# Segmentation-Based Masked Sampling for text-to-animated image synthesis in disaster scenarios

Ru-Bin Won
Information and Communication Engineering
University of Science and Technology, Korea
Daejeon, Republic of Korea
rubrub@etri.re.kr

Minji Choi
Information and Communication Engineering
University of Science and Technology, Korea
Daejeon, Republic of Korea
cmj@etri.re.kr

Ji Hoon Choi
Media Intellectualization Research Lab
ETRI (Electronics and Telecommunuications Research Institute)
Daejeon, Republic of Korea
cjh@etri.re.kr

Byungjun Bae
Media Intellectualization Research Lab
ETRI (Electronics and Telecommunications Research Institute)
Daejeon, Republic of Korea
1080i@etri.re.kr

***Abstract*— Generative AI has demonstrated significant capabilities in text-to-video synthesis, using advanced models characterized by extensive parameters. Given that current disaster alert services are mostly text-based and can be less accessible to many, this paper proposes a new approach. We aim to provide animated disaster images that precisely mirrors text prompts, thereby enhancing the efficiency and accessibility of disaster alerts. Our methodology combines the strengths of segmentation models and pre-trained Vision Transformer (ViT) mechanisms. By using a unique image selection based on CLIPScore and processing it with the CLIPSeg segmentation model, we generate an animated representation of disaster scenarios. This offers a simple, fast, and effective solution to the challenges of the current disaster alert systems.**

***Keywords—text-to-video synthesis, disaster alert system, Vision Transformer (ViT), masked sampling, image segmentation, text-to-animated image generative model***

## I. INTRODUCTION

In recent years, Generative AI has emerged as a pivotal advancement within the realm of text-to-video synthesis. Such systems often use complex models with numerous parameters, leveraging space and time super-resolution techniques to produce high-quality extended videos[1, 2]. However, upon examination of their practical implementation, especially in the critical context of disaster alert systems, the complexity of these models presents discernible challenges for efficient deployment.

Today's disaster alert services primarily use text formats. While this might seem efficient, it poses considerable accessibility barriers. Vulnerable groups, including the elderly, children, individuals with disabilities, and non-native speakers, might find it challenging to swiftly understand and act on these text alerts. Integrating these textual alerts into video format using conventional generative models could increase the time lag, which is unacceptable in urgent scenarios where timeliness is critical.

Given these constraints, it becomes crucial to reimagine a video generative model that operates efficiently and simply. This study addresses these challenges. Rather than aligning with broader research efforts that primarily focus on video generation, we zero in on crafting animated images. These animations not only demand less memory but are also quicker to produce. Through the utilization of segmentation models[3] and the incorporation of the advanced functionalities offered by pre-trained Vision Transformer (ViT) mechanisms[4], our approach presents a unique perspective.

Initiating the process starts with creating a limited image set. From this, the image that offers the closest representation to the provided textual description, as determined by its CLIPScore[5], is selected for further processing. The chosen image is then segmented using the CLIPSeg model[3], resulting in a binary mask that filters out static elements. Combining this mask with a randomized one and using ViT capabilities, we craft animations mirroring the original disaster text prompt.

Our key contributions can be outlined as follows:

- We employed CLIPScore to select the generated image that most accurately aligns with the given textual description—a crucial component for effective disaster alert systems.
- We propose a Segmentation-Based Masked Sampling technique. This method utilizes a combined mask, which integrates a binary mask (excluding objects identified by the CLIPSeg segmentation) and a randomized mask.

979-8-3503-1327-7/23/$31.00 ©2023 IEEE

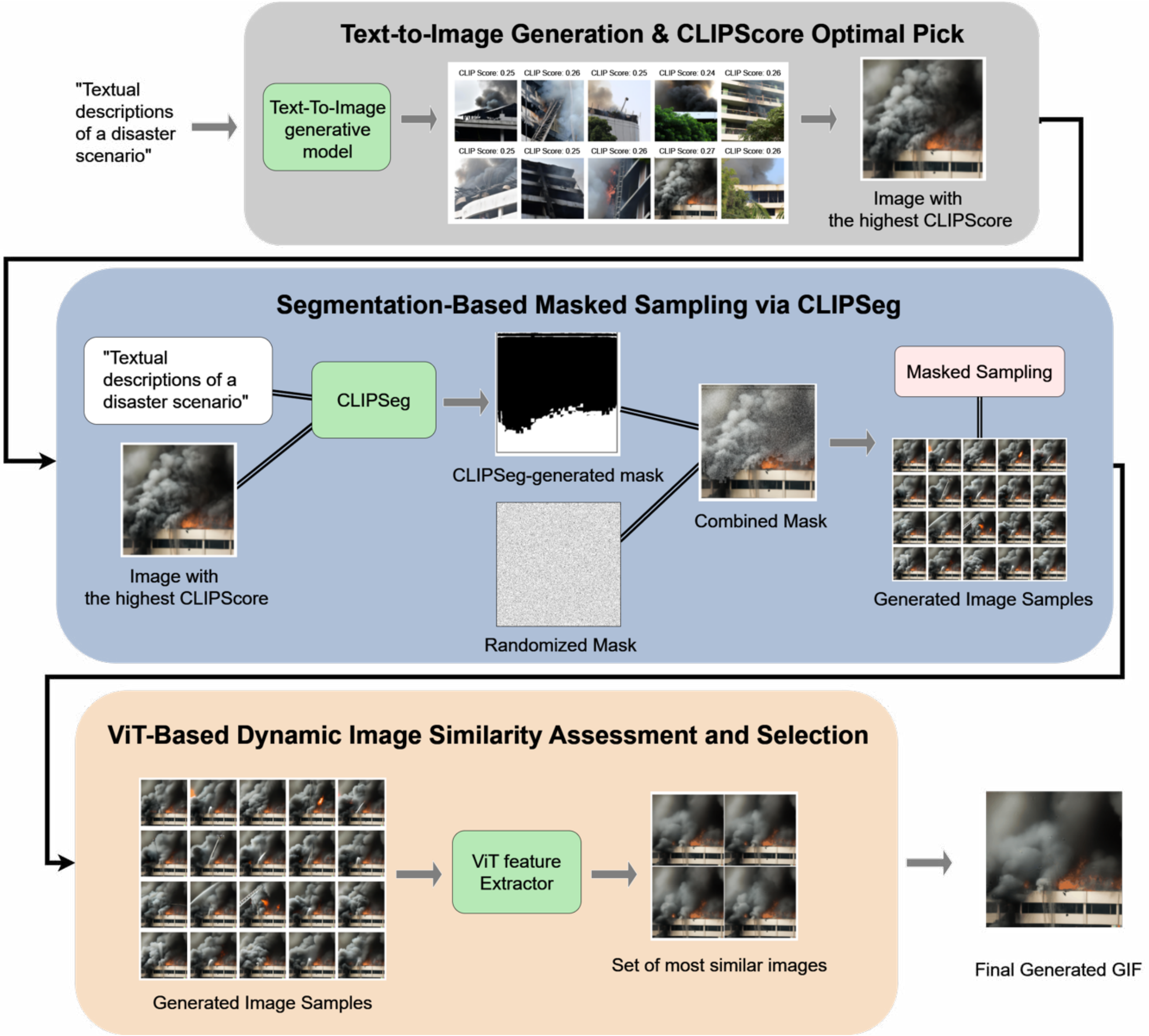
Text-to-Image Generation & CLIPScore Optimal Pick
"Textual descriptions of a disaster scenario"
Text-To-Image generative model
CLIP Score: 0.25
CLIP Score: 0.26
CLIP Score: 0.25
CLIP Score: 0.24
CLIP Score: 0.26
CLIP Score: 0.25
CLIP Score: 0.25
CLIP Score: 0.26
CLIP Score: 0.27
CLIP Score: 0.26
Image with the highest CLIPScore
Segmentation-Based Masked Sampling via CLIPSeg
"Textual descriptions of a disaster scenario"
CLIPSeg
CLIPSeg-generated mask
Masked Sampling
Image with the highest CLIPScore
Combined Mask
Generated Image Samples
Randomized Mask
ViT-Based Dynamic Image Similarity Assessment and Selection
ViT feature Extractor
Set of most similar images
Generated Image Samples
Final Generated GIF

Fig. 1. Proposed Method for Process Steps in the Model

## II. BACKGROUND KNOWLEDGE

### A. CLIPScore

CLIPScore[5] offers a reference-independent metric for image captioning that mirrors human evaluations. CLIPScore provides a distinctive approach by sidestepping the often cumbersome and cost-intensive task of collecting reference captions, unlike its counterparts. Instead, it employs the CLIP model to process both the image and the proposed caption, subsequently determining a similarity metric between the two. The metric is defined as:

$$\mathrm{CLIP}-\mathrm{S}(\boldsymbol{I},\boldsymbol{C}) = w * \max\left(\cos(\mathbf{E}i,\mathbf{E}c),0\right) \qquad (1)$$

This represents the cosine similarity between the visual CLIP embedding, $\mathbf{E}i$, associated with image $i$, and the textual CLIP embedding $\mathbf{E}c$, corresponding to caption $c$. The resultant score ranges between 0 and 100, with a score proximate to 100 indicating superior performance.

### B. CLIPSeg Model for Segmentation-Based Mask

In the realm of image segmentation, the CLIPSeg model[3] extends the capabilities of the CLIP transformer. It is innovatively designed for both zero-shot and one-shot segmentation tasks, leveraging a lightweight, transformer-based decoder. Unique to this model is its ability to segment images based on either a text query or a reference image, adeptly addressing three primary segmentation tasks. It makes use of the joint text-visual embedding space of CLIP for conditioning, allowing for seamless processing of both text and visual cues. The architecture, purely transformer-based, integrates U-Net-inspired skip connections to the CLIP encoder, optimizing the decoder to be both effective and parameter-efficient. In its operation, CLIPSeg generates a binary segmentation mask from the provided image and text.

Importantly, this paper introduces the term “segmentation-based mask” to denote the mask produced through the CLIPSeg image segmentation process. This mask is subsequently merged with a randomized mask that consistently masks 15% of the original image using a logical OR operation. The resultant mask preserves the essential areas of the initial mask while introducing variability in other areas. Using this refined mask as a reference, the model produces several image samples, a process we term "Segmentation-Based Masked Sampling."

### *C. ViT image transformer feature extraction*

The Vision Transformer (ViT)[4] is a transformer-based architecture specifically tailored for image feature extraction. It segments input images into linear 2D patches, processed by an encoder reminiscent of transformers used in natural language processing. While both Convolutional Neural Networks(CNNs) and ViTs are prominent in feature extraction, ViTs have shown superior performance, especially when pre-trained on large datasets. Their self-attention mechanism, which captures intricate image patterns, also offers greater flexibility and interpretability compared to CNNs.

For our study, we utilized the vit-base-patch16-224 variant of ViT, chosen for its computational efficiency and simpler design in comparison to more massive models like ViT-Large or ViT-Huge. The "Base" denotes its 12 transformer layers with 768-dimensional hidden states, while "Patch16" and "224" refer to its image segmentation and size specifications, respectively.

A key strength of ViT in the context of image similarity detection lies in its global perspective. Unlike traditional methods that focus on localized patterns, ViTs capture the whole image landscape, ensuring a more contextually-aware feature extraction. Such features provide deeper semantic insights, enabling nuanced image comparisons that extend beyond basic pixel similarities. This capability enhances the accuracy of detecting image similarities, capturing details that might be missed by conventional techniques.

## III. PROPOSED METHOD

Our objective is to create an efficient model for fast disaster alerts that generates animated image from text, ensuring the information is both accessible and comprehensible to a broad audience. We tried to maximize the benefit from the pre-trained model such as CLIPSeg and ViT. In the initial stages, we carry out masked sampling using a composite mask — a combination of the mask generated from the CLIPSeg segmentation process and a randomized mask with a 15% masking ratio. Following this, we pinpoint the most analogous images using ViT feature extraction and dynamic programming.

To demonstrate our concept of text-to-animated image generation, our experimentation concentrated on creating an animated image depicting a fire disaster, anchored by the textual prompt, "*A fire has broken out in the building*."

### *A. Image Generation via Text-to-Image Generative Model*

For a given textual description of a disaster scenario as input, we utilize the Text-to-Image (T2I) generative model to produce a set of images. In our experimental setup, we limited the number of images to a minimum of 10 ($10 \leq x$), as our findings indicated that this quantity sufficed in capturing a representative reflection of the textual description. Our experimentation was conducted using the Stable Diffusion[6] and DALL·E[7] models. However, it is worth highlighting that this system is designed to incorporate any T2I generative model seamlessly, ensuring flexibility in model selection.

### *B. Selection of the Image with the highest CLIPScore*

Given the critical nature of disaster scenarios, ensuring that the visual representation closely aligns with the textual description is paramount. To ensure this precision, we employ the CLIPScore metric. From the set of images generated in the preceding step, we select the one that registers the highest CLIPScore, as this signifies the most accurate reflection of the textual details. For the textual cue, "*A fire has broken out in the building*," images produced by DALL·E[7] achieved an average CLIPScore of 25.5%, whereas those from Stable Diffusion[6] reached 25.6%. From these results, we deduced that the choice of image generation model doesn't significantly impact our system's efficacy.

### *C. Segmentation-Based Masked Sampling*

Utilizing the image that secured the highest CLIPScore and its corresponding textual description, we identify the core subject of the description. All other components in the image are then masked out. This masking is executed via CLIPSeg[3], a model capable of generating image segmentations based on the given textual prompt. For instance, given the description "*A fire has broken out in the building*," the language model identifies the subject—namely, "fire"—and introduces it to CLIPSeg[3] in conjunction with the original disaster image. This interaction yields a binary segmentation mask where static entities like the building are rendered opaque, while dynamic facets such as the fire are kept transparent.

Following this, the CLIPSeg-generated mask is integrated with a randomized mask, maintaining a consistent 15% masking ratio across the initial image. Our experimentation determined that a masking ratio of 15% strikes an optimal balance between image diversity and accuracy. The final mask aims to uphold the salient portions of the initial image using the CLIPSeg-generated mask while infusing select randomness into remaining areas. Due to its dependency on segmentation-centric masking for sampling, this method is coined Segmentation-Based Masked Sampling.

The culmination of this process is the generation of multiple images that diverge from the original, while retaining specific still elements. In our example, features like buildings and trees are cloaked, leaving only fire variations across all image samples.

### *D. Leveraging ViT feature extraction to dynamically select and merge the most analogous images*

After generating a variety of image samples using the Segmentation-Based Masked Sampling technique, our aim is to select and group the most analogous images to produce a natural-looking animated image. For this purpose, we employ the Vision Transformer (ViT)[4] model, particularly its 'vit_base_patch16_224' configuration, which we refer to as ViT-B in this context. Upon loading our pre-trained ViT-B model, we reconfigure it to serve as a feature extractor by discarding the classification head and replacing it with an identity function. This modification allows us to obtain a compact feature representation of input images. Additionally, we define a preprocessing pipeline that entails image resizing, normalization, and tensor transformation.

For each sample, the ViT-B extracts features used to compute pairwise Euclidean distances. This produces a distance matrix indicating the similarity between samples. To determine the most similar images, we implement a dynamic programming strategy that seeks image combinations with the lowest aggregate pairwise distance. This algorithm evaluates the image combinations, iteratively determining their cumulative distances and utilizing memoization to optimize the computation and eventually find the set of most similar images. This procedure ensures that the most analogous images, based on the feature representations obtained from ViT-B, are selected for the subsequent integration process.

To conclude the process, our model employs a straightforward linear interpolation method to produce a coherent animated representation, specifically in GIF format.

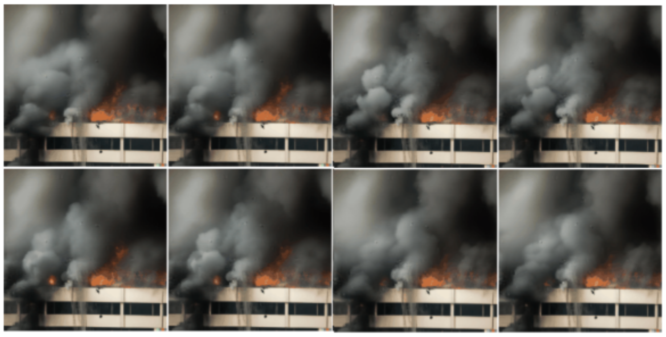

Fig. 2. Frame sequence of the generated animated image

## IV. RESULTS

From the given textual prompt, "*A fire has broken out in the building*," the Text-To-Image generative model produced 10 images. Of these, the one with the highest CLIPScore was selected for further processing. The chosen image accurately encapsulated the information from the text, with both '*fire*' and '*building*' elements clearly represented, highlighting the efficacy of the CLIP model.

Subsequently, the segmentation provided by CLIPSeg was commendably precise, distinctly demarcating the '*fire*' region within the image. A composite mask, with a masking ratio of 15%, was then passed through our mask sampling module, which yielded high-quality, varied image samples. From the 20 generated samples, variations in the fire and smoke regions were evident. Employing the ViT feature extractor, we identified 5 images with the shortest feature distances, indicating high similarity among them. These images were then seamlessly merged using linear interpolation.

The resulting animated image was both dynamic and cohesive. The transitions between frames were fluid, providing a vivid depiction of the evolving fire scenario. The nuanced animations captured subtle shifts in the fire and smoke dynamics, underscoring the effectiveness of selecting highly similar images for the animation process. Moreover, the final animated representation retained the essence of the original textual prompt, ensuring that the viewer could easily discern the narrative of a building engulfed in flames.

## V. CONCLUSION

This study proposes a novel approach to address the limitations of current text-based disaster alert systems, emphasizing accessibility and efficiency. By merging the potentials of generative AI, segmentation models, and the pre-trained Vision Transformer (ViT) mechanisms, we have crafted a methodology to convert textual descriptions of disaster scenarios into dynamic animated images. Our unique application of CLIPScore ensures the precision of image selection in alignment with the provided text, while our Segmentation-Based Masked Sampling, anchored by the CLIPSeg model, introduces variability essential for animation. The subsequent use of the ViT model facilitates the selection of the most analogous images, ensuring fluidity in the resulting animation.

In practice, our approach effectively translated the textual prompt "*A fire has broken out in the building*" into a coherent and vivid animated depiction of the disaster, offering a testament to the methodology's efficacy. Such an integration of text-to-animated-image generation promises to revolutionize disaster alert systems, making them more intuitive and accessible, especially for vulnerable populations who might find text-based alerts challenging to interpret. This approach not only champions rapid information dissemination in critical scenarios but also offers a blueprint for further advancements in text-to-video synthesis.

## ACKNOWLEDGMENT

"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)"

## REFERENCES

[1] S. Yin, C. Wu, H. Yang, J. Wang, X. Wang, M. Ni*, et al*., "Nuwa-XL: Diffusion over diffusion for extremely long video generation," *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 2023. doi:10.18653/v1/2023.acl-long.73.

[2] R. Yang, P. Srivastava, and S. Mandt, "Diffusion Probabilistic Modeling for Video Generation," *arXiv preprint arXiv:2203.09481*, 2022.

[3] T. Luddecke and A. Ecker, "Image segmentation using text and image prompts," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/cvpr52688.2022.00695

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, *et al.,* "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, 2020.

[5] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A reference-free evaluation metric for image captioning," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. doi:10.18653/v1/2021.emnlp-main.595

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with Latent Diffusion Models," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* doi:10.1109/cvpr52688.2022.01042

[7] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, *et al.*"Zero-Shot Text-to-Image Generation," *arXiv preprint arXiv:2102.12092*, 2021.