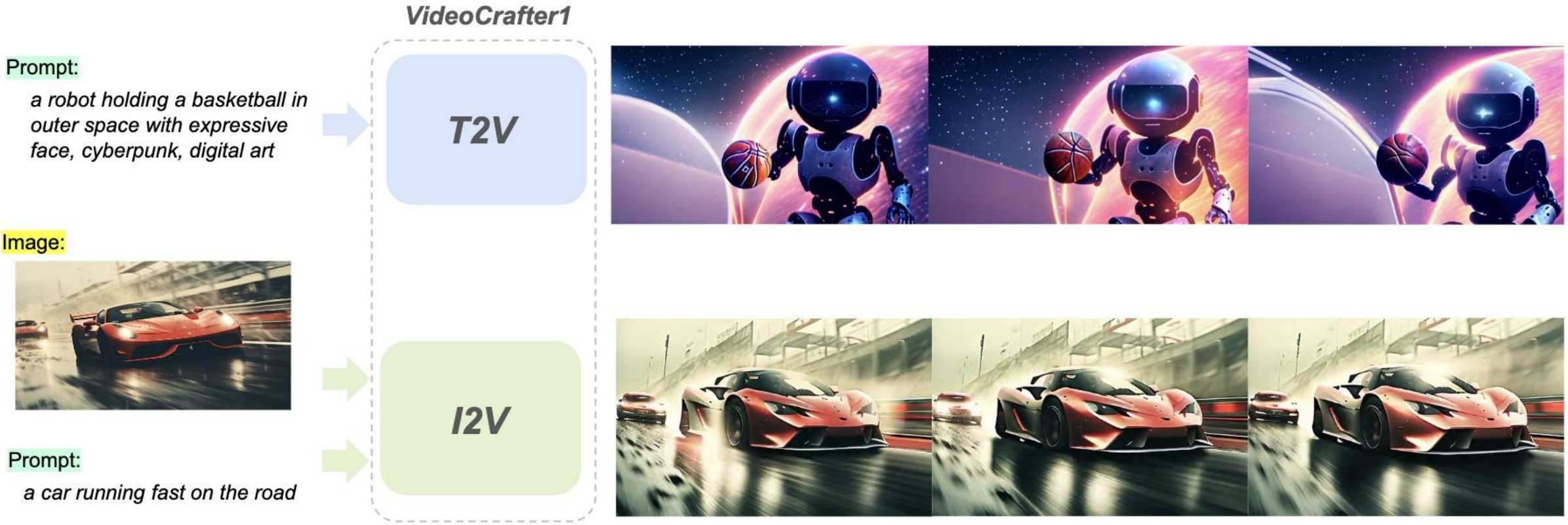
VideoCrafter1: Open Diffusion Models for High-Quality Vidoe Generation

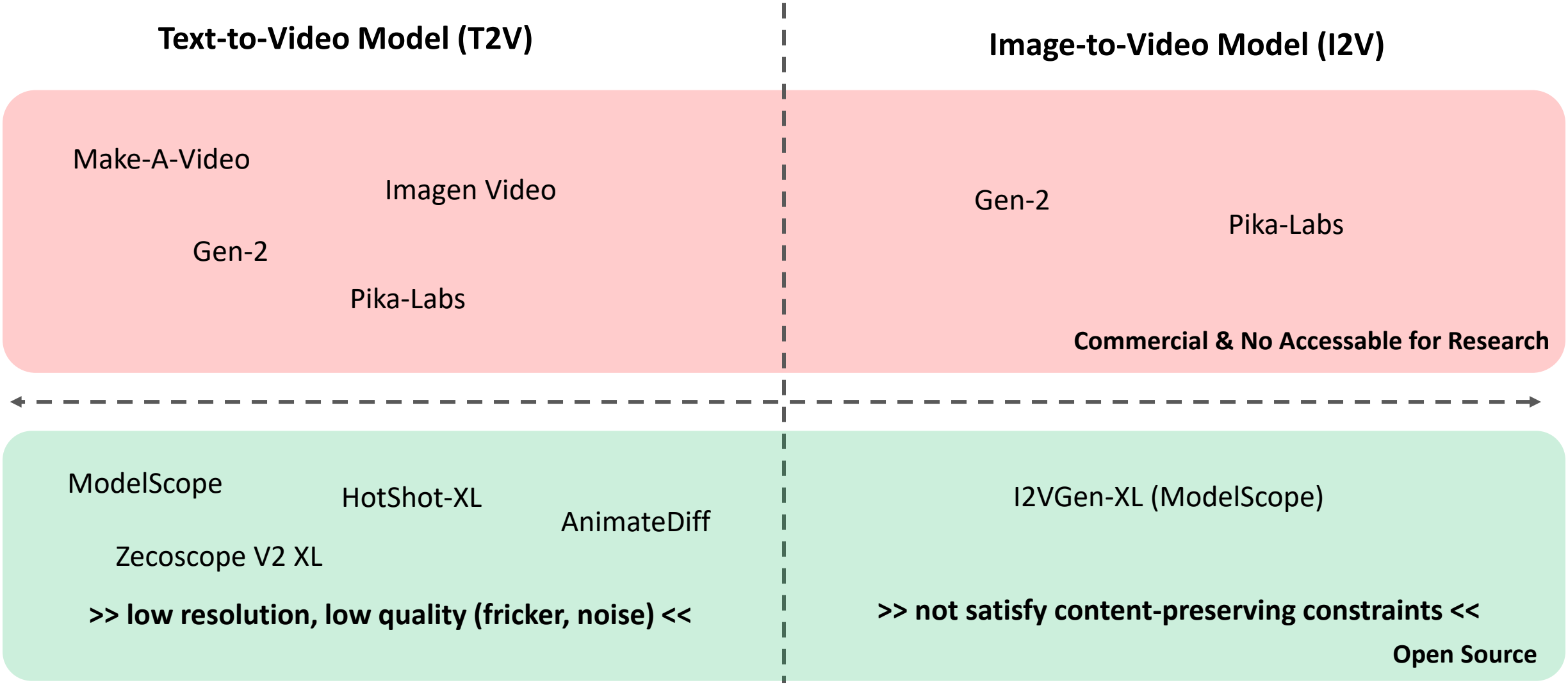
Chen, Haoxin, et al., Arxiv 2023

2024.09.24 Jihun Chae

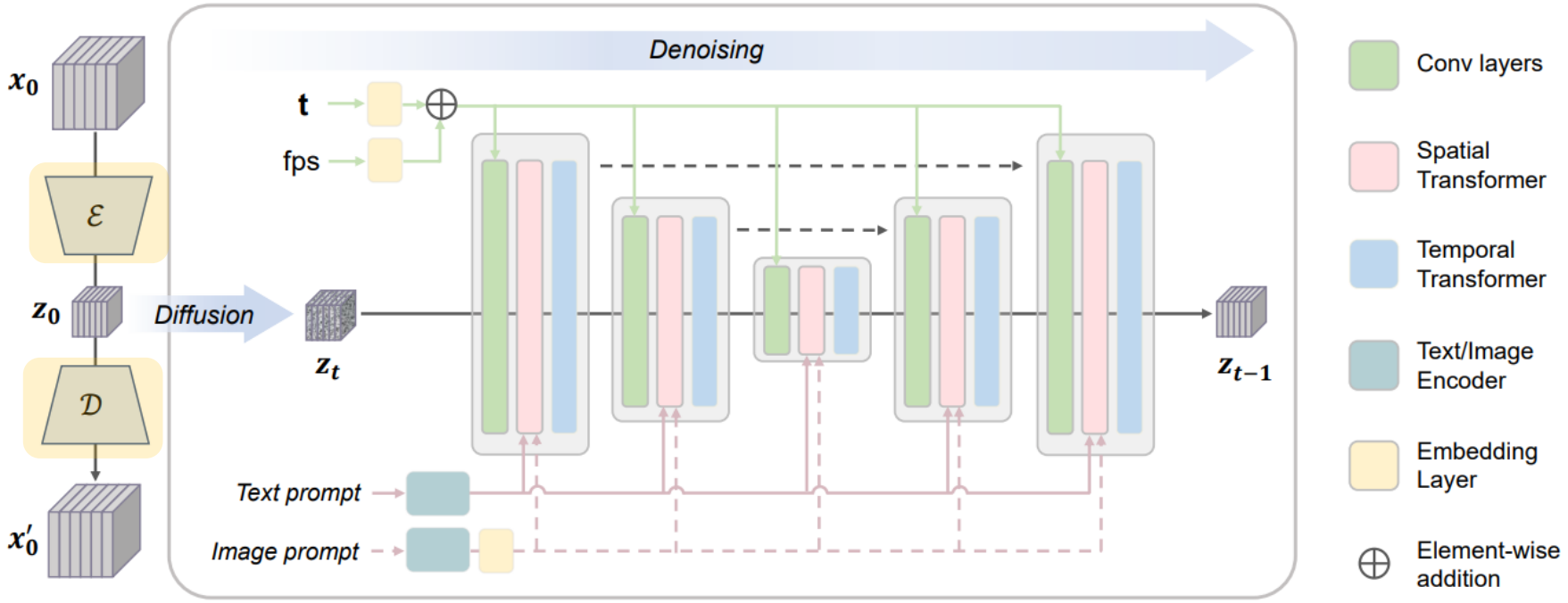
VideoCrafter1: Open Diffusion Models for High-Quality Vidoe Generation



Background


Text-to-Video Model (T2V)
Image-to-Video Model (I2V)
Make-A-Video
Imagen Video
Gen-2
Pika-Labs
Gen-2
Pika-Labs
Commercial & No Accessable for Research
ModelScope
HotShot-XL
AnimateDiff
Zecoscope V2 XL
>> low resolution, low quality (fricker, noise) <<
I2VGen-XL (ModelScope)
>> not satisfy content-preserving constraints <<
Open Source

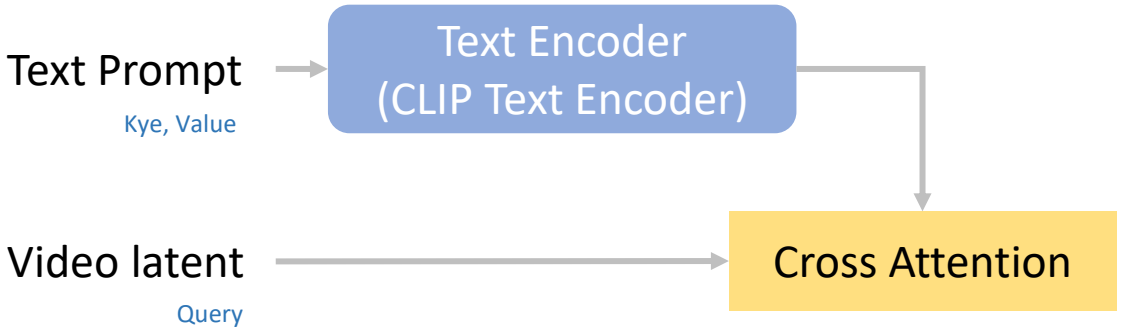
VideoCrafter1: Text-to-Video Model (T2V)

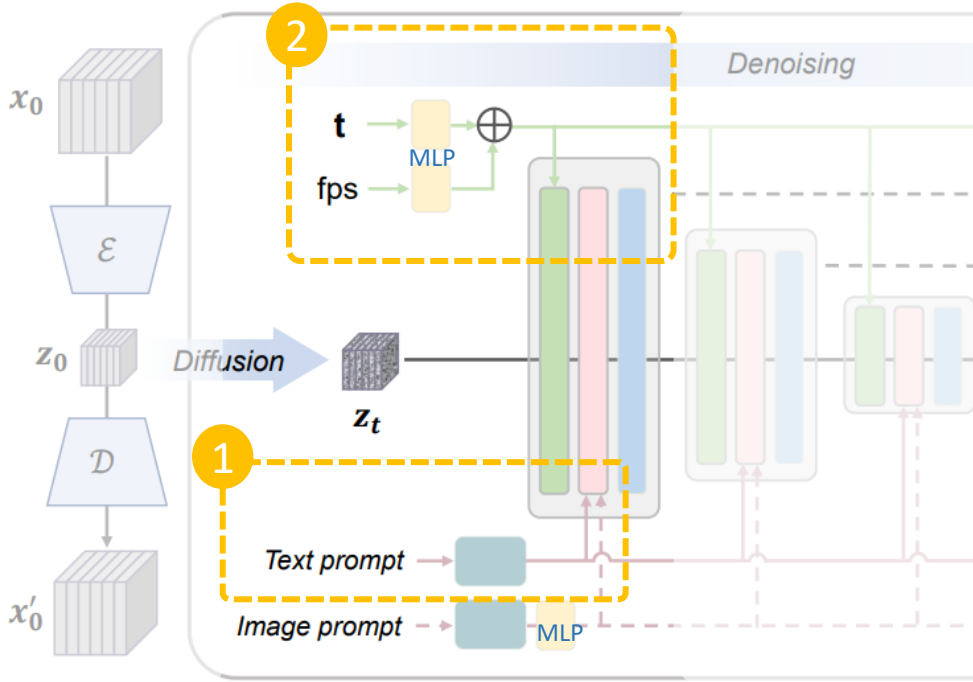


- Latent Video Diffusion Model
  - Video VAE + Video Latent Diffusion Model
- Denoising 3D U-Net
  - Spatial-Temporal Blocks (Convolutional layer, Spatial Transformer, Temporal Transformer)

VideoCrafter1: Text-to-Video Model (T2V)

- Structure Overview - Denoising 3D U-Net
  - Inject semantic control via cross-attention
    - semantic control : text prompt / motion speed control / video FPS
    - Text prompt ⟪ cross-attention
    - Motion speed control & video FPS ⟪ MLP & element-wise addiction
  - Cross Attention Mechanism 1
    - $Cross\ Attention\ (Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V$





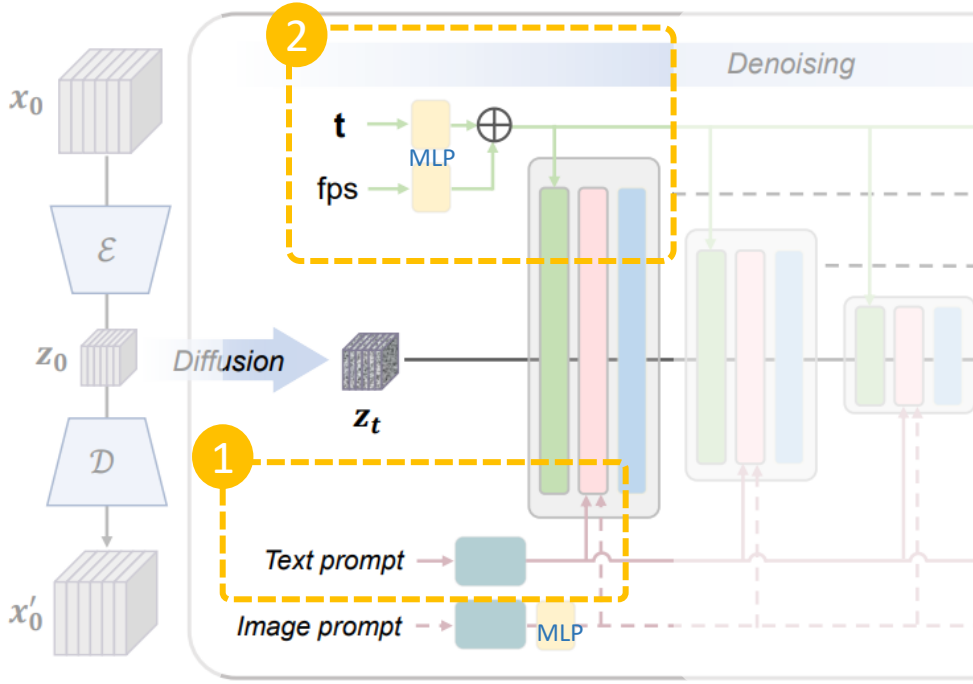
VideoCrafter1: Text-to-Video Model (T2V)

- Structure Overview - Denoising 3D U-Net
  - Inject semantic control via cross-attention
    - semantic control : text prompt / motion speed control / video FPS
    - Text prompt ⟪ cross-attention
    - Motion speed control & video FPS ⟪ MLP & element-wise addiction
  - Motion Speed control with FPS 2

```
def forward(self, x, timesteps, context=None, features_adapter=None, fps=16, **kwargs):

    t_emb = timestep_embedding(timesteps, self.model_channels, repeat_only=False)
    emb = self.time_embed(t_emb)

    if self.fps_cond:
        if type(fps) == int:
            fps = torch.full_like(timesteps, fps)
        fps_emb = timestep_embedding(fps,self.model_channels, repeat_only=False)
        emb += self.fps_embedding(fps_emb)
```



VideoCrafter1: Image To Video Model (I2V)

- Image To Video Model (I2V)
  - Text prompt offer highly flexible control for content generation
  - But, **focus on semantic-level** rather than detail appearance

  → Aim to integrate an additional conditional input : *image input*

  Text 정보로는 세세한 표현을 전달하기 힘드니, Image 정보를 추가하자!

  - Text-Align Rich Image Embedding
    - CLIP의 Text Encoder에 대응되는 Image Encoder 사용하여 Image feature 추출
    - full patch visual tokens $F_{vis} = \{f_i\}_{i=0}^{K}$ vs. global semantic token f_{cls} from CLIP Image Encoder
      - Global semantic token represents visual content at a semantic level, less capable of capturing details.
      - Full patch visual tokens be obtained last layer of the CLIP image ViT

Image source:https://theaisummer.com/vision-transformer/

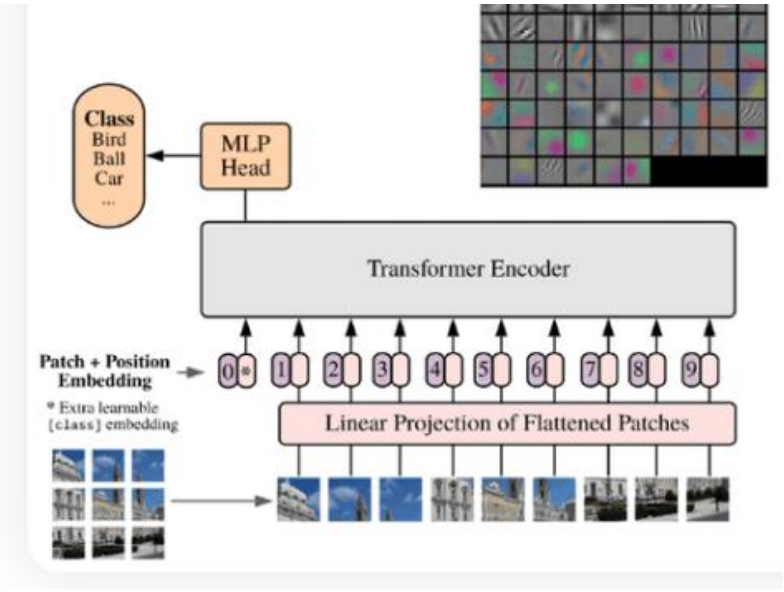
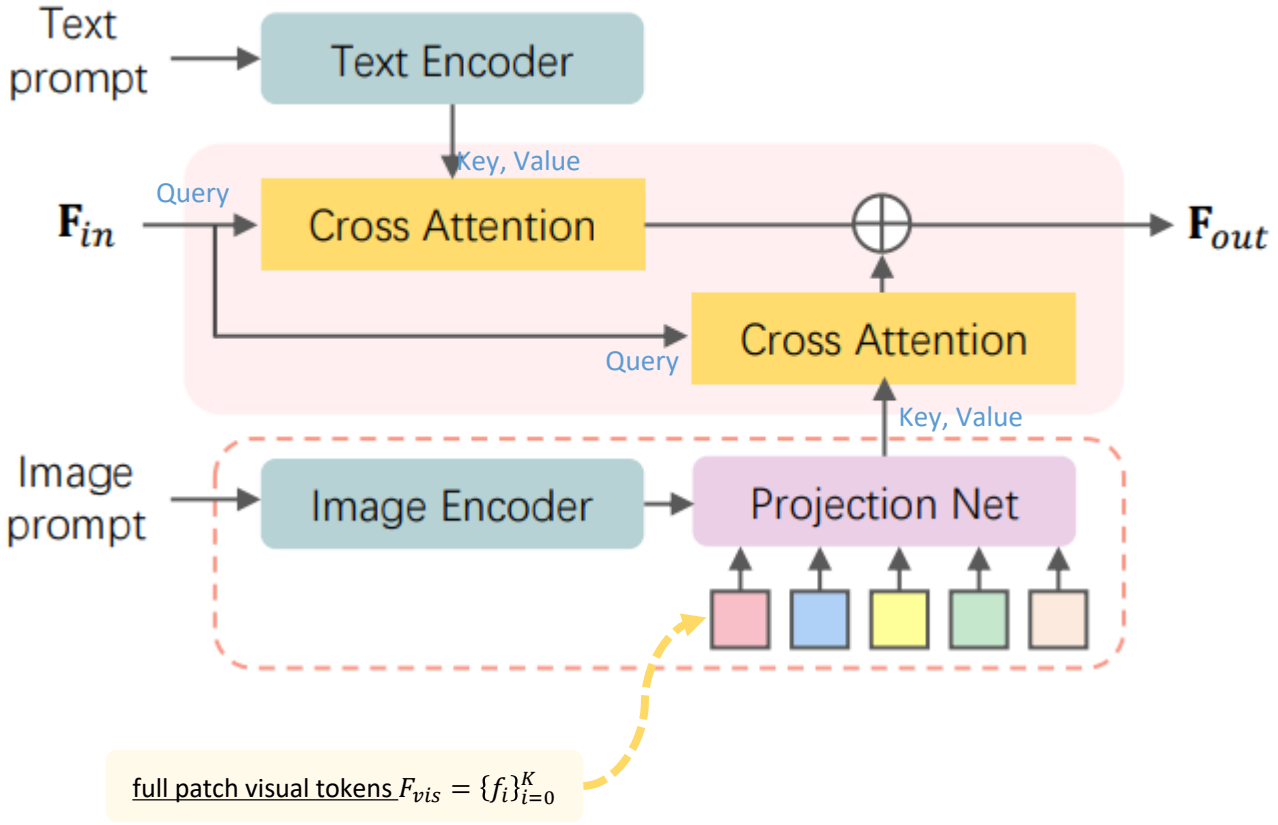


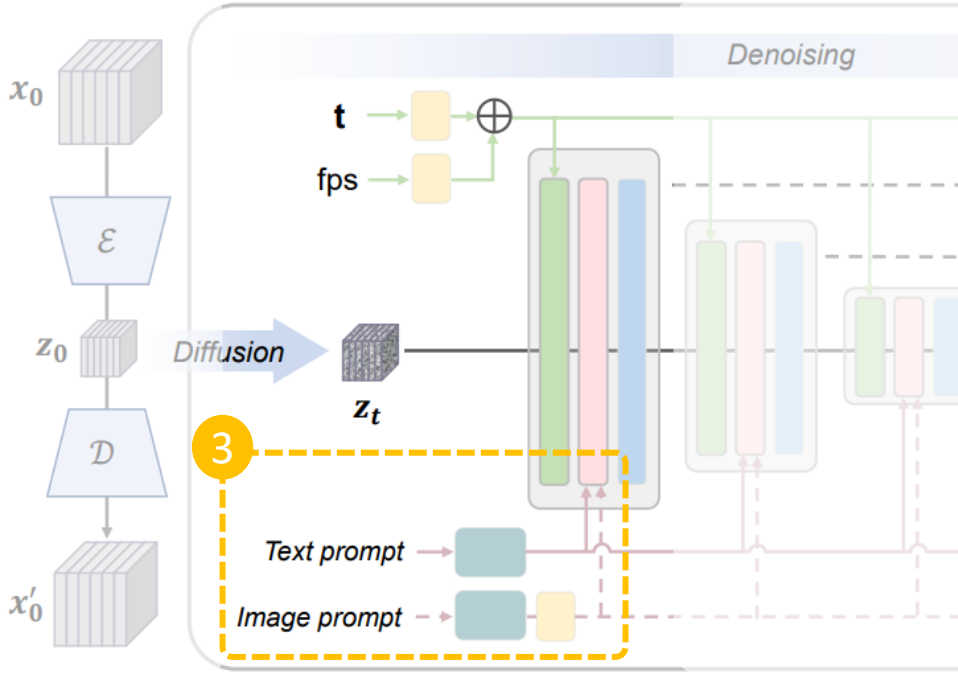
Figure 5. Image-conditioned text-to-video generation comparison. (a) Conditional image input. (b) Generation with the global semantic token conditioned. (c) Generation with the full patch visual tokens conditioned. The used text prompt is "a beautiful girl with colorful hair".

VideoCrafter1: Image To Video Model (T2V)

- Text Aligned Rich Image Embedding

$$\mathbf{F}_{out} = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}_{text}^{\top}}{\sqrt{d}})\mathbf{V}_{text} + \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}_{img}^{\top}}{\sqrt{d}})\mathbf{V}_{img} \quad (7)$$

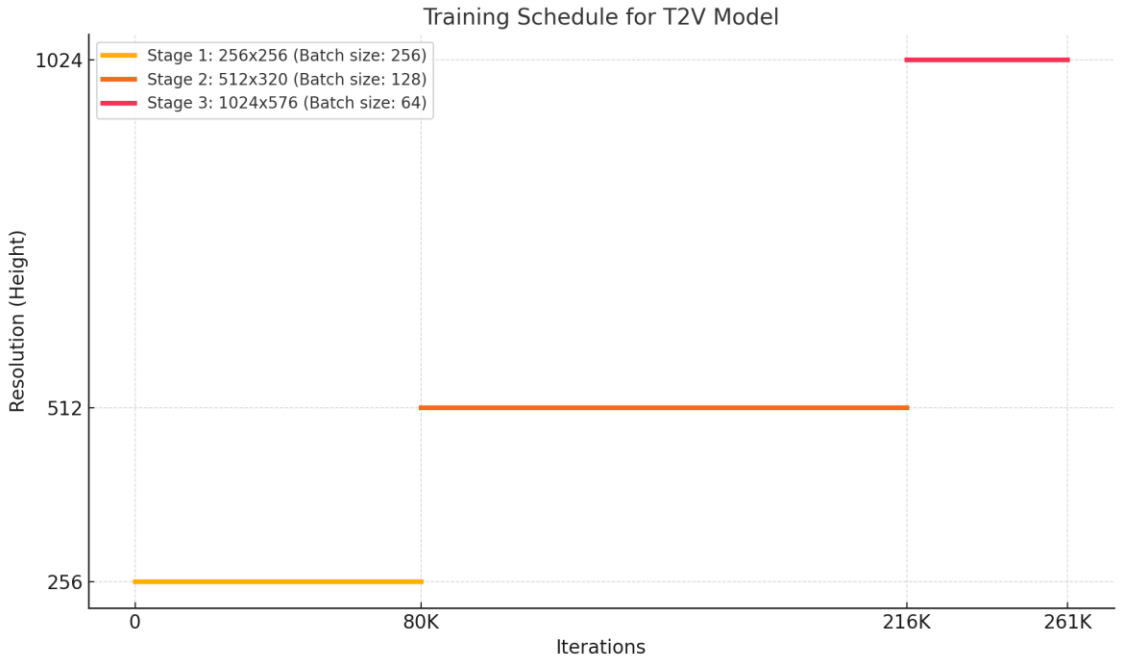




Experiment

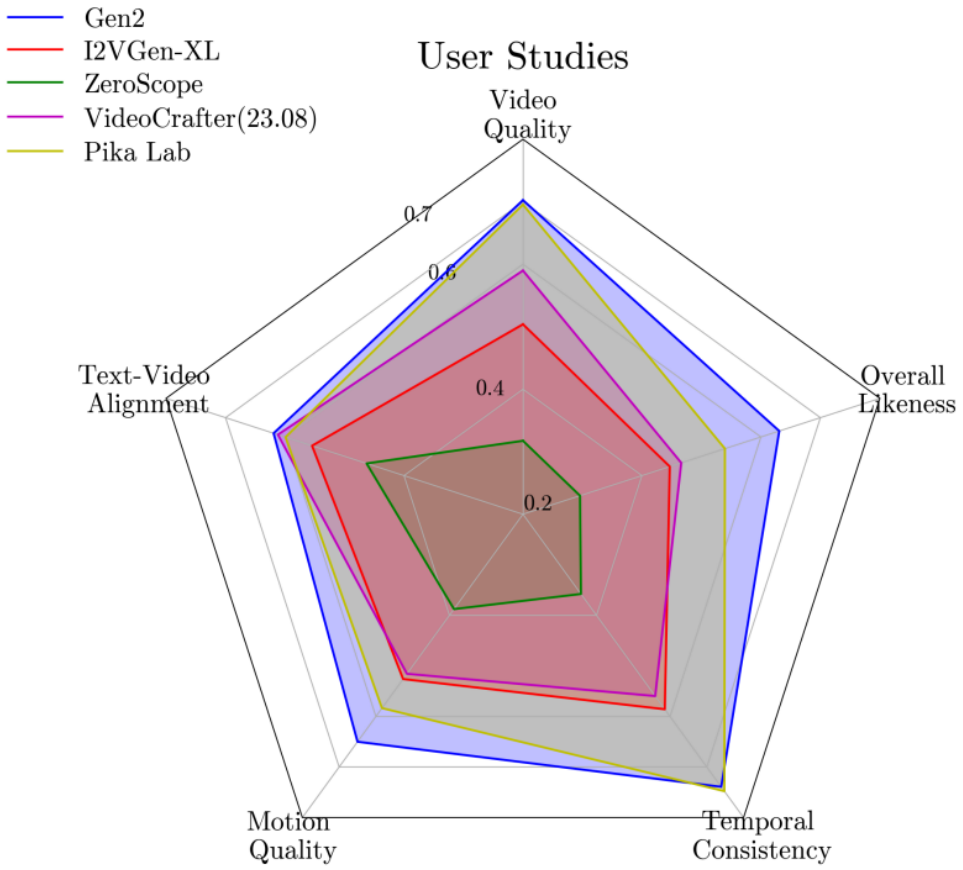
- Dataset
  - LAION COCO
    - A large text-image dataset
    - 600 million generated high-quality captions for web
  - WebVid-10M
    - A large scale short video with textual discriptions
    - 10 million
  - Custom Video Dataset
    - Large scale high-quality video
    - 10 million video with resolution greater than 1280 x 720
- Evaluation
  - EvalCrafter: a benchmark for evaluation video generation models
    - VideoCrafter1 vs. Gen-2
      Pika-Labs
      ModelSCopes

- Training Scheme
  - T2V Model
    - Employ training strategy used in Stable Diffusion
      - Training from low resolution to high resolution



Comparisons

- Benchmark: EvalCrafter

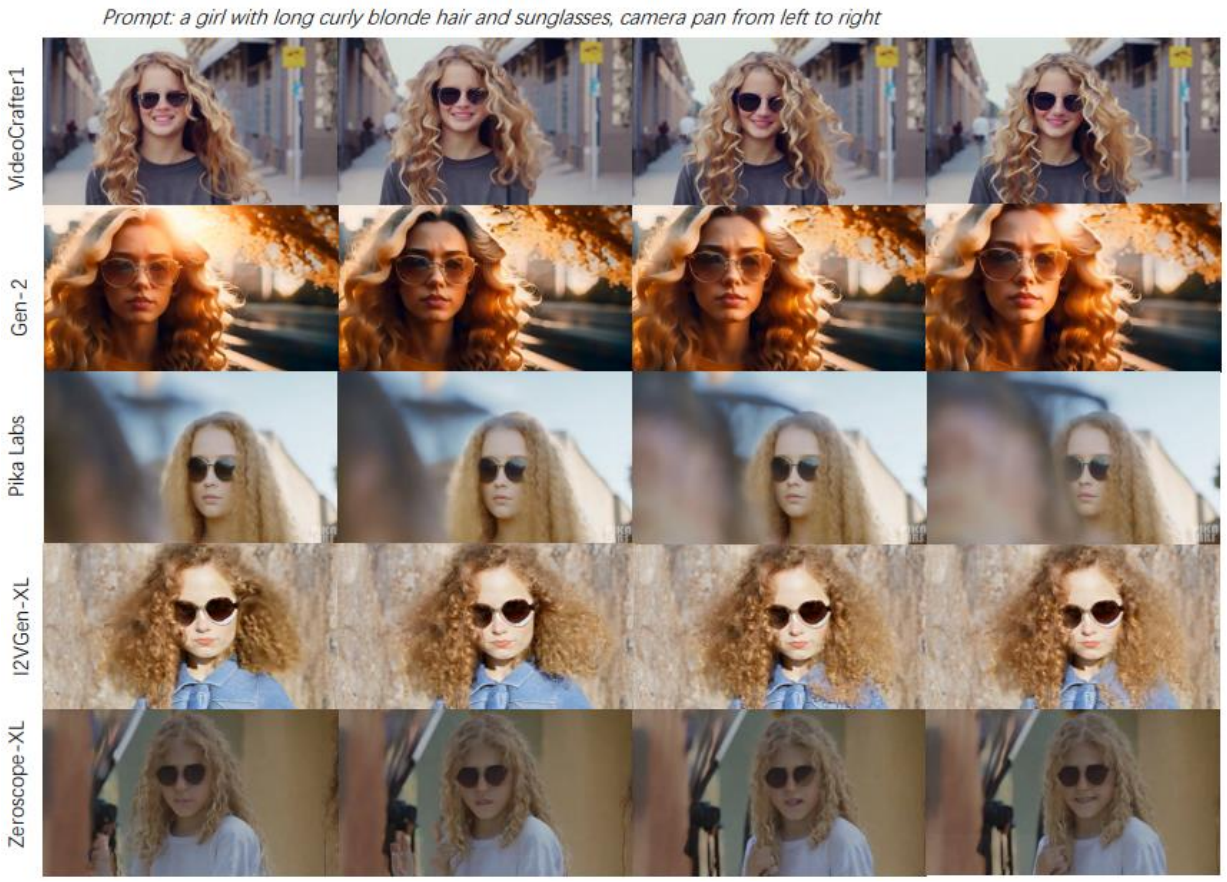


	Visual Quality	Text-Video Alignment	Motion Quality	Temporal Consistency
I2VGen-XL†	55.23	47.22	59.41	59.31
ZeroScope	56.37	46.18	54.26	61.19
PikaLab*	63.52	**54.11**	57.74	69.35
Gen2*	**67.35**	52.30	**62.53**	**69.71**
VideoCrafter$^{23.04}$	46.88	41.56	**56.24***	55.78
VideoCrafter$^{23.08}$	59.53	51.29	51.97	56.36
VideoCrafter$^{23.10}$	**61.64**	**66.76**	56.06	**60.36**

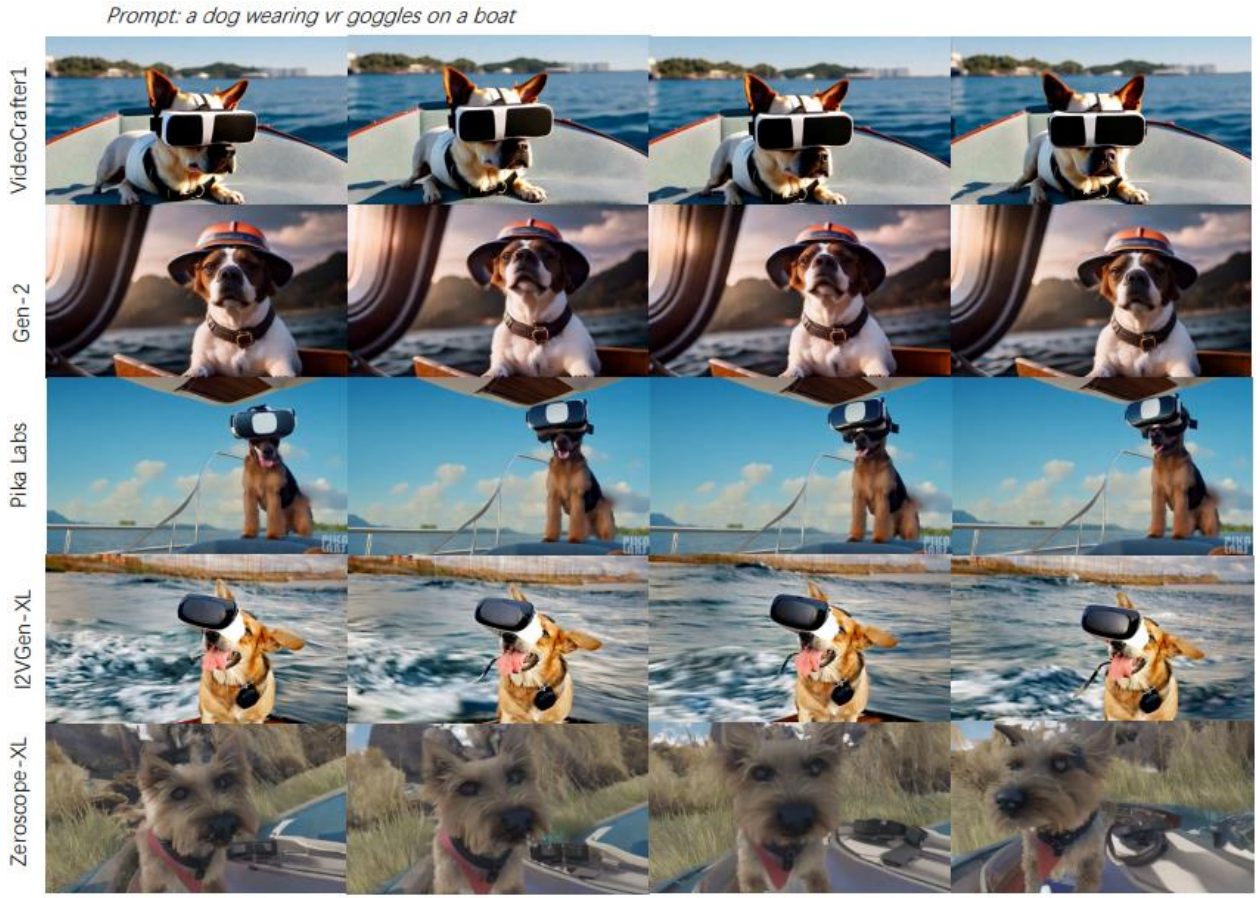
Table 1. Human-preference aligned results from four different aspects, with the rank of each aspect in the brackets. * indicated these models are not open-sourced.

Comparisons (Text-to-Video Model)

Prompt: a girl with long curly blonde hair and sunglasses, camera pan from left to right

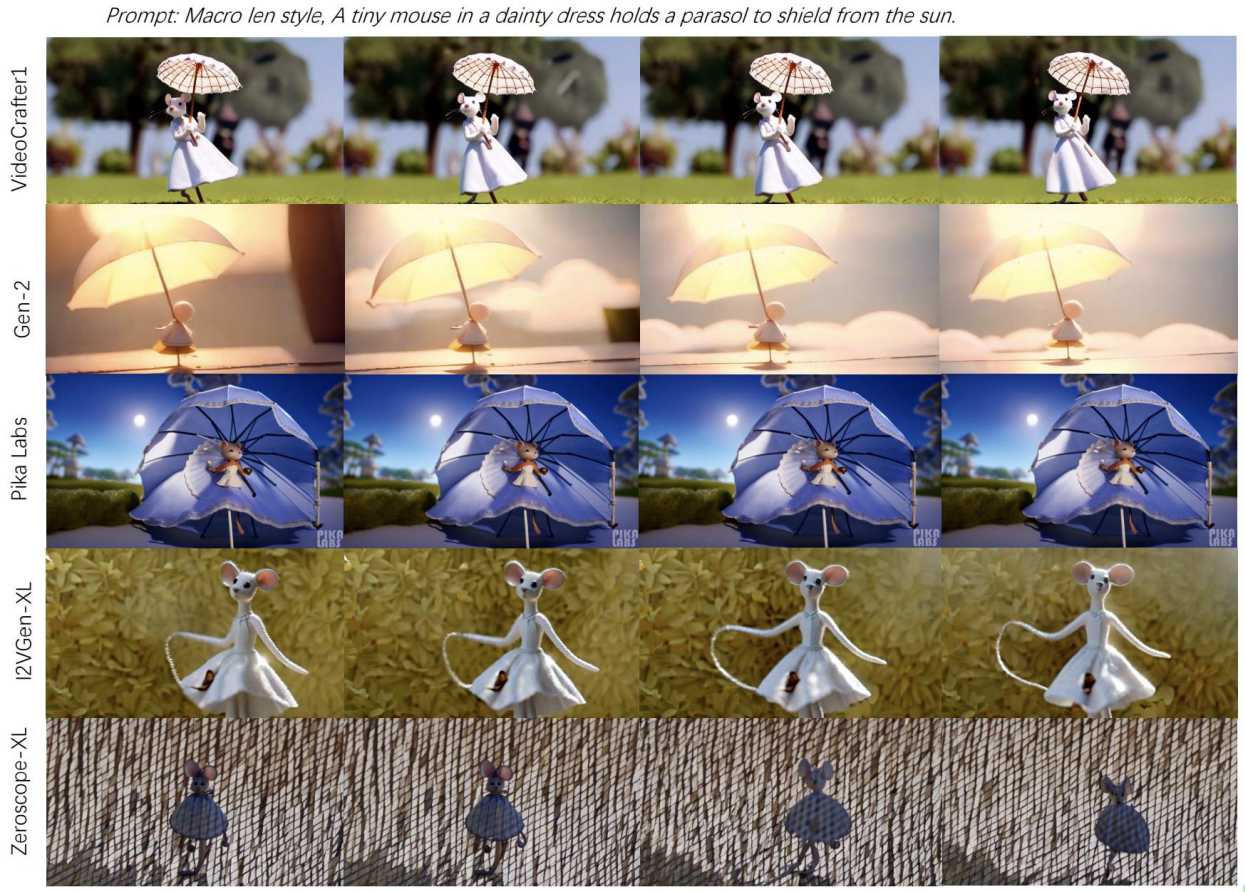


Prompt: a dog wearing vr goggles on a boat

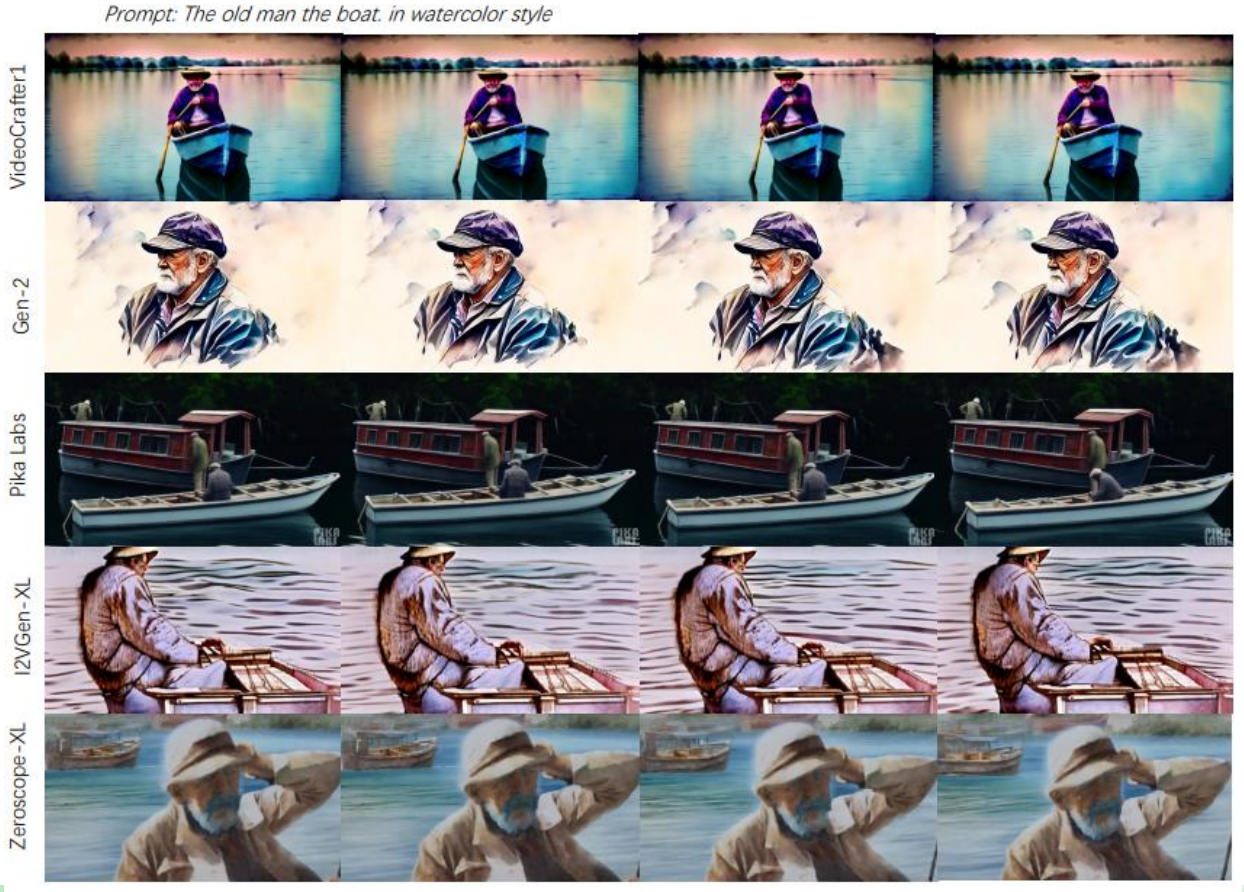


Comparisons (Text-to-Video Model)

Prompt: Macro len style, A tiny mouse in a dainty dress holds a parasol to shield from the sun.



Prompt: The old man the boat, in watercolor style



Comparisons (Image-to-Video Model)

Figure 9. Visual comparisons with image-to-video approaches: VideoComposer, I2VGen-XL, Pika, Gen-2 and our I2V model.