# Introduction to Generative Models and Text-to-Image and Video Models

Rubin Won
UST-ETRI
MS Student
rubrub@etri.re.kr
October 5th, 2023

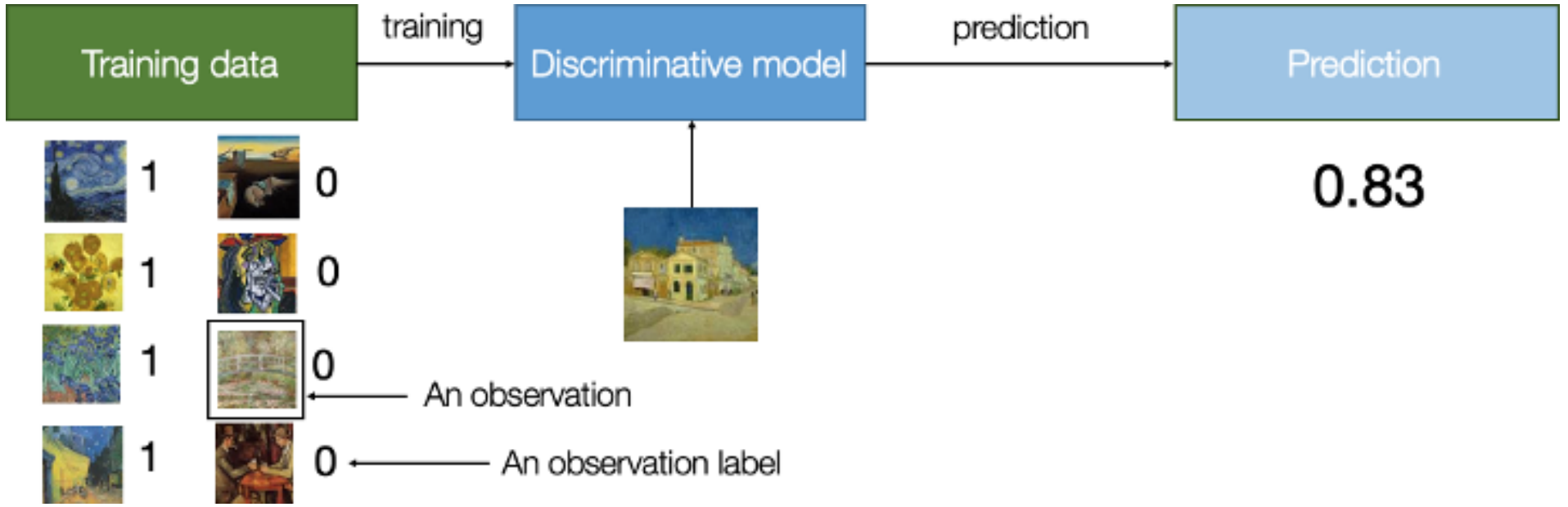
# Contents

- Introduction & Background
- Related Works
- Motivation / What’s Next

# Introduction & Background

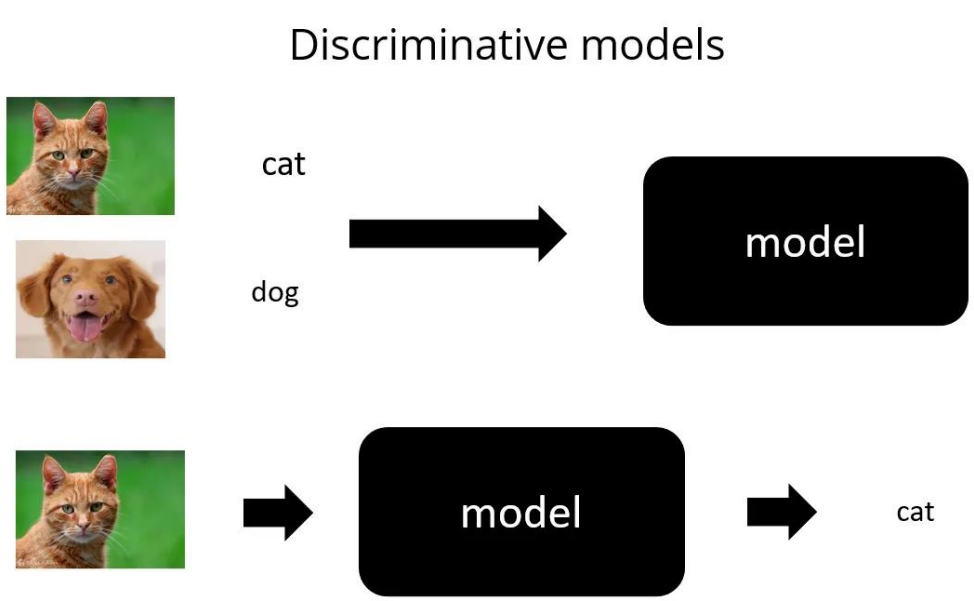
➢ A **generative model** can be defined as follows:

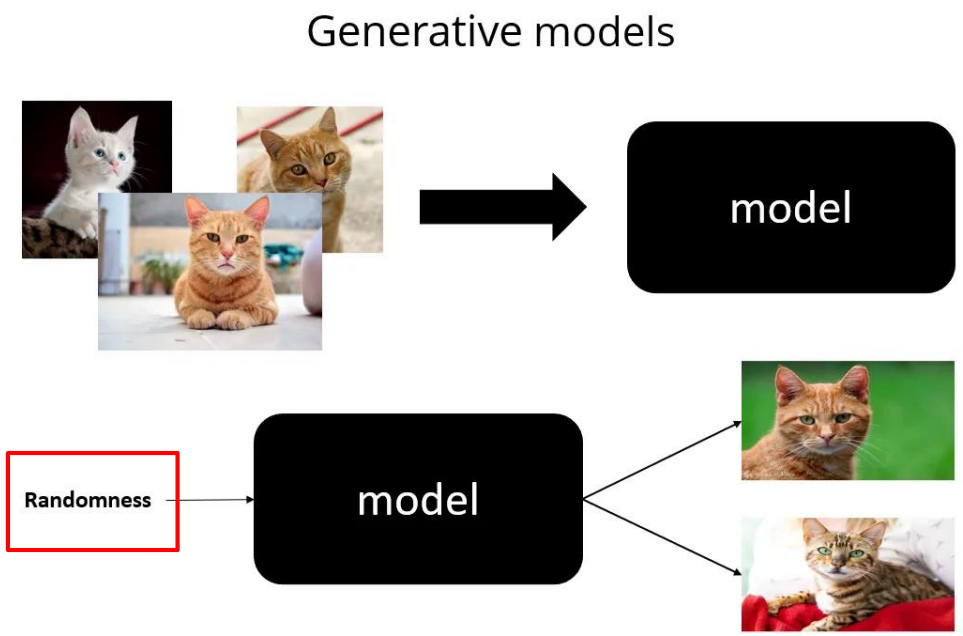
- A generative model describes how a dataset is **generated**, in terms of a **probabilistic model.** By **sampling** from this model, we are able to **generate new data.**




**< Generative modeling process >**

➢ A generative model must also be **probabilistic** rather than deterministic.



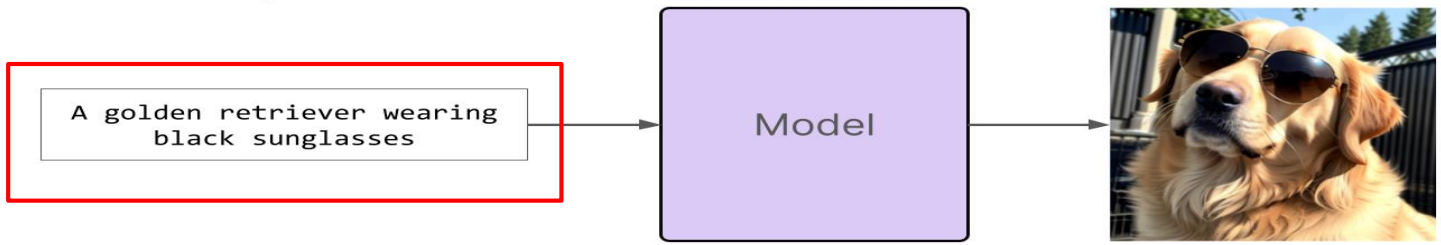
# Introduction & Background - **What is Text-to-Image Model?**

➢ **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.

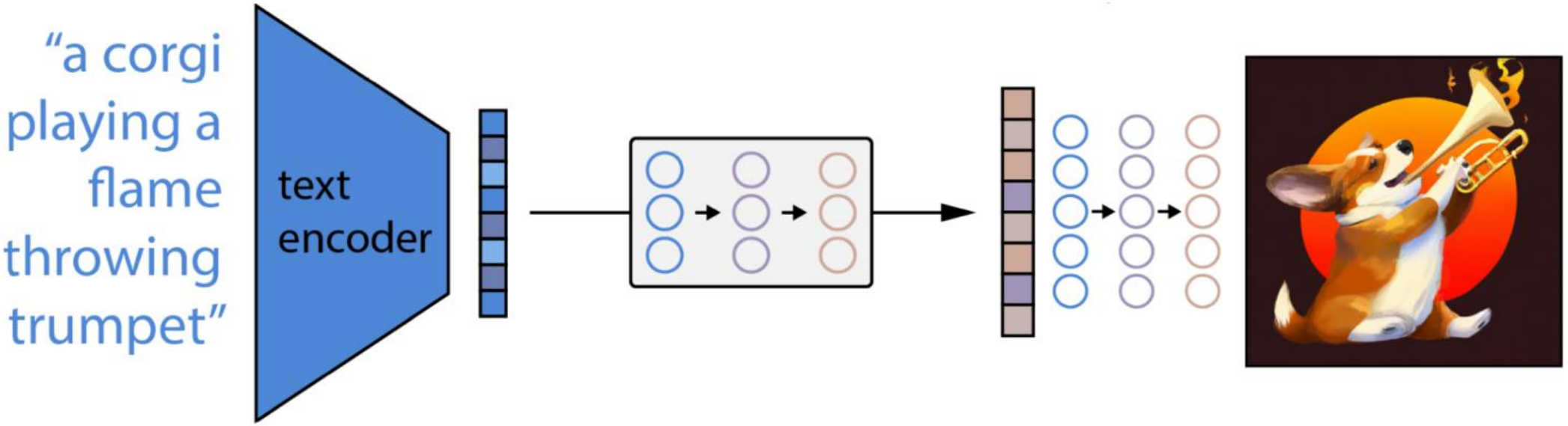
**Basic Generative Models**


**Text-to-Image Models**



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

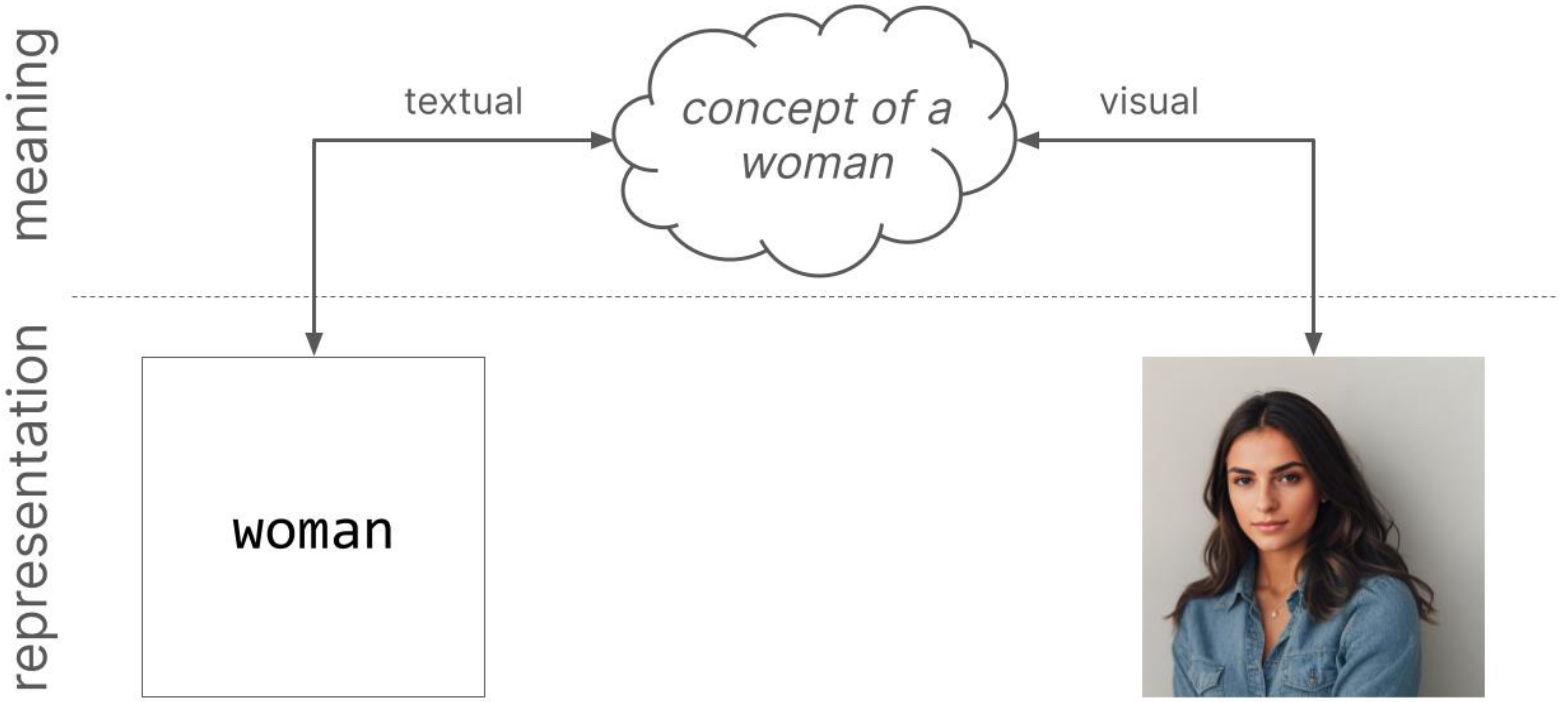
- **Component 2.**
  - A **decoder model** that decodes this “meaning vector” into an image




**Component 1** **Component 2**

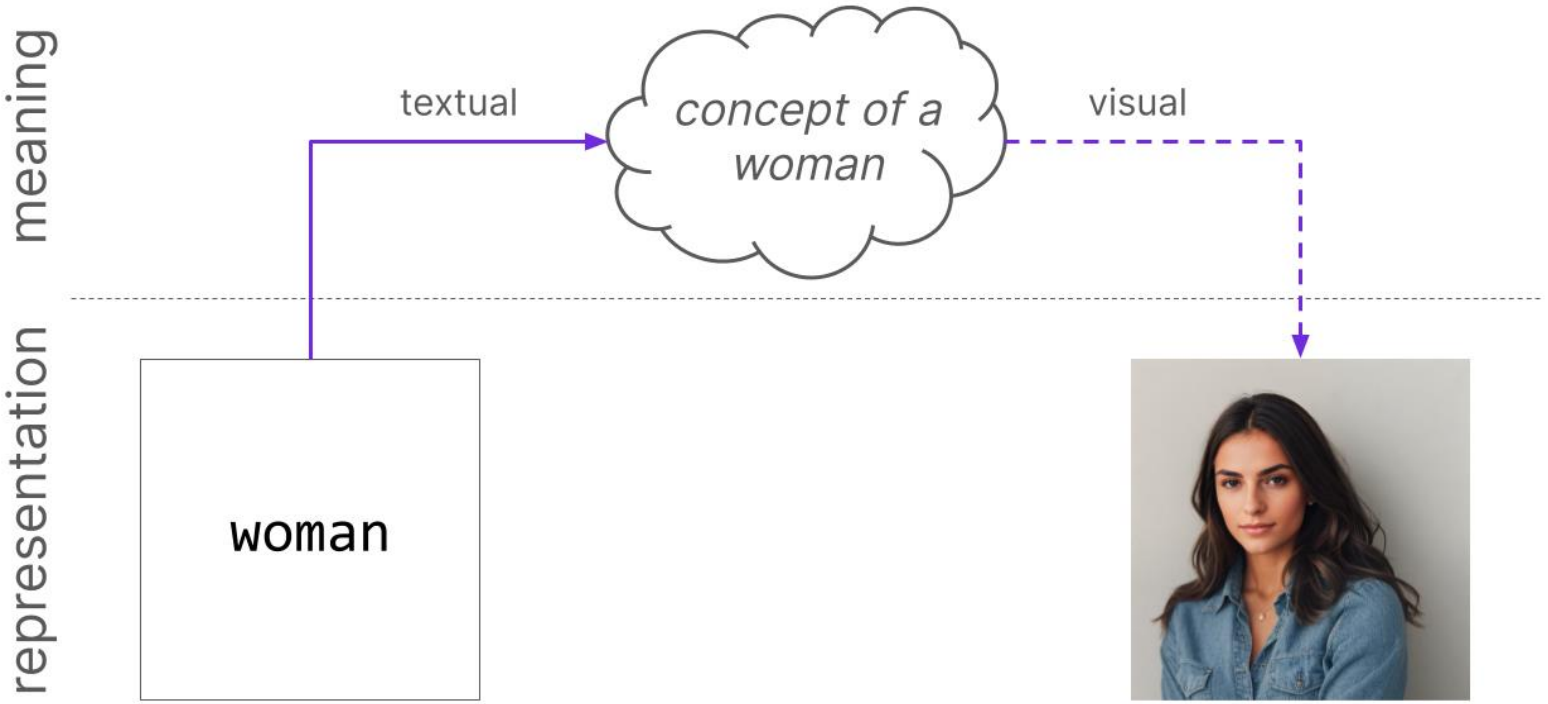
# Introduction & Background - **High Level look of the T2I process**

- First, extract the **meaning(concept)** from the text by using a text encoder
- Then we learn how to map from words to meaning.



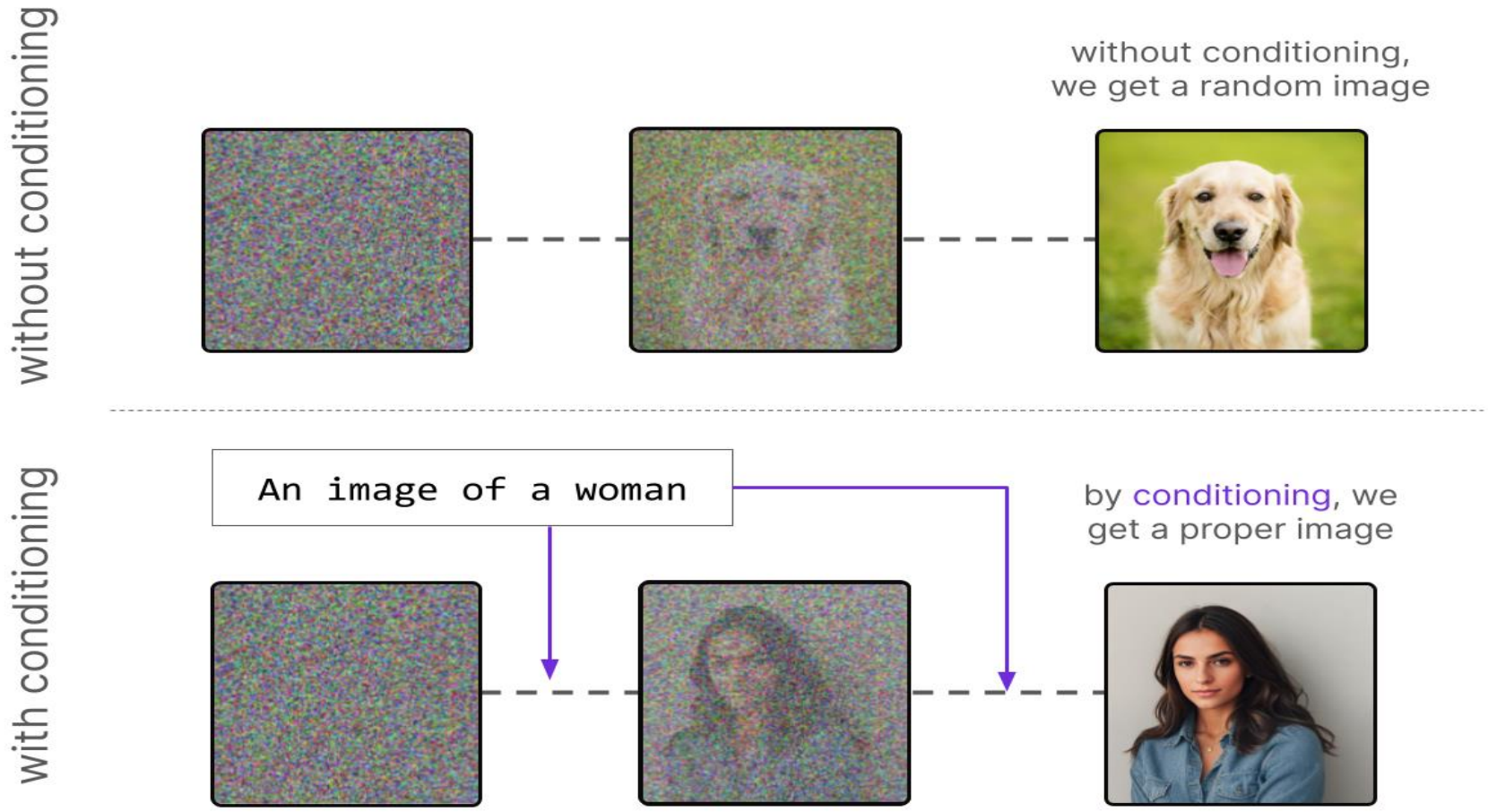
# Introduction & Background - **High Level look of the T2I process**

- Now we learn to map from meaning to images(visual space).
- We use “**conditioning”** using the meaning vector to condition the generation process.



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

➢ **Conditioning** can be considered the practice of providing additional information to a process to impose a condition on its outcome.



Given a textual Condition: **“An image of a woman” :**

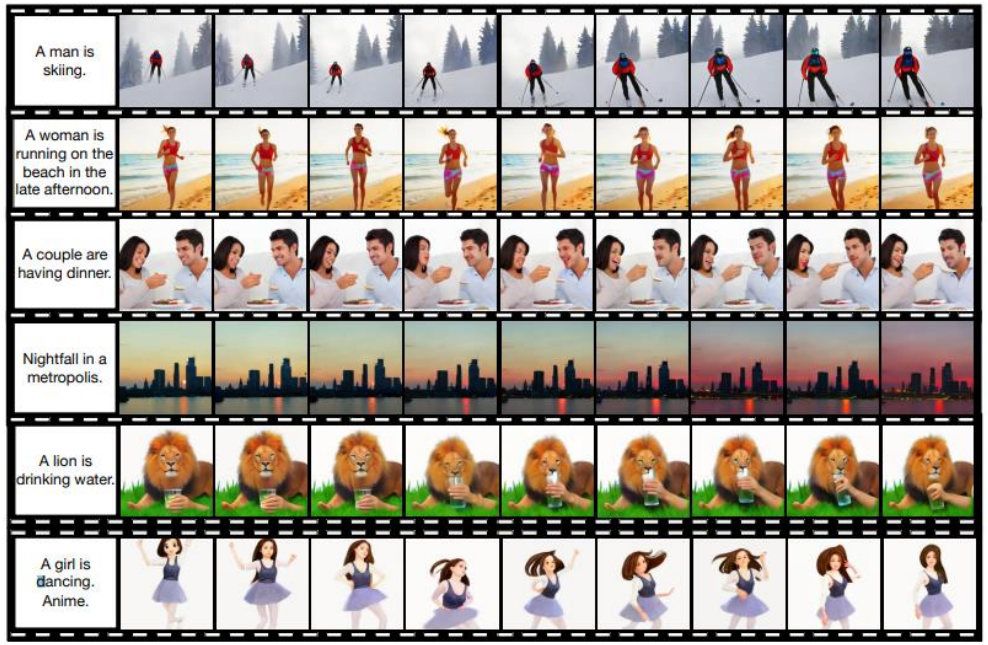
➢ Using the same **"meaning vector"**, generative model (Stable Diffusion) can produce multiple images that capture the intended meaning.

➢ **Text-to-Video (T2I) model** is an extension of the Text-to-Image (T2I) concept but focuses on generating videos instead of static images based on textual descriptions.

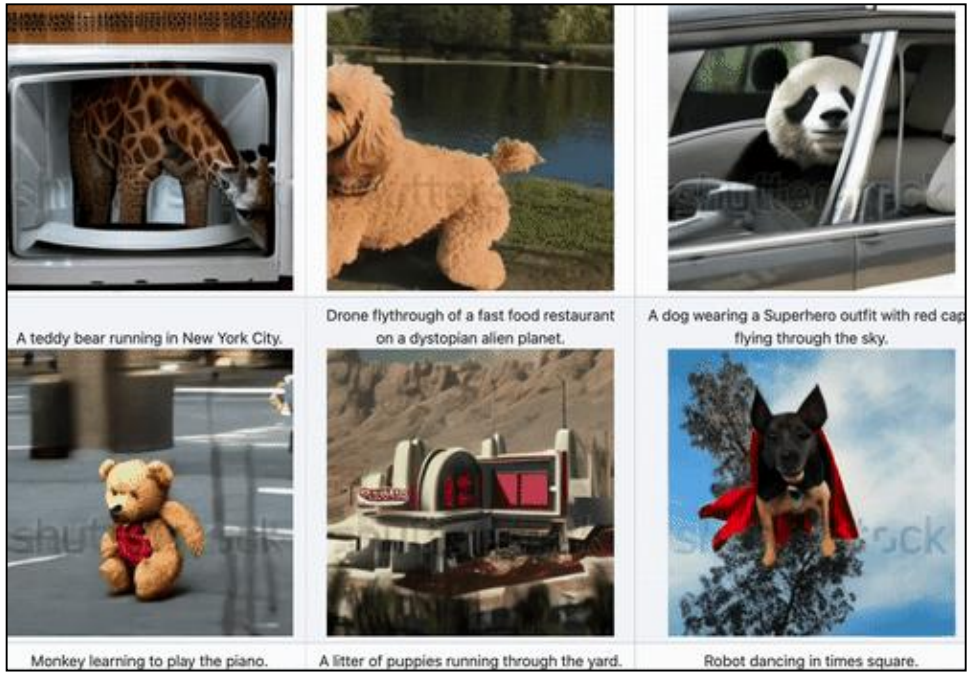
**“A teddy bear painting a portrait”** ⟹

**Method 1: Learning from Image-Video Pair Datasets**

- This method relies on learning from datasets that contain pairs of videos and associated text descriptions. It uses this data to generate videos based on text inputs.



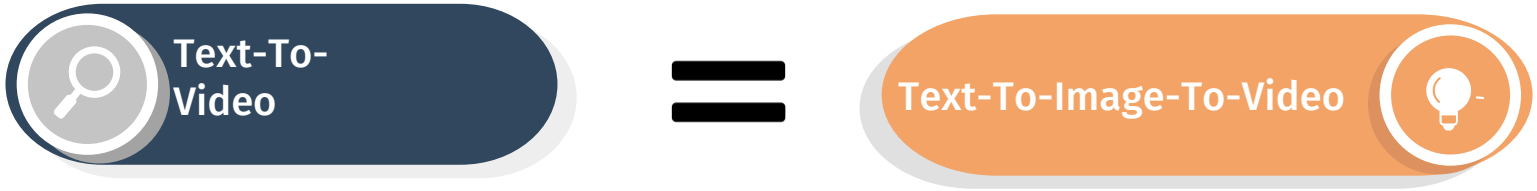

**Trained on video-text pair datasets**



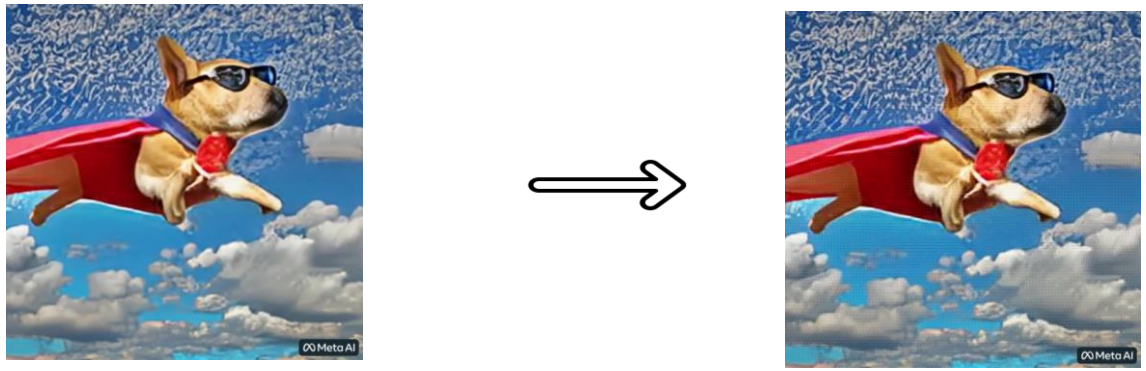

**Generated Videos (text provided)**

**Method 2: Text-To-Image + Motion → Text-To-Video**

- This method combines the capabilities of Text-To-Image generation with motion generation to create Text-To-Video systems.




**"A dog wearing a Superhero outfit with red cape flying through the sky,"**



# Related Works

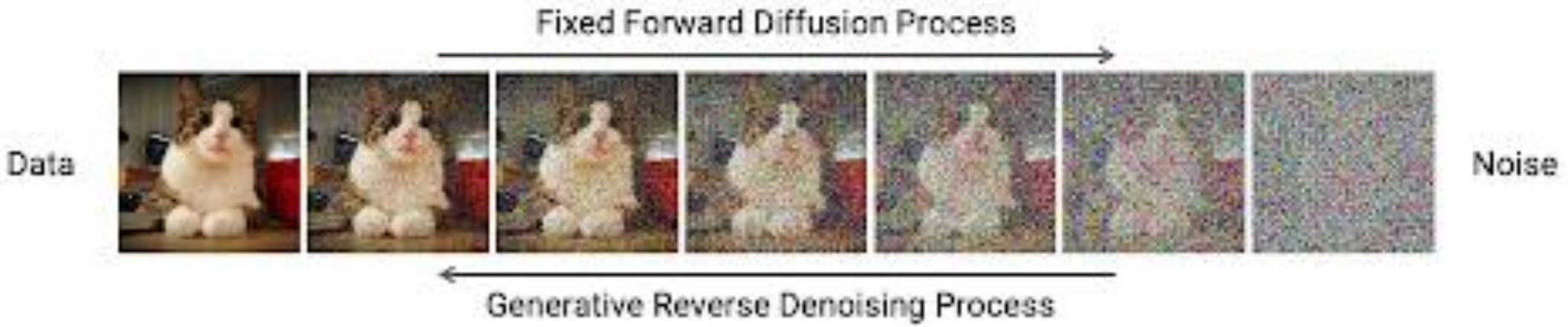
## Stable Diffusion

- Stable Diffusion is a text-to-image model released in 2022 based on diffusion techniques.

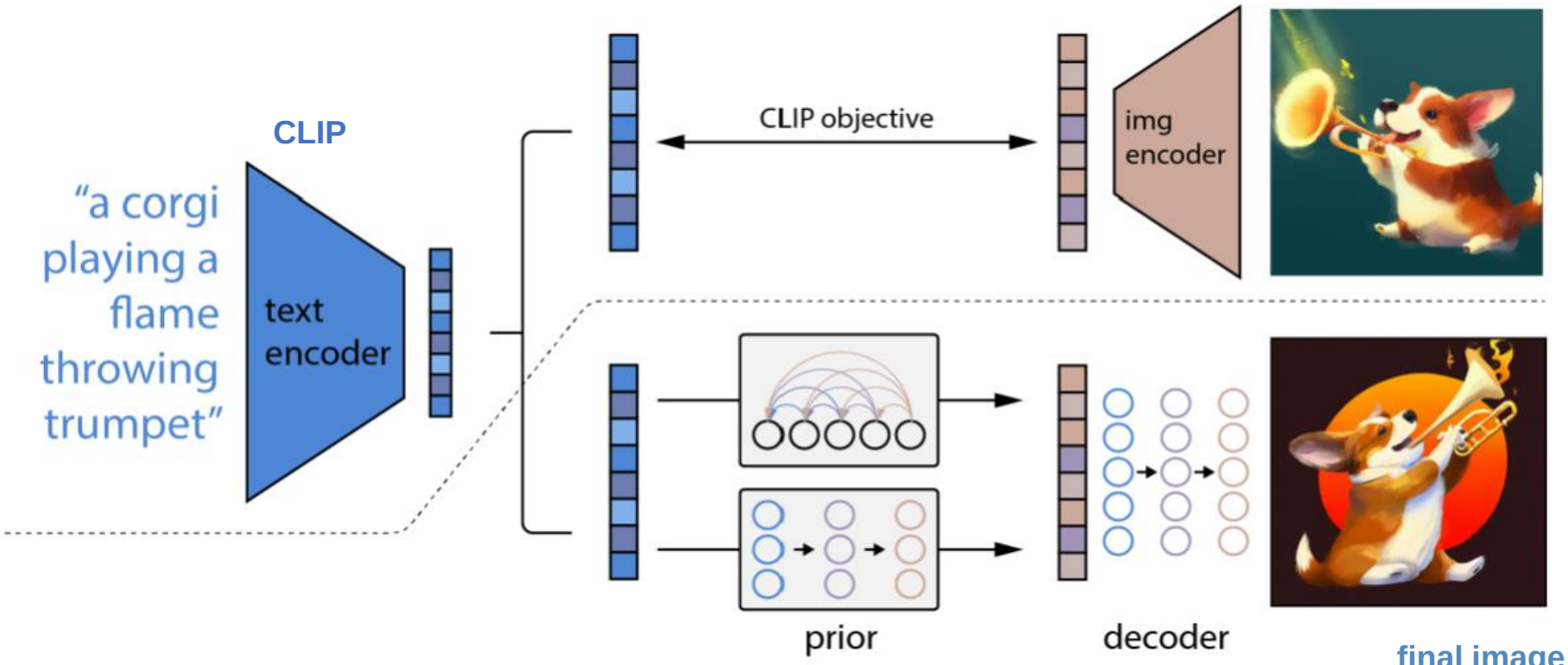

**"a photograph of an astronaut riding a horse"**

## Diffusion Process

➢ ML system that are trained to denoise random Gaussian noise step by step, to get sample of interest (image)



# Related Works / Application of T2I models - **Dall-E 2**




**CLIP training process**
- to learn join representation space (text & image)

**text-to-image process**
- autoregressive / diffusion prior to produce image embedding

**step1**. CLIP model training == (text, image pair)

**step2**. Train model "prior" == (text → image embedding)

**step3**. Decoder == (image embedding → image)

# Related Works / Application of T2V models - **Google Imagen Video (2022.10.5)**

➢ **Google Imagen Video** is a **Cascaded Diffusion Model,** that is, a step-by-step diffusion model.
- (Initially create roughly, and gradually enhance the resolution and frame rate)

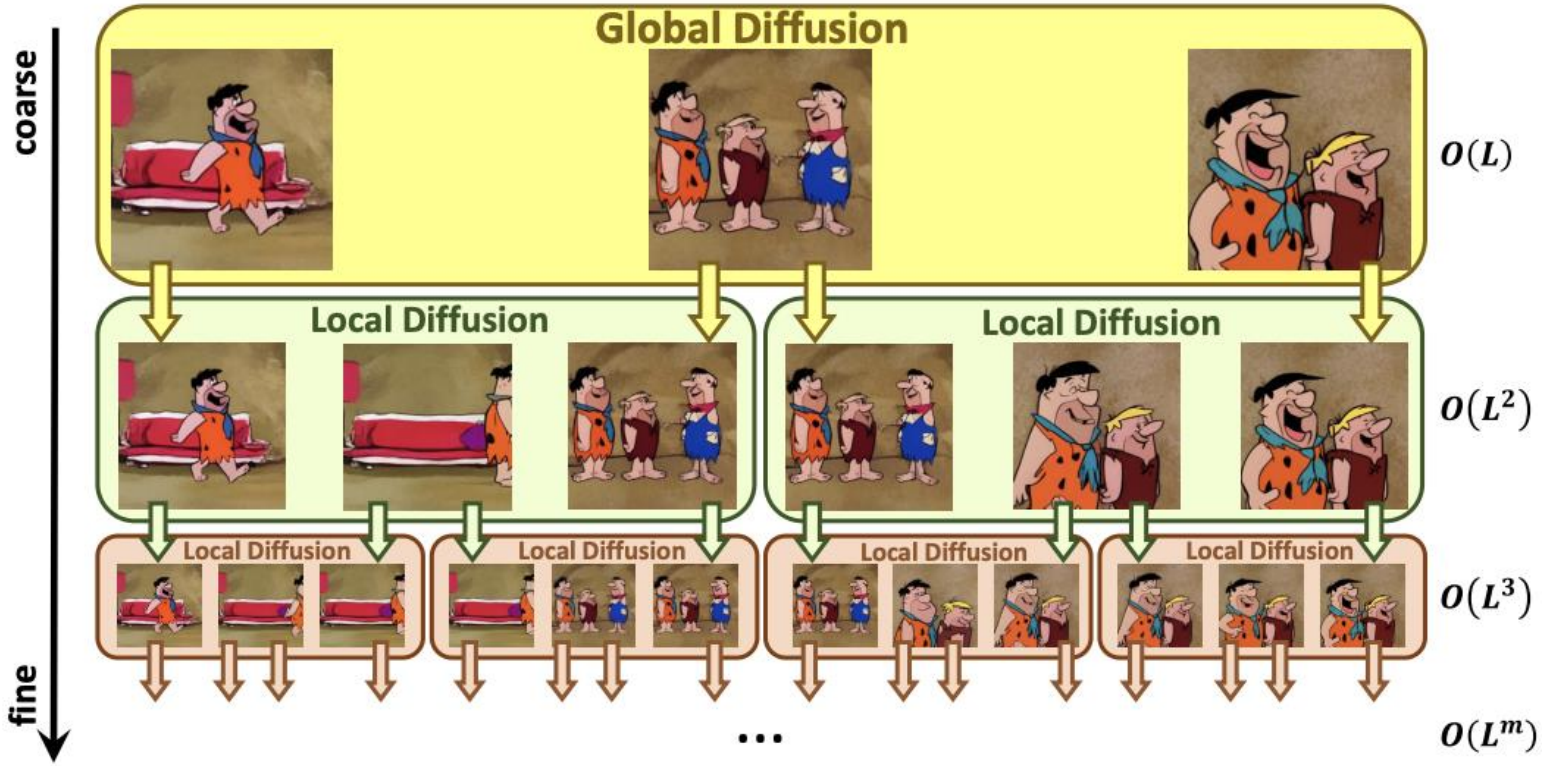

**Use of models:**

[1] Time Super Resolution(TSR)
2] Spatial Super-Resolution(SSR)

- **Repeat!**

# Related Works / Application of T2V models - **NUWA-XL (2023.3.22)**

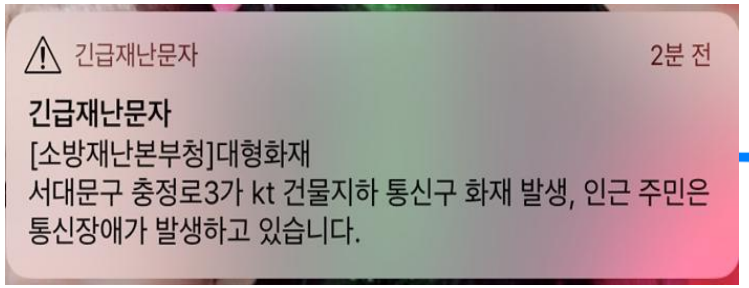
➢ **NUWA-XL** is an architecture based on the "**Diffusion over Diffusion**" approach that generates long videos through a "**coarse-to-fine**" process.



# Motivation

# 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**

Text-To-Image generation model

**visual information for vulnerable populations**

# 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.**

# What’s Next?

- Detail Technology of Generative models: VAEs, GANs, diffusion, transformers
- Detail Technology used for my research

## Thank You!

# Q & A

# References


[1] Microsoft Research Asia. (n.d.). NUWA-XL. Retrieved from https://msra-nuwa.azurewebsites.net/#/

[2] Make a Video Studio. (n.d.). Retrieved from https://makeavideo.studio/

[3] ArXiv. (2022). [Title of the Paper]. Retrieved from https://arxiv.org/abs/2209.14792

[4] WandB. (n.d.). A Gentle Introduction to Dance Diffusion. Retrieved from https://wandb.ai/wandb_gen/audio/reports/A-Gentle-Introduction-to-Dance-Diffusion--VmlldzoyNjg1Mzky

[5] Towards Data Science. (n.d.). What Are Stable Diffusion Models and Why Are They a Step Forward for Image Generation? Retrieved from https://towardsdatascience.com/what-are-stable-diffusion-models-and-why-are-they-a-step-forward-for-image-generation-aa1182801d46

[6] Fotor Blog. (n.d.). What Is Stable Diffusion? Retrieved from https://www.fotor.com/blog/what-is-stable-diffusion/

[7] AssemblyAI Blog. (n.d.). Modern Generative AI Images. Retrieved from https://www.assemblyai.com/blog/modern-generative-ai-images/

[8] OpenReview. (n.d.). [Title of the Paper]. Retrieved from https://openreview.net/pdf?id=n7XbkHOwKn6

[9] Medium. (n.d.). An In-Depth Look at the Transformer-Based Models. Retrieved from https://medium.com/@yulemoon/an-in-depth-look-at-the-transformer-based-models-22e5f5d17b6b#:~:text=Autoregressive%20(AR)%20and%20autoencoding%20(,abstractive%20summarization%20and%20question%2Danswering.

[10] Hugging Face. (n.d.). Text-to-Video. Retrieved from https://huggingface.co/blog/text-to-video