster type－flood, landslide, drought, wildfire, heavy snow, typhoon, and earthquake－sourced exclusively from Korean domain websites. Automated tools were used for the initial collection to gather a diverse range of images relevant to each disaster type.

A manual review followed to eliminate images that did not accurately represent the disaster scenarios. This process ensured the dataset only included relevant images, removing those with embedded text, excessive blurriness, or a high presence of human subjects to maintain quality and relevance.

2. Training the LLM Model Using Metadata from a Real Image

In this research, we fine-tuned a cutting-edge large language model. Our base model, Meta's LLaMA-2 Model allows the detailed modifications of transformer layers for language tasks. The model features a token embedding layer with a vocabulary of 32,000 and an embedding size of 4,096, providing substantial representational power and effectively managing varying input sizes. The core component is a series of layers with unique attention mechanisms, enhanced with low-rank adapters (LoRA)[18] to dynamically adjust attention patterns, reducing the need for comprehensive retraining. Dropout regularization and low-rank matrices improve the model's ability to adapt to data patterns.

We fine-tuned the LLaMa-2 model specifically for natural disasters, including floods, landslides, droughts, wildfires, heavy snow, typhoons, and earthquakes. Fine-tuning involves adjusting LLaMa-2's parameters to better reflect the specialized vocabulary and contextual nuances characteristic of disaster scenarios.

3. Images Using Metadata Generator and Text-to-Image (T2I) Model

Figure 3 illustrates the metadata generation inference process using the fine-tuned LLaMa-2 model. Once the keywords extracted from the CAP data are input into the model, the metadata is produced as output. This generated metadata not only categorizes but also enriches the data, ensuring that subsequent visual data generation processes are contextually aligned with the specifics of the emergency situations depicted in the textual data.

The pre-trained LLM model, based on LLaMA-2, generates disaster-related metadata matching our metadata template. Using input from the Emergency Common Alerting Protocol (CAP), the disaster type is identified and processed by the pre-trained metadata generator. This metadata is then fed into the Text-To-Image (T2I) generative model as a prompt, resulting in the final disaster scenario image.

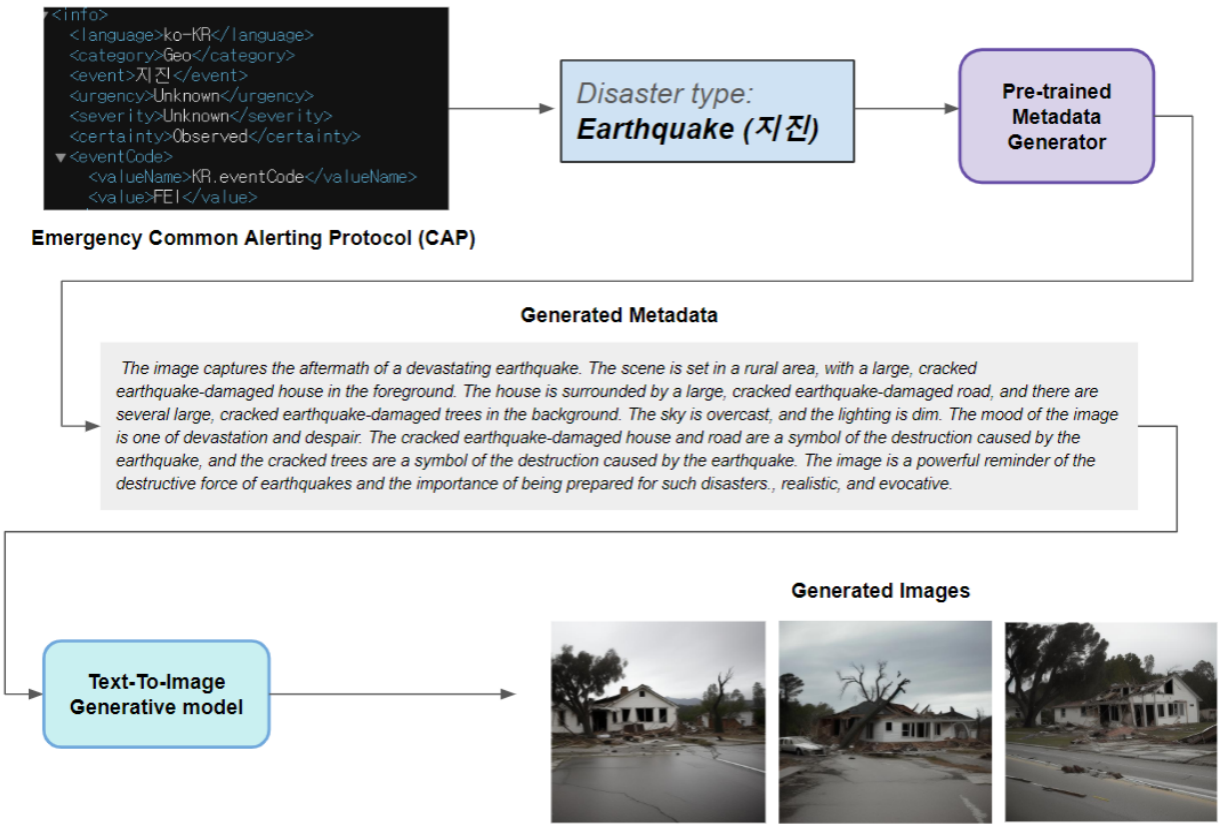

그림 3. 재난 시나리오 생성 진행 절차
Fig. 3. Workflow for Generating Disaster-Scenario

4. Producing 3D Cinemagraph: Applying Animation Effects to the Image

Our approach to 3D cinematography begins with an image input and a manually defined mask to control the animated regions. This framework is built on the pre-trained Stable Diffusion model. We utilize the Language Segment-Anything (LangSAM) model, which integrates instance segmentation with text prompts to generate masks for specific objects.

Once an image is generated from a text prompt using Stable Diffusion, it is passed to the Mask Generator Network, where the LangSAM model creates a mask that delineates areas for motion based on keywords related to fluid elements, such as water or river for fluid, and smoke or fire for wildfire disaster types. LangSAM analyzes both the text prompt and the generated image to identify and segment objects or areas for animation according to the description. This automated mask creation effectively differentiates between static and animated parts, enhancing the cinamgraph creation process. After the automatic mask generation, the mask and motion labels are produced as outputs and are passed to the Motion Estimation Network.

Animating a still image involves converting 2D data into a 3D structure and manipulating it sequentially. Creating a 3D scene starts by producing detailed depth maps from a single image to approximate the scene's geometry using the dense prediction transformer (DPT)[8]. The image is then transformed into a Layered Depth Image (LDI) format, segmenting the scene into layers by depth discontinuities

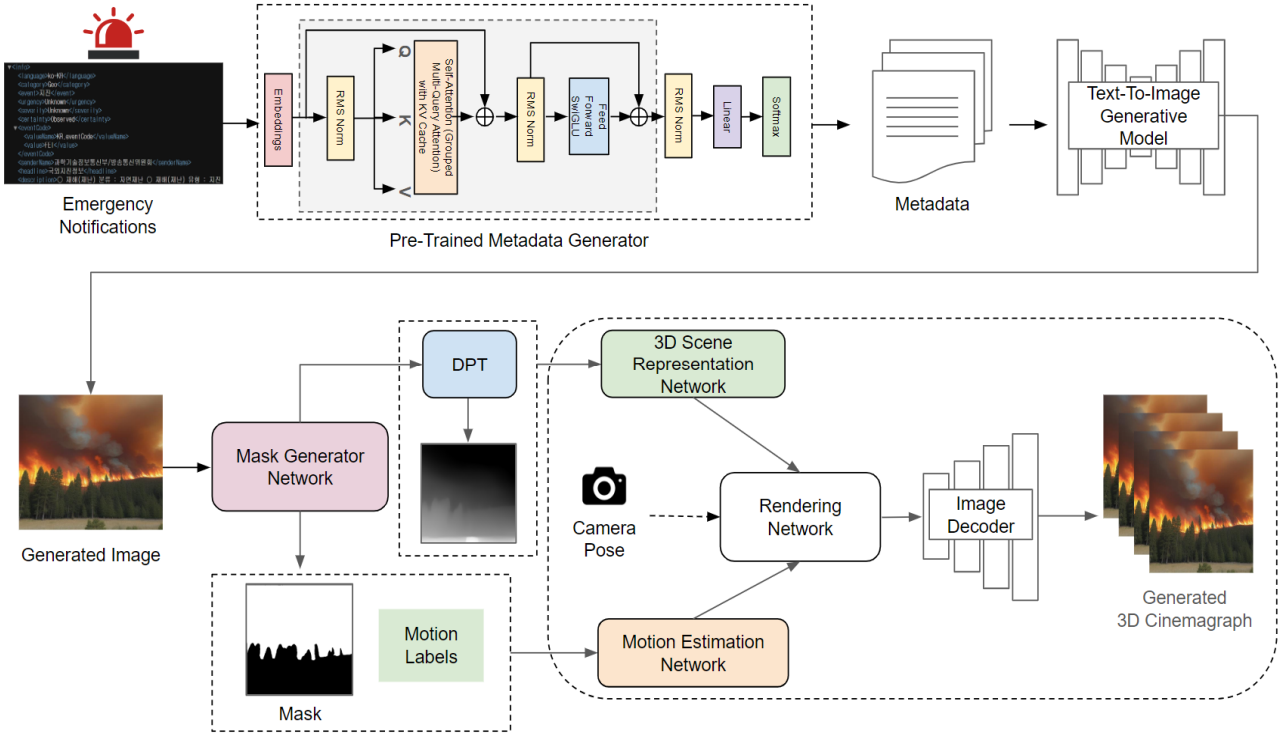

그림 4. 제안된 텍스트-투-3D 시네마그래프 프레임워크
Fig. 4. The proposed Text-to-3D Cinemagraph Framework

and filling hidden areas. Each layer is refined using a pre-trained model to inpaint obscured sections and extract 2D features, which are encoded into feature maps. These enhanced LDIs are projected into a 3D space, forming a feature point cloud. This cloud, with points that each have 3D coordinates and feature vectors, facilitates dynamic and realistic animations with depth and motion.

Concurrently, 2D optical flow is calculated to capture motion between video frames, which is transformed into a 3D space, forming RGB point clouds with scene flows detailing each point's movement. These RGB point clouds are animated over time according to the scene flow and projected back into various viewing angles. However, this can lead to gaps in the scene. To mitigate this, a 3D symmetric animation technique (3DSA) enhances continuity and fills gaps in the point cloud animation.

For a refined approach to animating static images, the Eulerian flow field is used. This method assumes a constant-velocity motion field, simplifying the capture of dynamic real-world movements. The Eulerian flow field defines the movement of pixels from one frame to another, calculating future positions using Euler integration to forecast the path of each pixel through multiple frames, creating a consistent displacement field.

An Image-To-Image translation network transforms an RGB image directly into an optical flow map, predicting pixel movements to mimic natural scene dynamics. This boosts the realism of the animation and addresses issues like distortion and occlusion.

Figure 5 displays the generated image, its corresponding depth map created by the dense prediction transformer (DPT), and the LangSAM-generated mask for applying motion in fluid areas, enhancing 3D cinemagraphs for disasters like wildfires and floods.

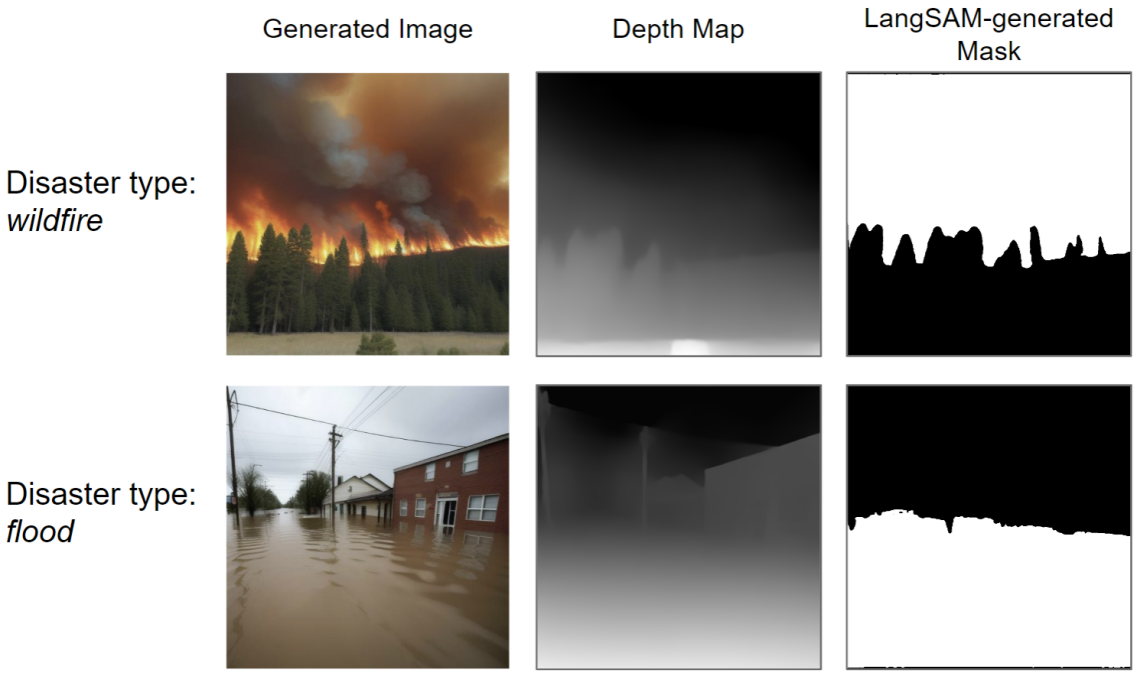

그림 5. 생성된 이미지, 3D 효과를 위한 Depth Map, 유동적인 영역(물, 불, 연기 등)의 모션 적용을 위한 LangSAM이 생성한 마스크
Fig. 5. Generated Image and Its Depth Map for 3D Effect, Alongside the LangSAM-Generated Mask for Applying Motion in Fluid Areas Such as Water, Fire, and Smoke

Ⅳ. Results

1. Text Generation

Figure 6 displays the outputs of a pretrained large language model (LLM) when provided with disaster-related keywords, covering all seven disaster types: flood, landslide, drought, wildfire, heavy snow, typhoon, and earthquake. The LLM generates detailed prompts specifying the mood, time of day, lighting conditions, and descriptions of each object in the image. Importantly, since the Text-To-Image model does not handle human-related images, the output prompts are modified to include the directive “Do not include any humans.”

For example, for the disaster type Earthquake, the generated prompt by the metadata generator was: “The image captures the aftermath of a devastating earthquake. The scene is set in a rural area, with a large, cracked earthquake-damaged house in the foreground. The house is surrounded by a large, cracked earthquake-damaged road, and there are several large, cracked earthquake-damaged trees in the background. The sky is overcast, and the lighting is dim. The mood of the image is…”. This detailed metadata enables the T2I generative model to accurately produce the disaster scenario image without unexpected results. However, extensive testing revealed that our model occasionally produces unexpected tokens, such as “0” or “α”, at the end of the prompt. Though these tokens do not affect the overall prompt or the image generation process.

Overall, the generated text is of high quality, accurately depicting the disaster scenarios, and is produced quickly and efficiently. However, through multiple test inferences, we noted that the prompts tended to be similar, showing less diversity. This consistency, while resulting in less variety, is advantageous for domain-specific, real-time scenario services.

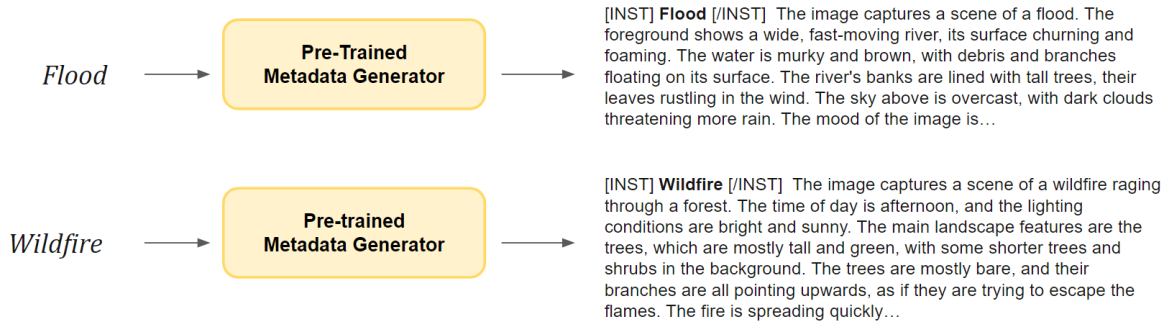

그림 6. 파인튜닝된 LLM의 생성 결과
Fig. 6. Generated Results from the Fine-Tuned Large Language Model (LLM)

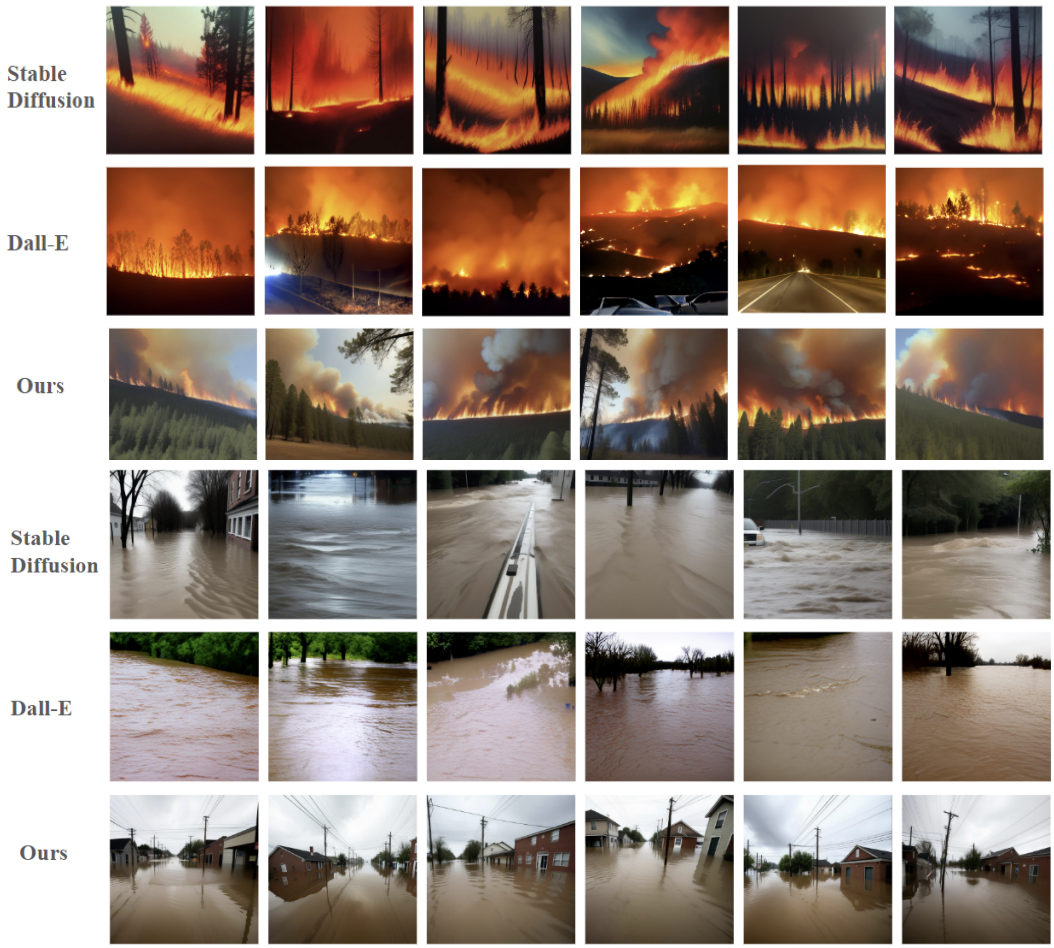

그림 7. ‘wildfire’(위)와 ‘flood’(아래) 키워드를 사용하여 생성된 이미지들
Fig. 7. Images generated using the keyword 'wildfire' (above), and ‘flood’ (below)

2. Image Generation

The top part of Figure 7 shows images generated with the keyword ‘wildfire’. The images generated using Stable Diffusion, capture the general idea of a wildfire but lack realism and accuracy. Dall-E's[7] images exhibit significant variation, sometimes including roads, trees, or cars. In contrast, our model accurately depicts the wildfire context, consistently producing similar images for the same prompt. The trees and smoke have consistent shapes, and the overall image maintains a uniform color and mood, free from extraneous objects.

The bottom part of Figure 7 shows images generated with the keyword ‘flood’. The Stable Diffusion images sometimes include random elements like roads and cars, while Dall-E’s images show less variation but often inaccurately depict floods, looking more like lakes. However, our model accurately captures the flood context with minimal variation, largely due to the use of metadata, which anchors the images within the targeted domain.

Table 1 presents the evaluation performance of DALL-E, Stable Diffusion, and our model using FID Score, CLIP Score, LPIPS Score, and SSIM Score. Disaster types include drought, earthquake, heavy snow, landslide, wildfire, flood, and typhoon. As highlighted in Table 1, our model outperforms DALL-E and Stable Diffusion in key metrics across various disaster scenarios, particularly in FID and SSIM scores. The consistently higher SSIM scores ach-

표 1. 재난 시나리오별 T2I 모델 성능 비교
Table 1. Performance Comparison of T2I Models across Disaster Scenarios

Disaster Type	T2I Model	FID Score (↓)	CLIP Score (↑)	LPIPS Score (↓)	SSIM Score (↑)
drought	Dall-E	58.9621	0.302246	0.563841	0.096903283
	Stable Diffusion	47.8415	**0.305664**	**0.505445**	0.113314714
	Ours	**45.24**	0.291504	0.53953	**0.210081838**
earthquake	Dall-E	48.941	0.269775	**0.597162**	0.124740098
	Stable Diffusion	**45.3205**	**0.276367**	0.616291	0.243389424
	Ours	51.5152	0.256592	0.629937	**0.300104171**
heavy snow	Dall-E	42.8278	0.285156	**0.59649**	0.216841497
	Stable Diffusion	47.1569	0.288086	0.664955	**0.555356296**
	Ours	**39.8914**	**0.298584**	0.636955	0.508998988
landslide	Dall-E	46.9144	0.305908	**0.613208**	0.139957937
	Stable Diffusion	57.1535	0.274902	0.68916	0.293309681
	Ours	**38.3474**	**0.308105**	0.642025	**0.445461847**
wildfire	Dall-E	33.5569	0.30127	0.64909	0.427983221
	Stable Diffusion	36.2084	0.296387	0.658495	0.369551316
	Ours	**27.692**	**0.304199**	**0.622422**	**0.463339194**
flood	Dall-E	47.9652	**0.27832**	**0.613305**	0.354063124
	Stable Diffusion	44.0885	0.276855	0.639292	0.321260773
	Ours	**42.8126**	0.271484	0.627062	**0.495805458**
typhoon	Dall-E	58.4703	0.270752	**0.59471**	0.247385279
	Stable Diffusion	61.9068	0.264893	0.617294	0.177115488
	Ours	**51.6765**	**0.286621**	0.632521	**0.297621737**

ieved by our model indicate its superior ability to maintain structural similarity and accuracy in the generated images. These results underscore the effectiveness of our approach in producing high-quality, structurally accurate images that closely align with the provided textual descriptions, ensuring a reliable and consistent representation of disaster scenarios.

3. Producing 3D Cinemagraphs

Figure 8 showcases the final generated 3D cinemagraph with 7 frames for the disaster types 'Wildfire' and 'Flood'.

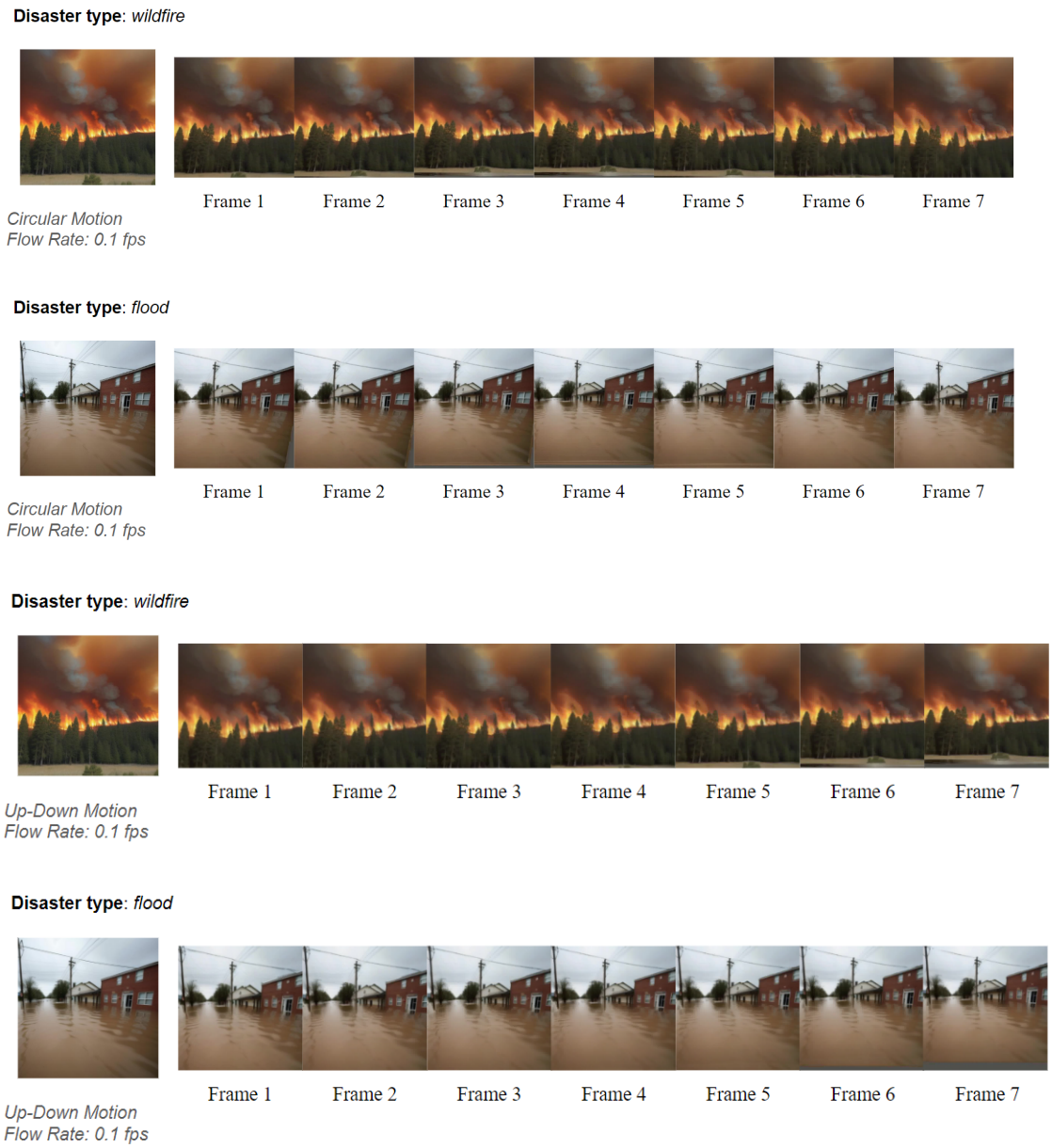

그림 8. 홍수와 산불의 최종 생성된 3D 시네마그래프
Fig. 8. Final Generated 3D Cinemagraph of Flood and Wildfire

The top two frames illustrate circular motion, while the bottom two exhibit an up-and-down camera motion effect. These animations are smooth and clear, greatly enhancing the overall image quality.

Ⅴ. Conclusion

In this thesis, we developed a novel method for generating cinemagraphs from textual prompts by integrating a fine-tuned large language model (LLM), a Text-To-Image (T2I) model, and an automated flow-generating network. Unlike conventional methods that produce video content from text—often requiring high quality and extensive processing—our approach simplifies the process by manipulating images directly and applying dynamic animation effects.

Our integrated framework has demonstrated significant effectiveness, particularly in disaster scenarios, by consistently producing high-quality images. The use of a Metadata Generator to create domain-specific metadata ensures that the generated images are contextually relevant and structurally accurate. This approach not only reduces variability but also enhances the precision of the generated visuals.

As evidenced in Table 1, our model outperforms existing methods like DALL-E and Stable Diffusion across various disaster scenarios, particularly in key metrics such as FID and SSIM scores. The consistently higher SSIM scores indicate our model's superior ability to maintain structural similarity and accuracy in the generated images. These results validate the effectiveness of our approach in producing high-quality, structurally accurate images that closely align with the provided textual descriptions, ensuring reliable and consistent representations of disaster scenarios.

We acknowledge limitations in our research, particularly the emergence of gray gaps and holes due to excessive camera movement. To address this, we suggest using video outpainting technology on 3D meshes as a potential solution.

Our integrated framework, combining an optimized prompt generator with the T2I model, has proven effective in consistently producing high-quality images. By employing a single-image flow estimator, our framework ensures relevance and efficiency, making it highly suitable for contexts benefiting from stable, high-quality automated image generation.

Moreover, while the model generally succeeds in generating accurate 3D cinemagraphs, challenges may arise when fluid elements like water and fire coexist in the same scene. For example, in images depicting both floods and wildfires, the LangSAM model may incorrectly animate overlapping areas, leading to unrealistic results. However, our framework minimizes such occurrences by clearly distinguishing between disaster types from the outset, significantly reducing the likelihood of these issues.

This framework can be generalized to a wide range of domains requiring precise information and automated generation capabilities, extending beyond disaster scenarios. Its applicability spans sectors such as entertainment and media, where it can rapidly create engaging content, and education, where customized images can significantly enhance learning experiences.

References

[1] Li, X., Cao, Z., Sun, H., Zhang, J., Xian, K., & Lin, G. (2023). 3d cinemagraphy from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4595-4605).
doi: https://doi.org/10.48550/arXiv.2303.05724

[2] Shih, M. L., Su, S. Y., Kopf, J., & Huang, J. B. (2020). 3d photography using context-aware layered depth inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8028-8038).
doi: https://doi.org/10.48550/arXiv.2004.0472

[3] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

[4] Li, J., Li, D., Xiong, C., & Hoi, S. (2022, June). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International conference on machine

learning (pp. 12888-12900). PMLR.

[5] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4015-4026).
doi: https://doi.org/10.1109/ICCV51070.2023.00371

[6] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).
doi: https://doi.org/10.48550/arXiv.2112.10752

[7] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021, July). Zero-shot text-to-image generation. In International conference on machine learning (pp. 8821-8831). Pmlr.
doi: https://doi.org/10.48550/arXiv.2102.12092

[8] Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision transformers for dense prediction. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 12179-12188).
doi: https://doi.org/10.1109/ICCV48922.2021.01196

[9] Holynski, A., Curless, B. L., Seitz, S. M., & Szeliski, R. (2021). Animating pictures with eulerian motion fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5810-5819).
doi: https://doi.org/10.48550/arXiv.2011.15128

[10] Mahapatra, A., Siarohin, A., Lee, H. Y., Tulyakov, S., & Zhu, J. Y. (2023). Synthesizing Artistic Cinemagraphs from Text.
doi: https://doi.org/10.48550/arXiv.2307.03190

[11] Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., ... & Sun, L. (2024). Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models.
doi: https://doi.org/10.48550/arXiv.2402.17177

[12] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., & Fleet, D. J. (2022). Video diffusion models. Advances in Neural Information Processing Systems, 35, 8633-8646.
doi: https://doi.org/10.48550/arXiv.2204.0345

[13] Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., ... & Taigman, Y. (2022). Make-a-video: Text-to-video generation without text-video data.
doi: https://doi.org/10.48550/arXiv.2209.14792

[14] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models.
doi: https://doi.org/10.48550/arXiv.2302.13971

[15] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models.
doi: https://doi.org/10.48550/arXiv.2307.09288

[16] Ruthotto, L., & Haber, E. (2021). An introduction to deep generative modeling. GAMM Mitteilungen, 44(2), e202100008.
doi: https://doi.org/10.48550/arXiv.2103.05180

[17] Mahapatra, A., & Kulkarni, K. (2022). Controllable animation of fluid elements in still images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3667-3676).
doi: https://doi.org/10.48550/arXiv.2112.03051

[18] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models.
doi: https://doi.org/10.48550/arXiv.2106.09685

[19] Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., ... & Salimans, T. (2022). Imagen video: High definition video generation with diffusion models.
doi: https://doi.org/10.48550/arXiv.2210.02303

[20] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., & Fleet, D. J. (2022). Video diffusion models. Advances in Neural Information Processing Systems, 35, 8633-8646.
doi: https://doi.org/10.48550/arXiv.2204.0345

[21] Croitoru, F. A., Hondru, V., Ionescu, R. T., & Shah, M. (2023). Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence.
doi: https://doi.org/10.48550/arXiv.2209.04747

[22] Oussidi, A., & Elhassouny, A. (2018, April). Deep generative models: Survey. In 2018 International conference on intelligent systems and computer vision (ISCV) (pp. 1-8). IEEE.
doi: https://doi.org/10.1109/ISACV.2018.8354080

[23] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models.
doi: https://doi.org/10.48550/arXiv.2303.18223

[24] Li, J., Tang, T., Zhao, W. X., Nie, J. Y., & Wen, J. R. (2024). Pre-trained language models for text generation: A survey. ACM Computing Surveys, 56(9), 1-39.
doi: https://doi.org/10.48550/arXiv.2201.05273

[25] Li, X., Lai, Z., Xu, L., Guo, J., Cao, L., Zhang, S., Dai, B., & Ji, R. (2024). Dual3D: Efficient and Consistent Text-to-3D Generation with Dual-mode Multi-view Latent Diffusion.
doi: https://doi.org/10.48550/arXiv.2405.09874

[26] Raj, A., Kaza, S., Poole, B., Niemeyer, M., Ruiz, N., Mildenhall, B., Zada, S., Aberman, K., Rubinstein, M., Barron, J., & Li, Y. (2023). Dreambooth3d: Subject-driven text-to-3d generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2349-2359).
doi: https://doi.org/10.48550/arXiv.2303.13508

[27] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K. (2023). Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 22500-22510).
doi: https://doi.org/10.48550/arXiv.2208.12242

[28] Poole, B., Jain, A., Barron, J. T., & Mildenhall, B. (2022). DreamFusion: Text-to-3D using 2D Diffusion.
doi: https://doi.org/10.48550/arXiv.2209.14988

저 자 소 개

원 루 빈

- 2022년 11월 : University of Alberta Computer Science 학사
- 2023년 3월 ~ 현재 : 과학기술연합대학원대학교(UST) 정보통신공학 석사과정
- ORCID : https://orcid.org/0009-0007-2664-0507
- 주관심분야 : Deep Generative model, Text-to-Image generation, Diffusion model

최 민 지

- 2022년 8월 : 홍익대학교 과학기술대학 전자·전기공학과 학사
- 2022년 9월 ~ 현재 : 과학기술연합대학원대학교(UST) 정보통신공학 석사과정
- ORCID : https://orcid.org/0009-0007-5851-6535
- 주관심분야 : 컴퓨터 비전, Deep Generative model, Text-to-Image generation

최 지 훈

- 1999년 2월 : 경희대학교 전자공학과 학사
- 2001년 2월 : 경희대학교 전자공학과 석사
- 2001년 3월 ~ 현재 : 한국전자통신연구원 미디어연구본부 미디어지능화연구실 책임연구원
- ORCID : https://orcid.org/0000-0002-3402-1921
- 주관심분야 : UHD 방송 기술, 재난정보미디어 서비스, AI 미디어 처리 기술

배 병 준

- 2006년 8월 : 경북대학교 전자공학과 박사
- 1997년 2월 ~ 2000년 10월 : ㈜엘지전자 주임연구원
- 2000년 11월 ~ 현재 : 한국전자통신연구원 미디어연구본부 미디어지능화연구실 책임연구원
- 2012년 3월 ~ 현재 : 과학기술연합대학원대학교(UST) ETRI스쿨 정보통신공학 책임교수
- 2024년 1월 ~ 현재 : 한국정보통신기술협회(TTA) PG802(지상파방송) 부의장
- 2024년 1월 ~ 현재 : 한국방송통신전파진흥원(KCA) 비상임이사
- ORCID : https://orcid.org/0000-0002-0872-325X
- 주관심분야 : UHD 방송 기술, 재난정보미디어 서비스, AI 미디어 처리 기술