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

Rubin Won
UST-ETRI
MS Student
rubrub@etri.re.kr
December 20th, 2023

# Contents

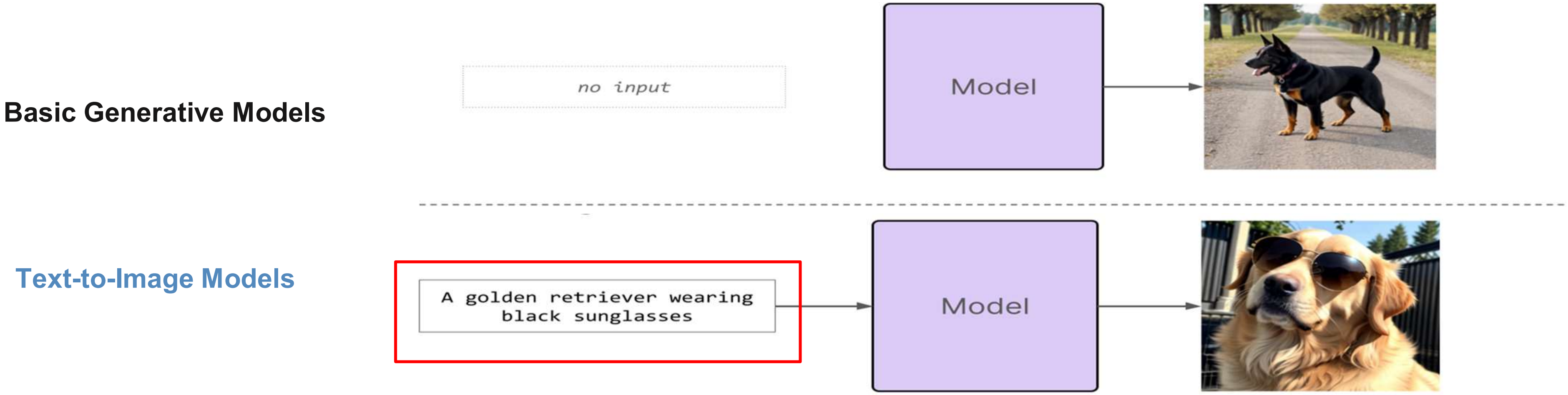
- Recap
- More Background Knowledge
- Proposed Method
- Results

# Recap

## Introduction & Background - **What is Text-to-Image Model?**

➢ **Text-to-Image (T2I) synthesis** leverages Generative AI to produce images based on **textual descriptions**.

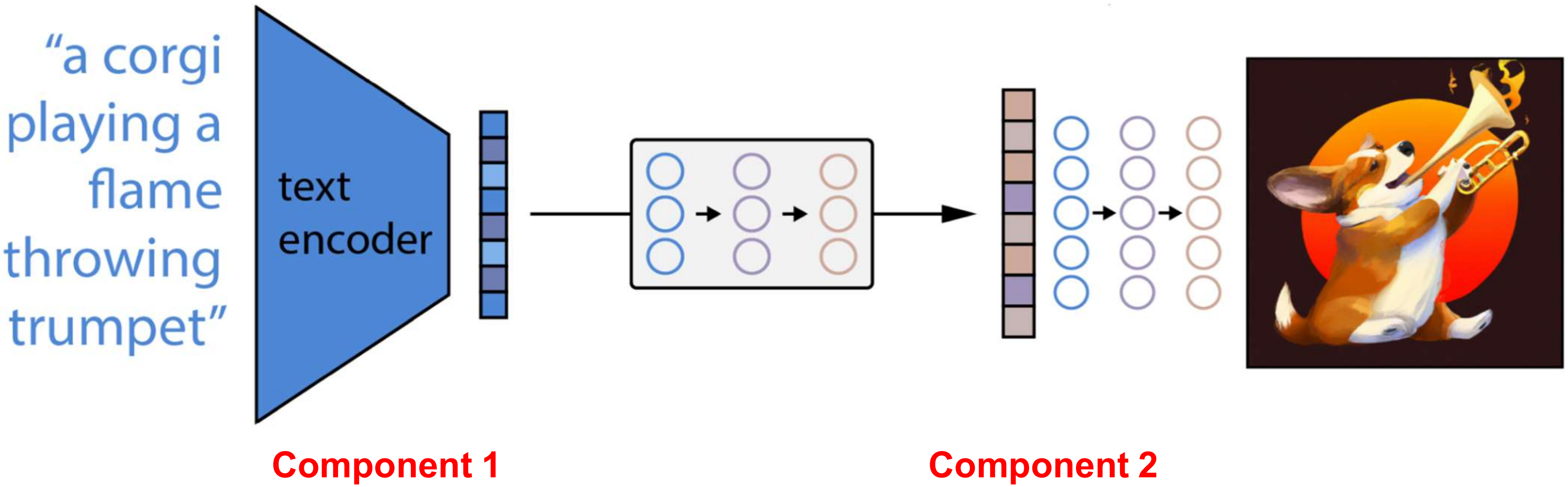
- Text-to-Image models use **a textual description to control** the image generation process in order to generate images that correspond to the description.



## Introduction & Background - **High Level look of the T2I process**

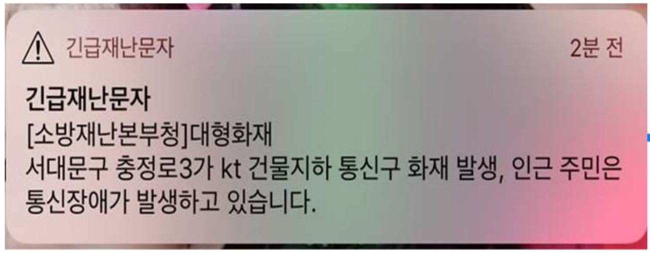
- **Component 1.**
  - A **textual encoder** that maps the text to a vector which captures the meaning of the text

- **Component 2.**
  - A **decoder model** that decodes this "meaning vector" into an image



## Motivation / Research Goals - **T2I and T2V applied on Disaster Alert System**

- ➢ Current emergency disaster alerts we receive during disaster scenarios are text-based and must be comprehended solely through text understanding.




**Text-based disaster alert**


**visual information for vulnerable populations**

# 01 Motivation / Research Goals - **Problems with current T2V models**

➢ The ongoing Text-To-Video research is focused on generating **high-quality, long** videos that include dazzling animation effects.

| Method | Parameters (Billion) | | | | | | | Speed (s) |
|---|---|---|---|---|---|---|---|---|
| | T2V Core | Auto Encoder | Text Encoder | Prior Model | Super Resolution | Frame Interpolation | Overall | |
| CogVideo [15] | 7.7 | 0.10 | – | – | – | 7.7 | 15.5 | 434.53 |
| Make-A-Video [31] | 3.1 | – | 0.12 | 1.3 | $1.4 + 0.7$ | 3.1 | 9.72 | – |
| Imagen Video [11] | 5.6 | – | 4.6 | – | $1.2 + 1.4 + 0.34$ | $1.7 + 0.78 + 0.63$ | 16.25 | – |

→ However, this approach results in **high spatial and temporal complexity.**

# More Background Knowledge

# More Background Knowledge - **CLIPScore**

**What is CLIPScore?**

- ➢ Unique, reference-independent metric for image captioning.
- ➢ Aligns closely with human evaluations.

**Contrast with Traditional Methods**

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

- ➢ No need for collecting reference captions.
- ➢ Utilizes the CLIP model for assessing similarity.

**Metric Calculation**

- ➢ Measures cosine similarity.
- ➢ Between image's visual CLIP embedding (Ei) and caption's textual CLIP embedding (Ec).

**Scoring**

- ➢ Range: 0 to 100.
- ➢ Closer to 100 indicates optimal performance.

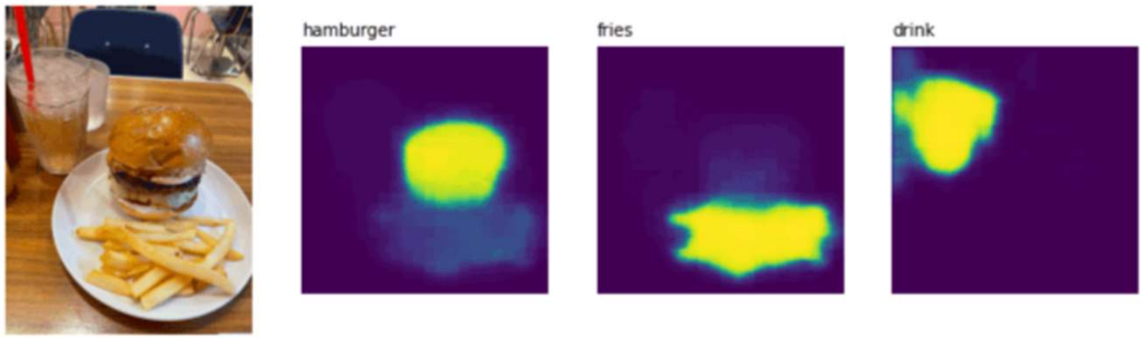
# 02 More Background Knowledge - **CLIPSeg Model for Segmentation-Based Mask**

**What is CLIPSeg?**

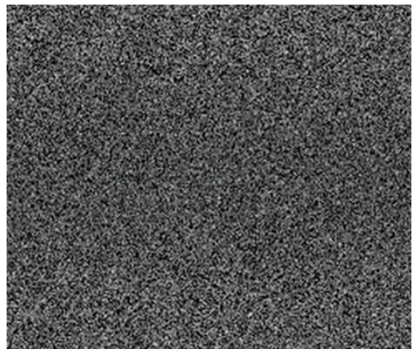
- Enhances the segmentation abilities of the CLIP transformer.
- Suitable for both zero-shot and one-shot tasks.
- [Functionality] CLIPSeg is capable of segmenting images via text query or reference image.
- [Output] Produces a binary mask from input text and images.

**My Experiment**

- Introduced the term "segmentation-based mask" for masks derived from CLIPSeg.
- Combined with a 15% randomized mask for preserving crucial areas and adding variability.
- This approach, named "Segmentation-Based Masked Sampling," aids in generating image samples.




CLIPSeg model example



randomized mask

# More Background Knowledge - **ViT image transformer feature extraction**

**Vision Transformer (ViT) Overview**

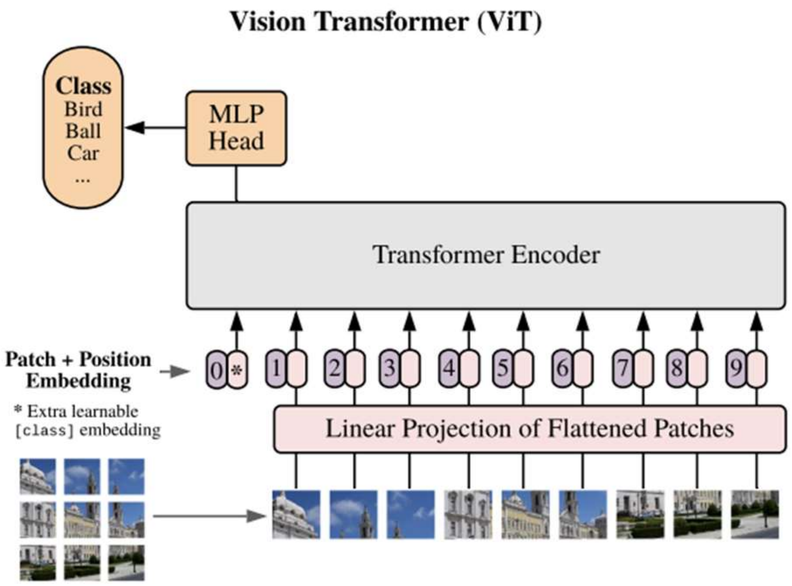
- ➢ Transformer architecture for image feature extraction.
- ➢ Divides images into 2D patches for processing.
- ➢ Uses an NLP-inspired encoder, differing from traditional CNNs.

**The Selected ViT Variant: vit-base-patch16-224**

- ➢ Chosen for its efficiency and simpler design.
- ➢ Comprises 12 transformer layers and 768-dimensional hidden states.

**ViT's Advantage in Image Similarity Detection**

- ➢ Offers a holistic view of images.
- ➢ Enables global image understanding for context-rich feature extraction.
- ➢ Provides deeper semantic insights, surpassing basic pixel comparisons.
- ➢ Enhances accuracy in image similarity detection.



# Proposed Method

# Proposed Method - **Overview**

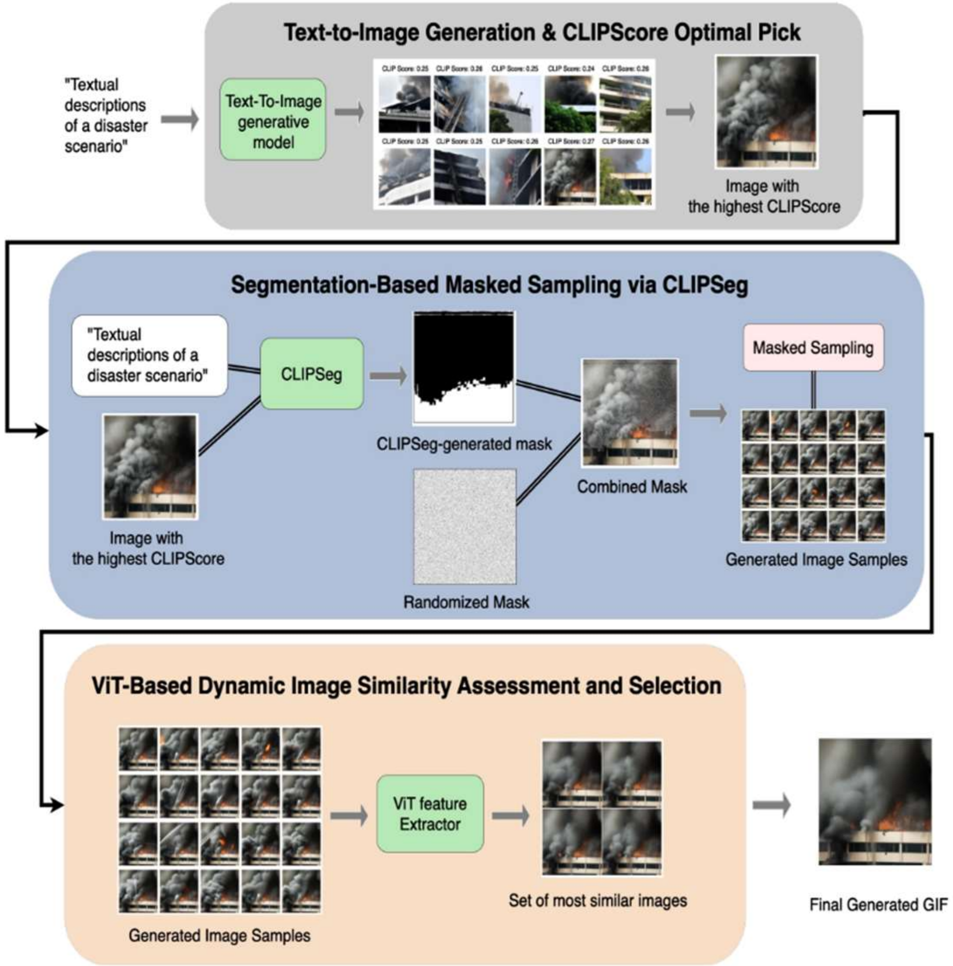



Fig. 1. Proposed Method for Process Steps in the Model

**Step1**.

Image Generation via Text-to-Image Generative Model

**Step2**.

Selection of the Image with the highest CLIPScore

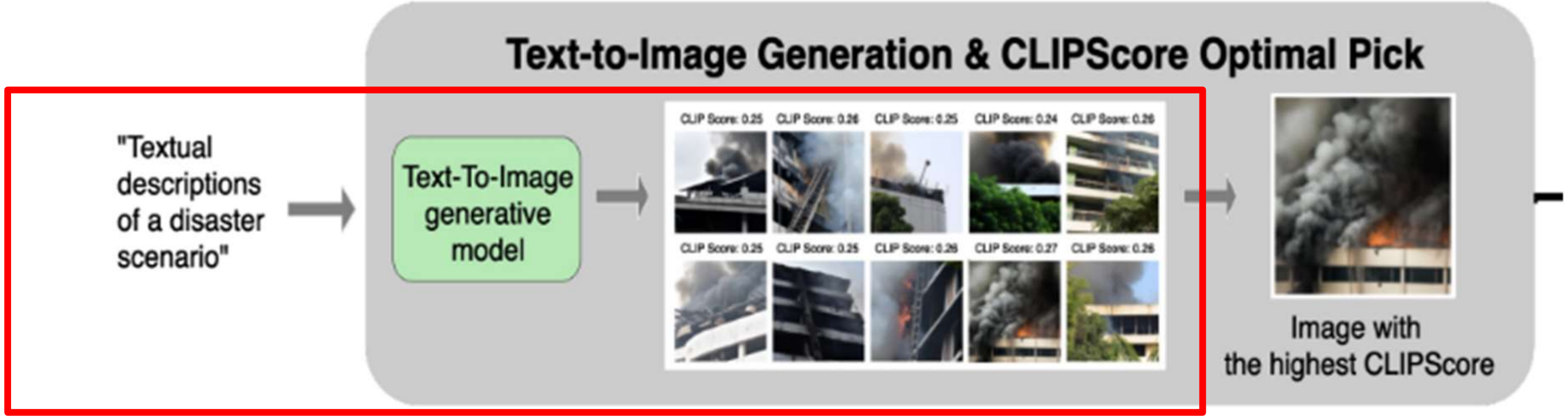
**Step3**.

Segmentation-Based Masked Sampling

**Step4**.

Leveraging ViT feature extraction to dynamically select and merge the images

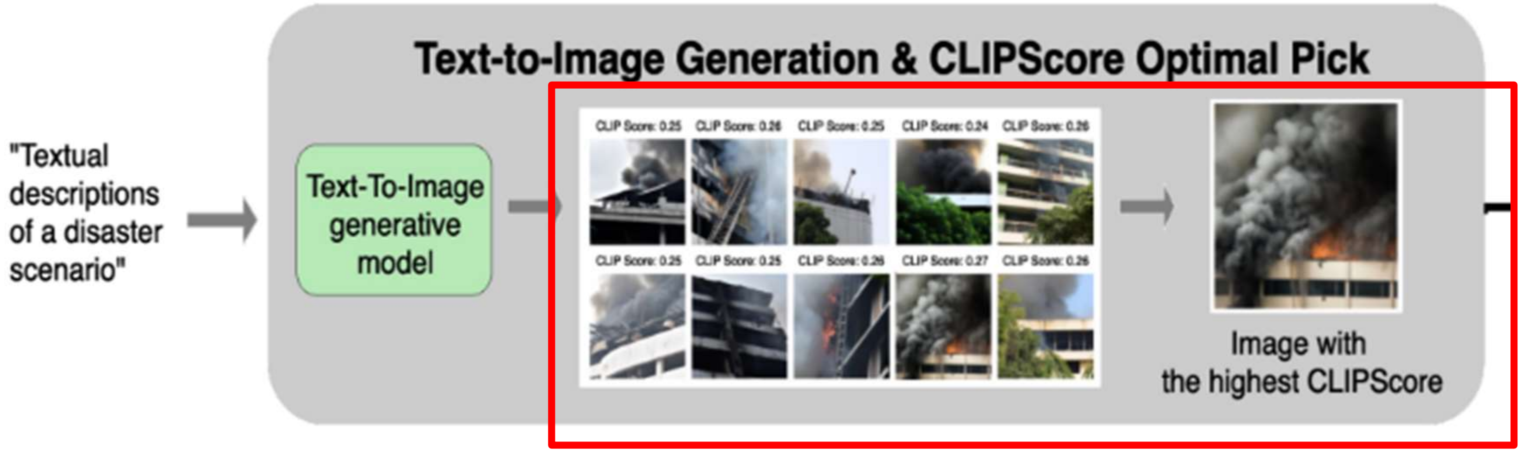
# Proposed Method - **Image Generation via Text-to-Image Generative Model**




**Text-to-Image Generation**

- Utilizes Text-to-Image (T2I) generative model.
- Input: Disaster scenario text.
- **[Output]** Produces a set of images based on the input text.
- **[Output]**  Minimum of 10 images found effective in representing the text.
- **[Used Models]** Stable Diffusion and DALL·E models were used for testing.

## 03 Proposed Method - **Selection of the Image with the highest CLIPScore**



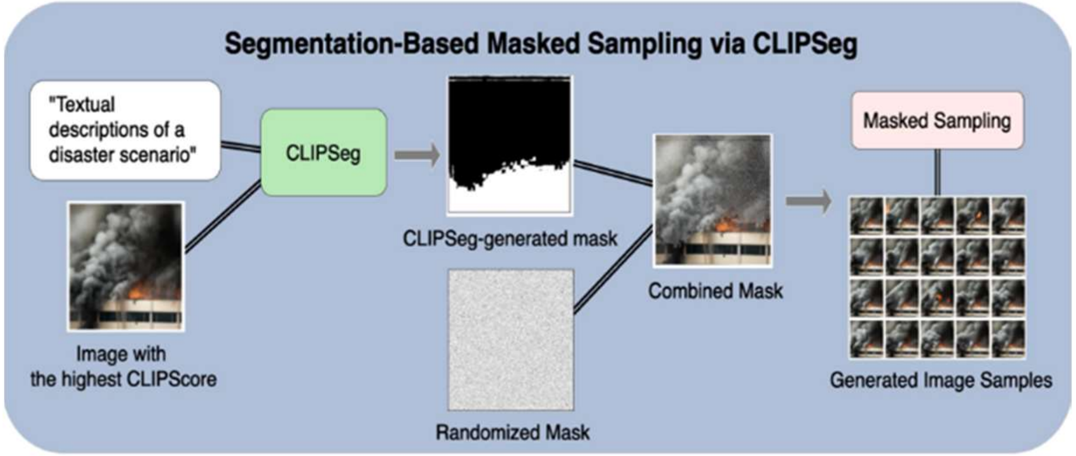

- ➢ Importance in Disaster Scenarios: Essential for visuals to closely match textual descriptions.
- ➢ CLIPScore is used for Accuracy Measurement.
- ➢ **Image Selection:** Chooses the image with the highest CLIPScore from the generated set.
- ➢ High CLIPScore indicates a close match to the text description.

### <u>Comparative Results</u>

- ➢ Example: For the scenario "A fire has broken out in the building":
    - ○ DALL·E images achieved a CLIPScore of 25.5%.
    - ○ Stable Diffusion images scored slightly higher at 25.6%.

# Proposed Method - **Segmentation-Based Masked Sampling**

- ➢ **Utilizing Top-Rated Images** - Selects the image with the highest CLIPScore.
- ➢ **Primary Subject Identification** - Identifies and highlights the main subject from the text.
  - ○ Example: "fire" in the scenario "A fire has broken out in the building".
- ➢ **Masking Secondary Components** - Uses CLIPSeg to mask out less important elements.
  - ○ Renders static elements like buildings opaque.
- ➢ **Blending with Randomized Mask** - Combines the primary mask with a 15% randomized mask.




**Segmentation-Based Masked Sampling Strategy**

- ➢ Focuses on segmentation for sampling.
- ➢ Creates diverse image variants that emphasize key features.
- ➢ Masks static or less relevant elements for clarity.

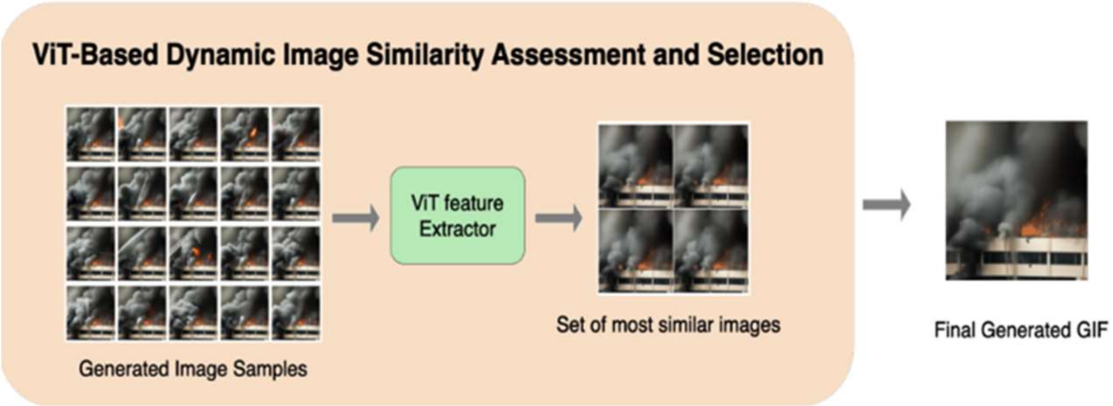
# 03 Proposed Method - **Leveraging ViT feature extraction to dynamically select and merge the most analogous images**

- **Segmentation-Based Masked Sampling** - Generates diverse image samples.
- **Grouping Similar Images for Animation** to group the most similar images for fluid animation.

- **Adapting ViT-B for Feature Extraction & Computing Image Similarity**
  - Similarity measured by pairwise Euclidean distances between features.

- **Dynamic Programming for Image Selection -**
  - Evaluates image combinations for cumulative distances. & Chooses most analogous images based on ViT-B features.

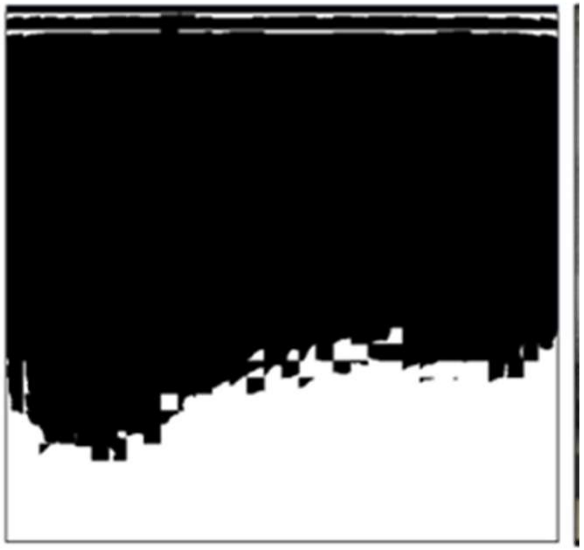
- **Creating Animated Representation** - Uses linear interpolation to craft a seamless animation.



# Results

Image Created by T2I model

CLIPSeg Model
Generated Mask



Combined Mask
(CLIPSeg + Randomized Mask)

## 04 Results - **Final Animated Image**

+

**Image**

**Combined Mask**

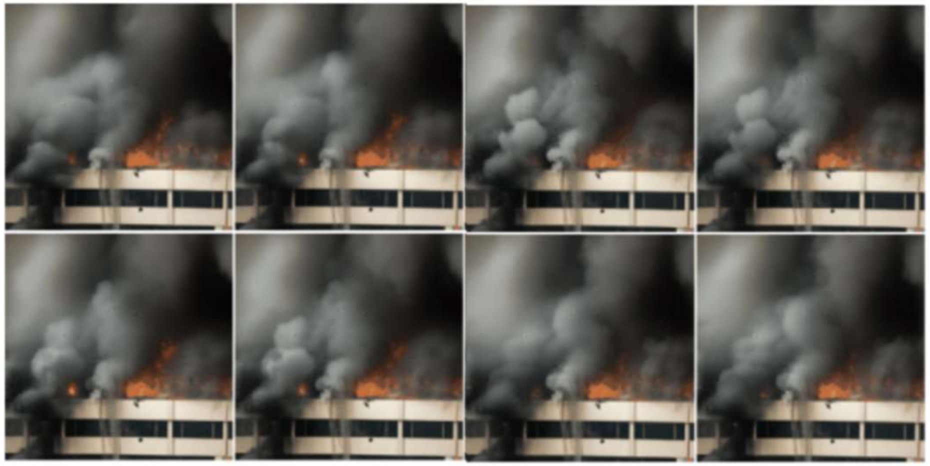


**Image created by Masked Sampling**

**Input text description:**

***"A fire has broken out in the building."***

**Final Result**

## Thank You!

# Q & A

# 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, “Highresolution 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."ZeroShot Text-to-Image Generation," arXiv preprint arXiv:2102.12092, 2021.