Text and Visual Generation based on Multimodal Pre-training
基于多模态预训练的文本和视觉生成

Nan DUAN (段楠)
Microsoft Research Asia
CNCC2021: 自然语言生成技术前沿与产业发展论坛
2021-12-18

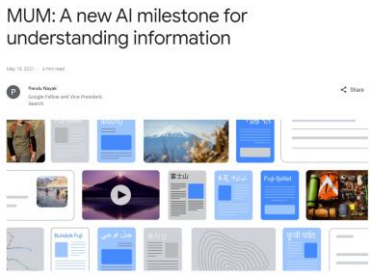
Visual contents become more important than ever before.

DALL·E: Creating Images from Text
OpenAI

01/21

05/21

MUM: Multitask Unified Model
Google



Metaverse
Facebook

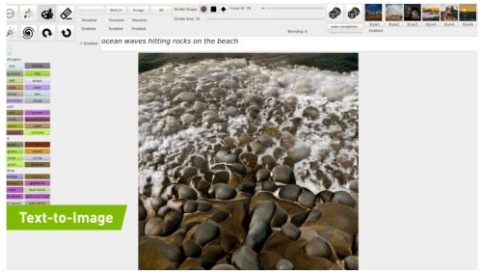
07/21

08/21

Omniverse
NVIDIA

Ego4D
Facebook

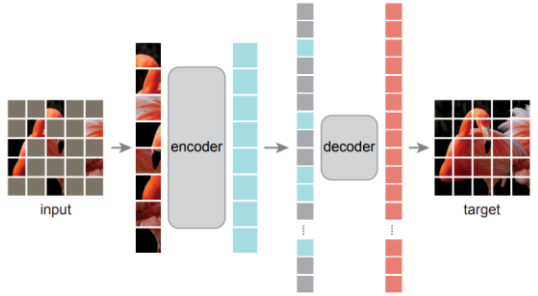
10/21



GauGAN AI Art
NVIDIA

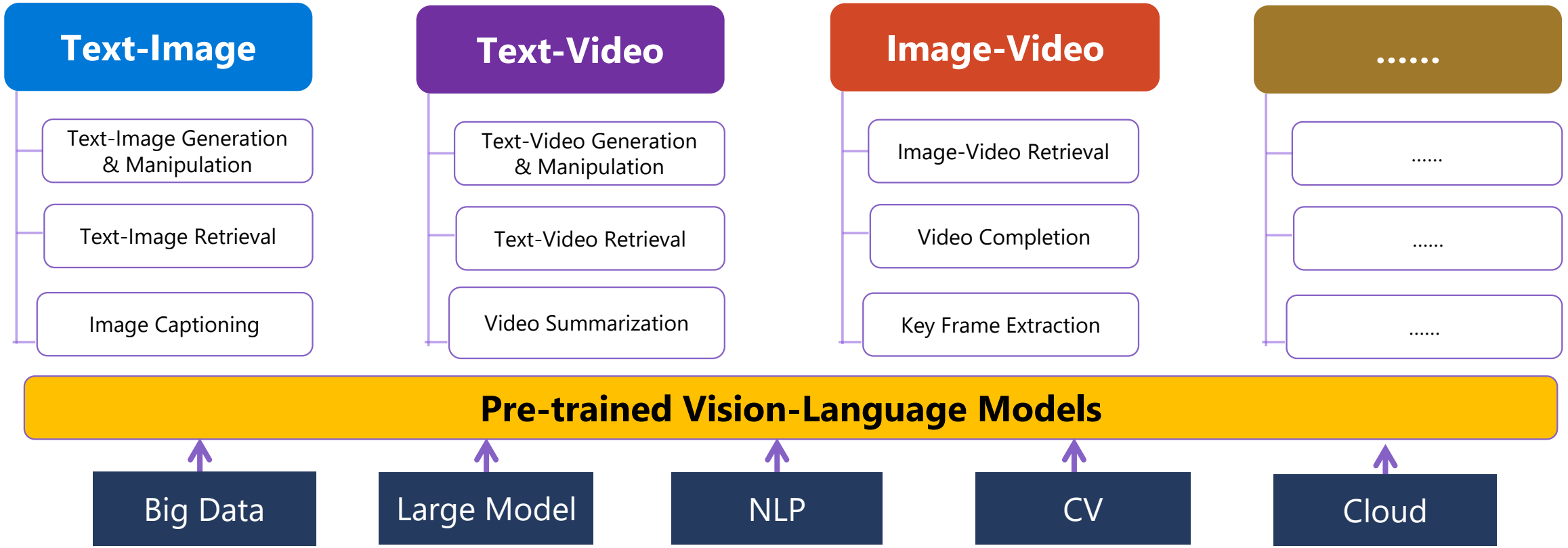
11/21

MAE
Facebook



Semantic-Driven Visual Content Creation Platform

Towards building large-scale pre-trained models to enable *controllable transformation* between text, image and video, and help creators to *improve* their content creation/editing *productivity*.



Part (1): Visual Synthesis

Text: There sea waves are beating the shore.

Text-to-Image

Image-to-Text

Video-to-Text

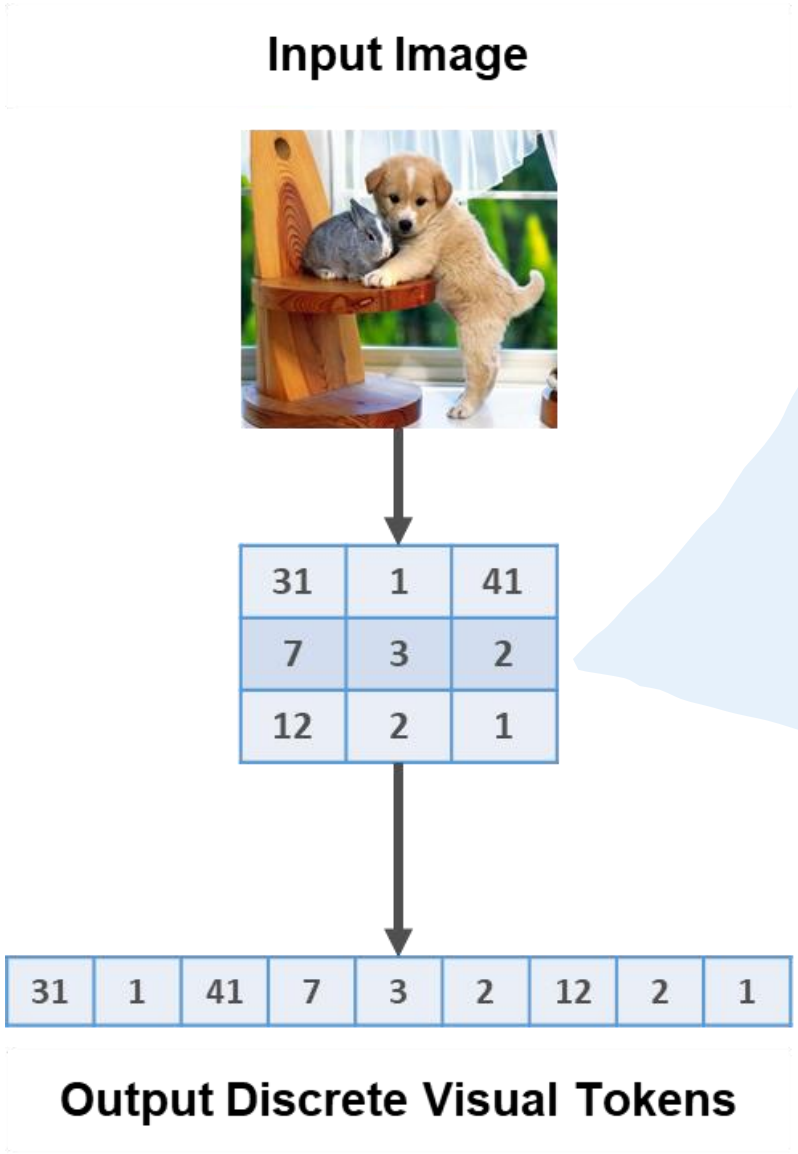
Text-to-Video

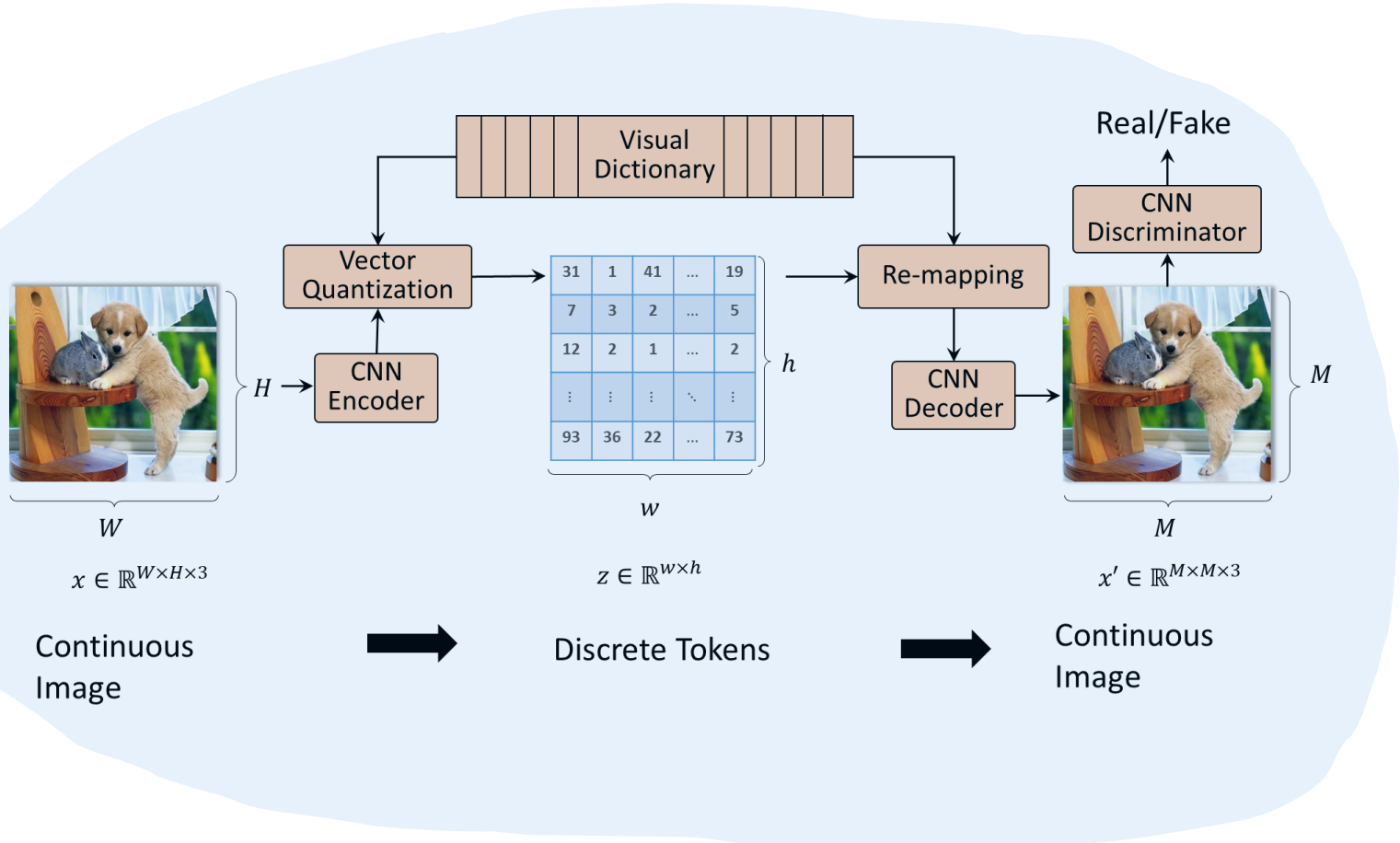
Image-to-Video

Image

Video

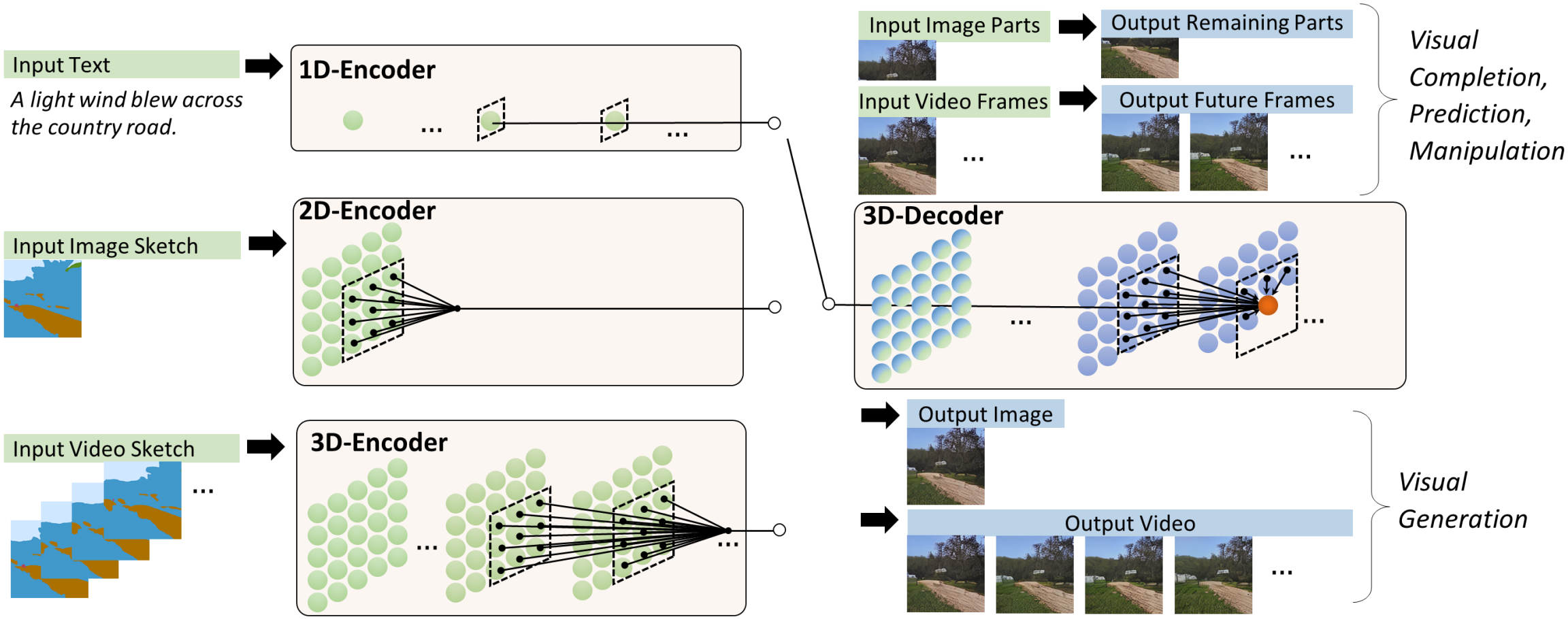
Discrete Visual Tokenization to Reuse NLP Technologies





VQ-VAE (Van den Oord et al., 2017), **VQ-GAN** (Esser et al., 2021)

A Unified Visual Synthesis Pre-training Framework

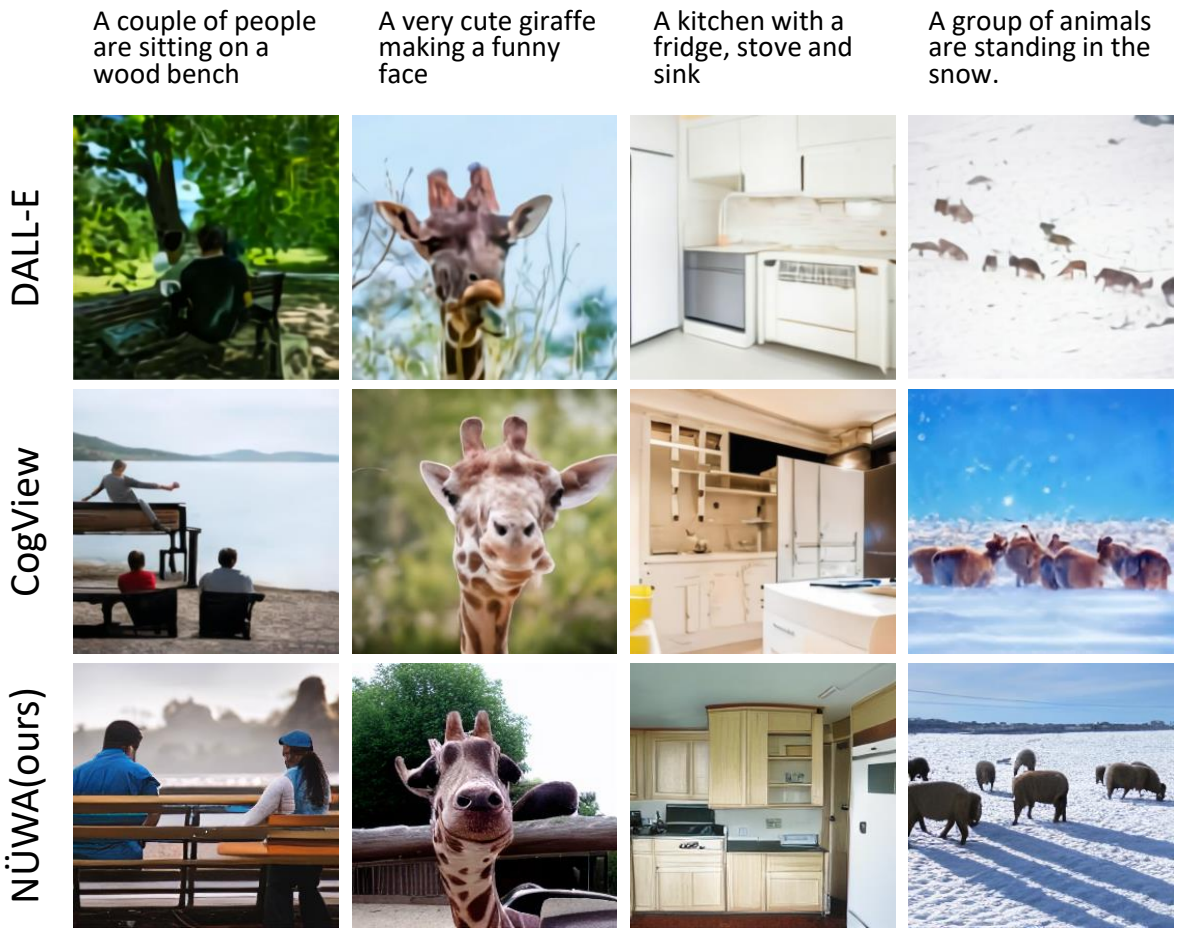


1. A general 3D transformer encoder-decoder framework is designed to cover language, image, and video.
2. A 3D sparse attention is proposed consider the characteristics of 3D data and reduce the computational complexity.
3. A multitask pre-training strategy using (large-scale) text-image pairs, (small-scale) text-video pairs, and (unlimited) videos.

Evaluation on Text-to-Image Generation

Table 1. Qualitative comparison with the state-of-the-art models for Text-to-Image (T2I) task on the MSCOCO dataset.

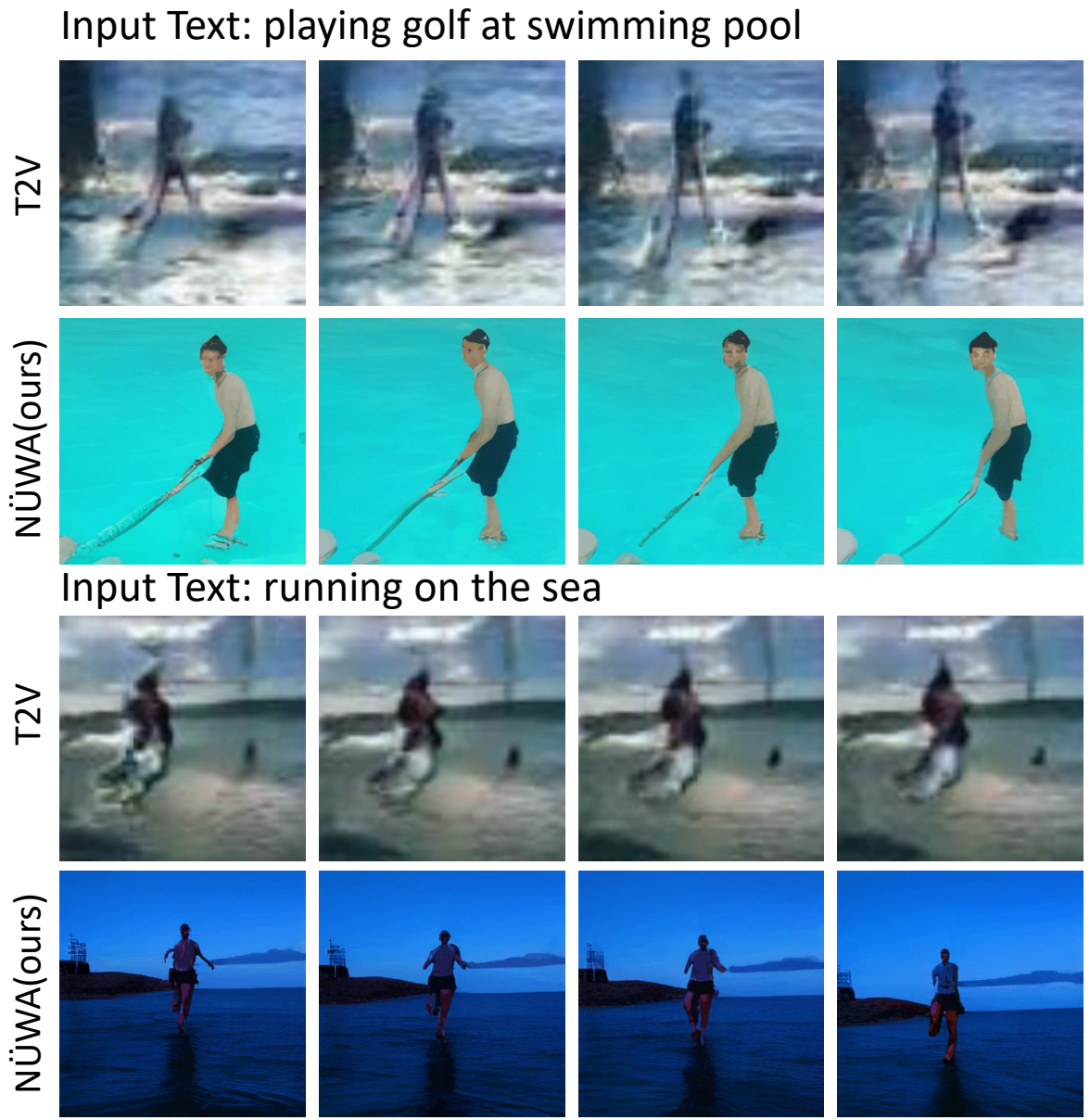
Model	FID-0↓	FID-1	FID-2	FID-4	FID-8	IS↑	CLIPSIM↑
AttnGAN [46]	35.2	44	72	108	100	23.3	0.2772
DM-GAN [49]	26	39	73	119	112.3	**32.2**	0.2838
DF-GAN [35]	26	33.8	55.9	91	97	18.7	0.2928
DALL-E [32]	27.5	28	45.5	83.5	85	17.9	-
CogView [9]	27.1	19.4	**13.9**	19.4	**23.6**	18.2	0.3325
NÜWA	**12.9**	**13.8**	15.7	**19.3**	24	27.2	**0.3429**



Evaluation on Text-to-Video Generation

Table 2. Quantitative comparison with state-of-the-art models for Text-to-Video (T2V) task on Kinetics dataset.

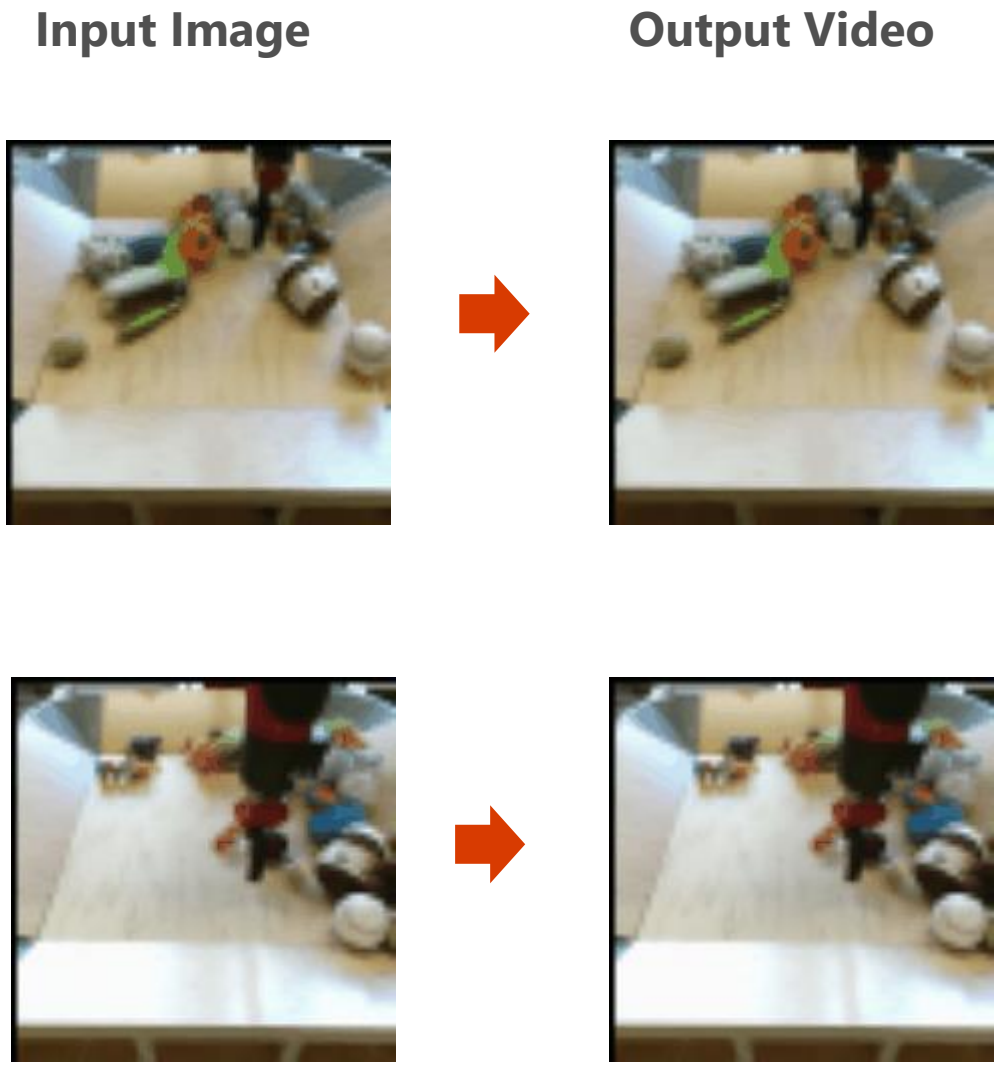
Model	Acc↑	FID-img↓	FID-vid↓	CLIPSIM↑
T2V [20]	42.6	82.13	14.65	0.2853
SC [2]	74.7	33.51	7.34	0.2915
TFGAN [2]	76.2	31.76	7.19	0.2961
NÜWA	**77.9**	**28.46**	**7.05**	**0.3012**



Evaluation on Video Prediction

Table 3. Quantitative comparison with state-of-the-art models for Video Prediction (V2V) task on BAIR (64×64) dataset. Cond. denotes the number of frames given to predict future frames.

Model	Cond.	FVD↓
MoCoGAN [36]	4	503
SVG-FP [8]	2	315
CNDA [12]	2	297
SV2P [1]	2	263
SRVP [13]	2	181
VideoFlow [17]	3	131
LVT [30]	1	126±3
SAVP [19]	2	116
DVD-GAN-FP [7]	1	110
Video Transformer (S) [43]	1	106±3
TriVD-GAN-FP [22]	1	103
CCVS [24]	1	99±2
Video Transformer (L) [43]	1	94±2
NÜWA	1	**86.9**



Ablation Study

Table 5. Effectiveness of multi-task pre-training for Text-to-Video (T2V) generation task on MSRVTT dataset.

Model	Pre-trained Tasks	FID-vid↓	CLIPSIM↑
NÜWA-TV	T2V	52.98	0.2314
NÜWA-TV-TI	T2V+T2I	53.92	0.2379
NÜWA-TV-VV	T2V+V2V	51.81	0.2335
NÜWA	T2V+T2I+V2V	**47.68**	**0.2439**

Table 6. Effectiveness of 3D nearby attention for Sketch-to-Video (S2V) task on VSPW dataset.

Model	Encoder	Decoder	FID-vid↓	Detected PA↑
NÜWA-FF	Full	Full	35.21	0.5220
NÜWA-NF	Nearby	Full	33.63	0.5357
NÜWA-FN	Full	Nearby	32.06	0.5438
NÜWA-AA	Axis	Axis	29.18	0.5957
NÜWA	Nearby	Nearby	**27.79**	**0.6085**

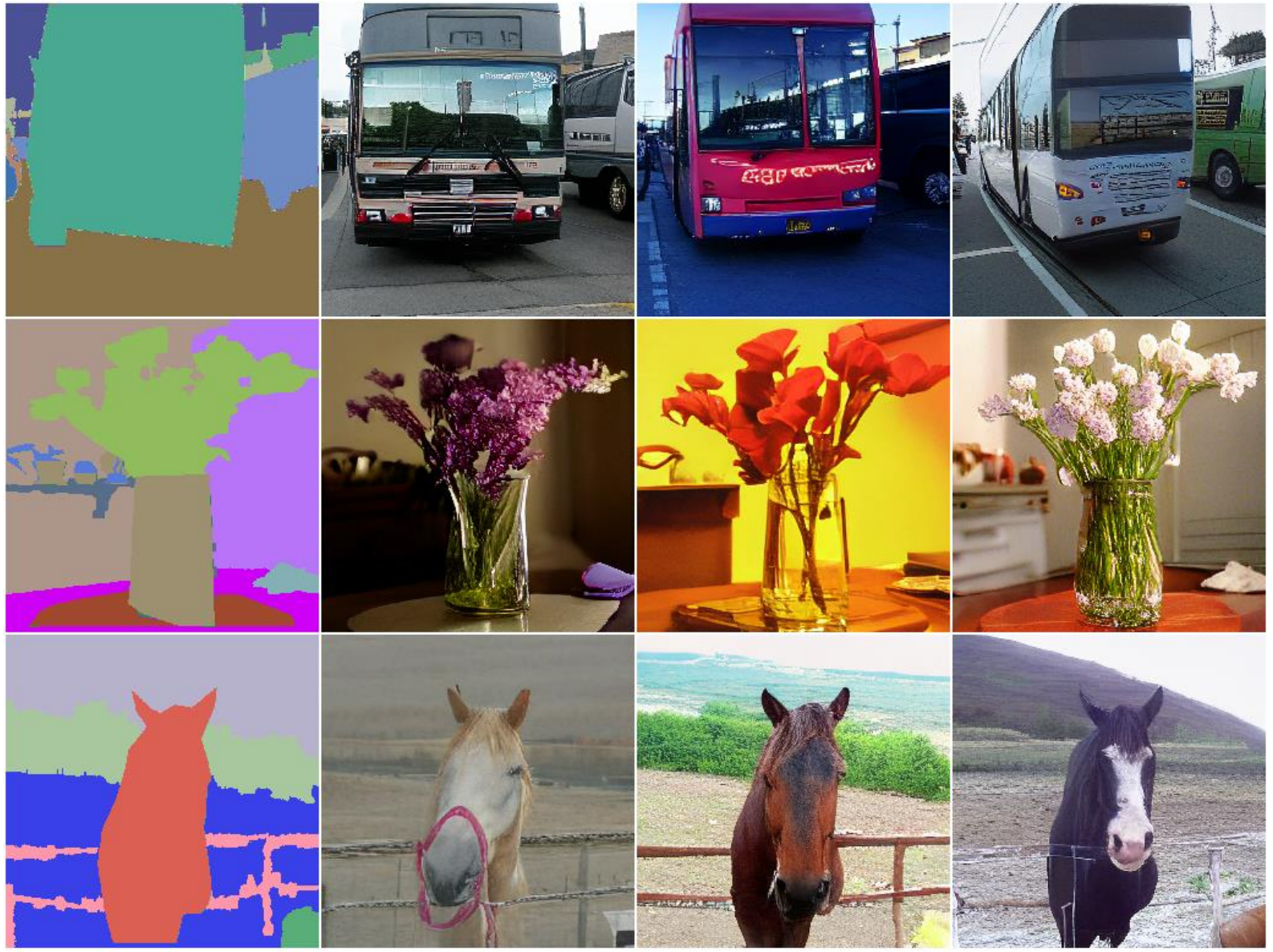
Example: Text-to-Image

A wooden house sitting in a field.

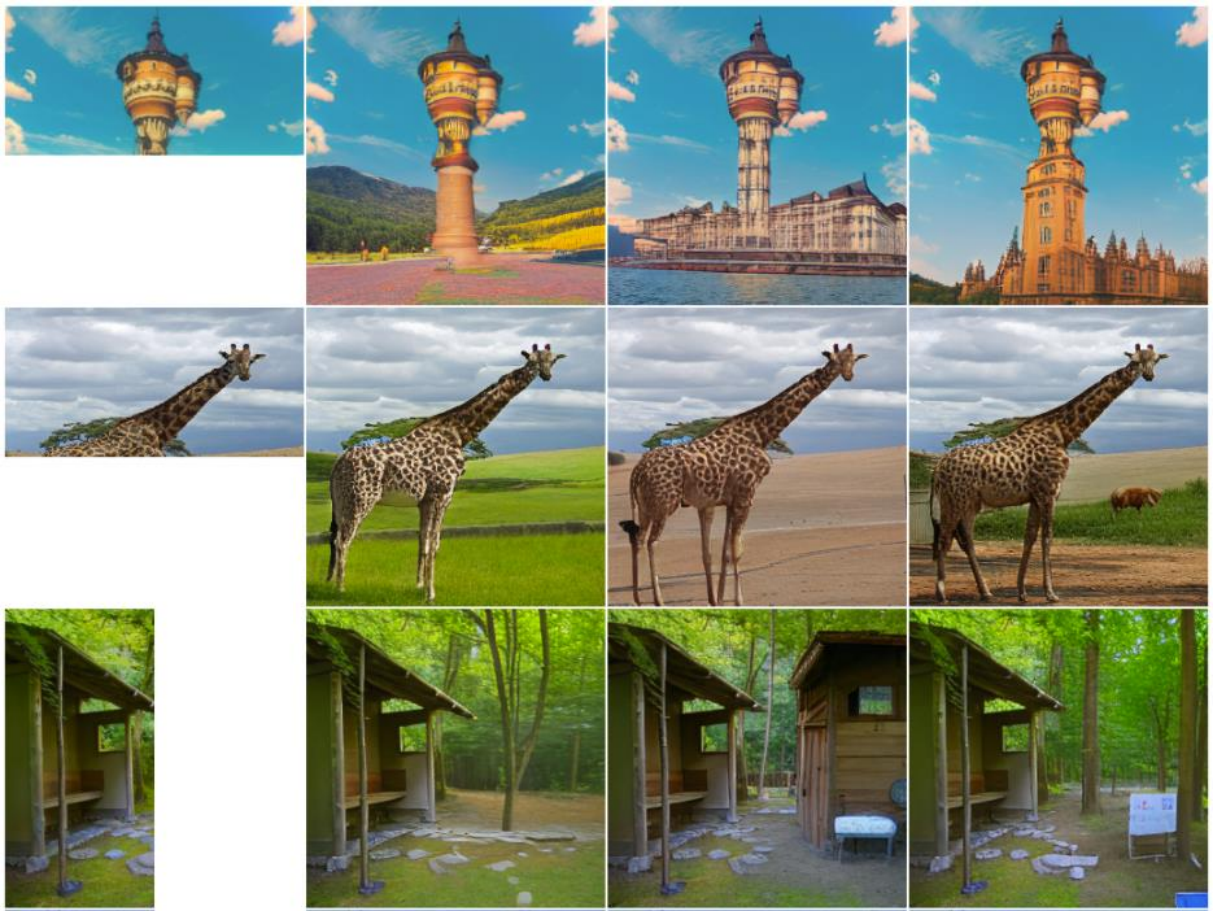
A young girl eating a very tasty looking slice of pizza.

Walnuts are being cut on a wooden cutting board.

Example: Sketch-to-Image

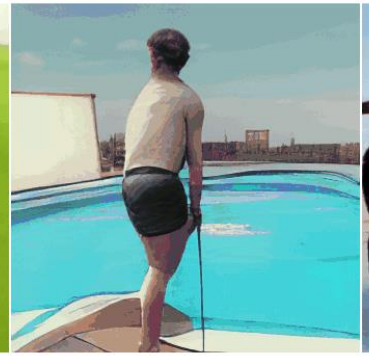
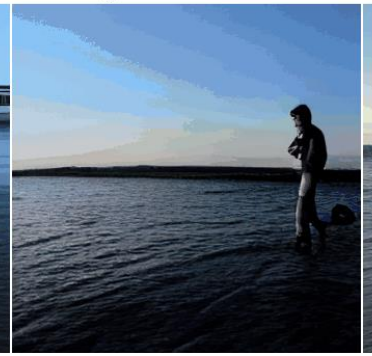
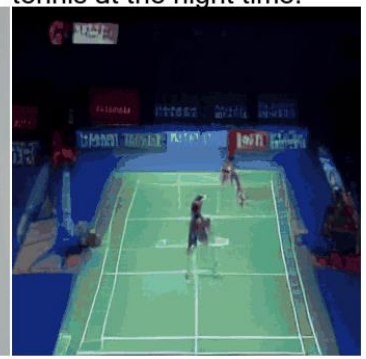


Example: Image Completion

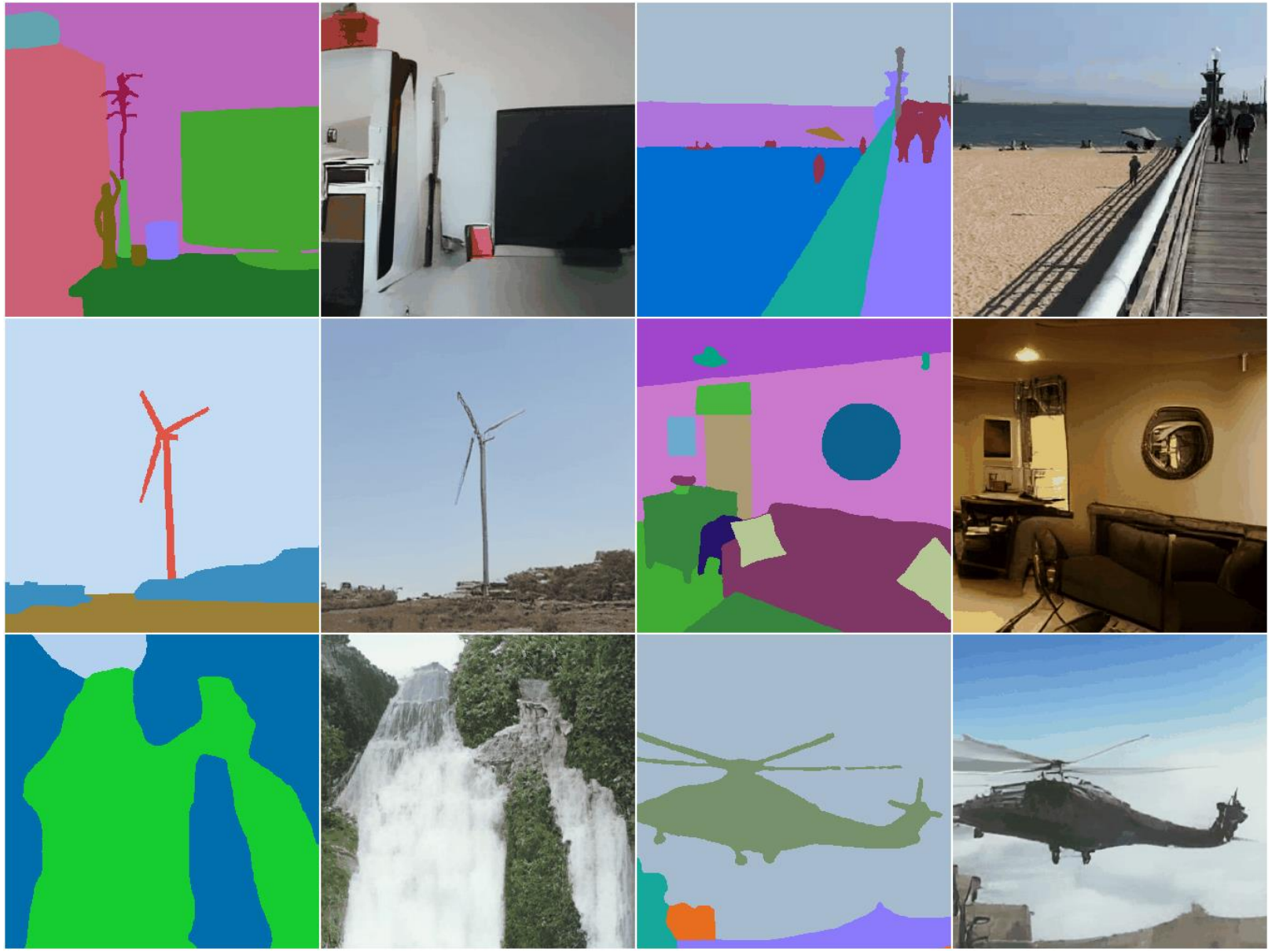


Example: Text-Guided Image Manipulation (Zero-Shot)

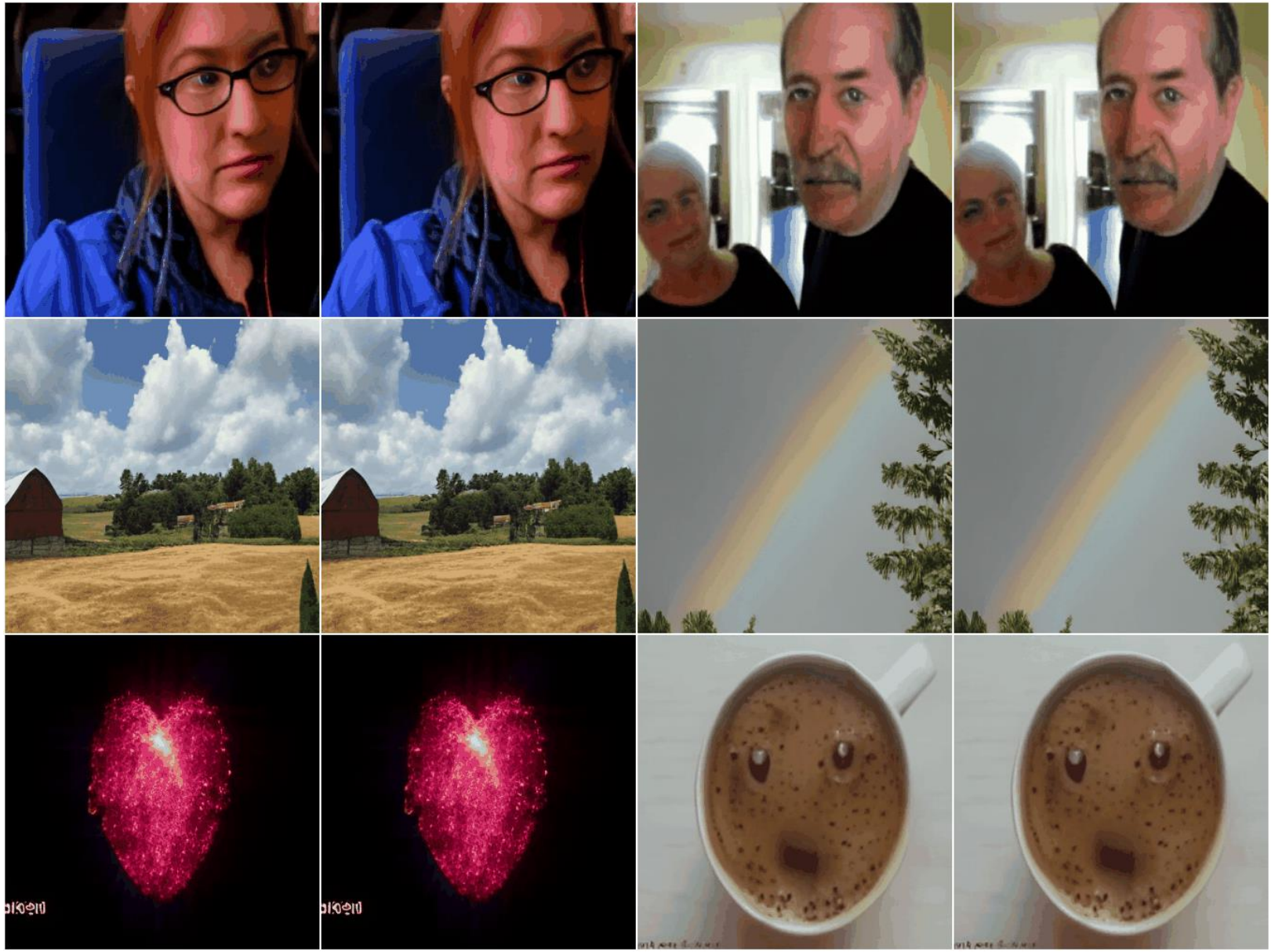
Example: Text-to-Video

Play golf on grass.	Play golf at swimming pool.	Play golf at swimming pool.	Play golf at swimming pool.
			
Sailing on the sea.	Running on the sea.	Running on the sea.	Running on the sea.
			
A suit man is talking from a studio with fun.	The whilte sailboat sailed on the sea.	A man is folding a piece of yellow paper.	Tennis players wearing blue and red t-shirts are playing tennis at the night time.
			

Example: Sketch-to-Video



Example: Video Prediction



Example: Text-Guided Video Manipulation (Zero-Shot)

Part (2): Visual Captioning

Text: There sea waves are beating the shore.

Text-to-Image

Image-to-Text

Video-to-Text

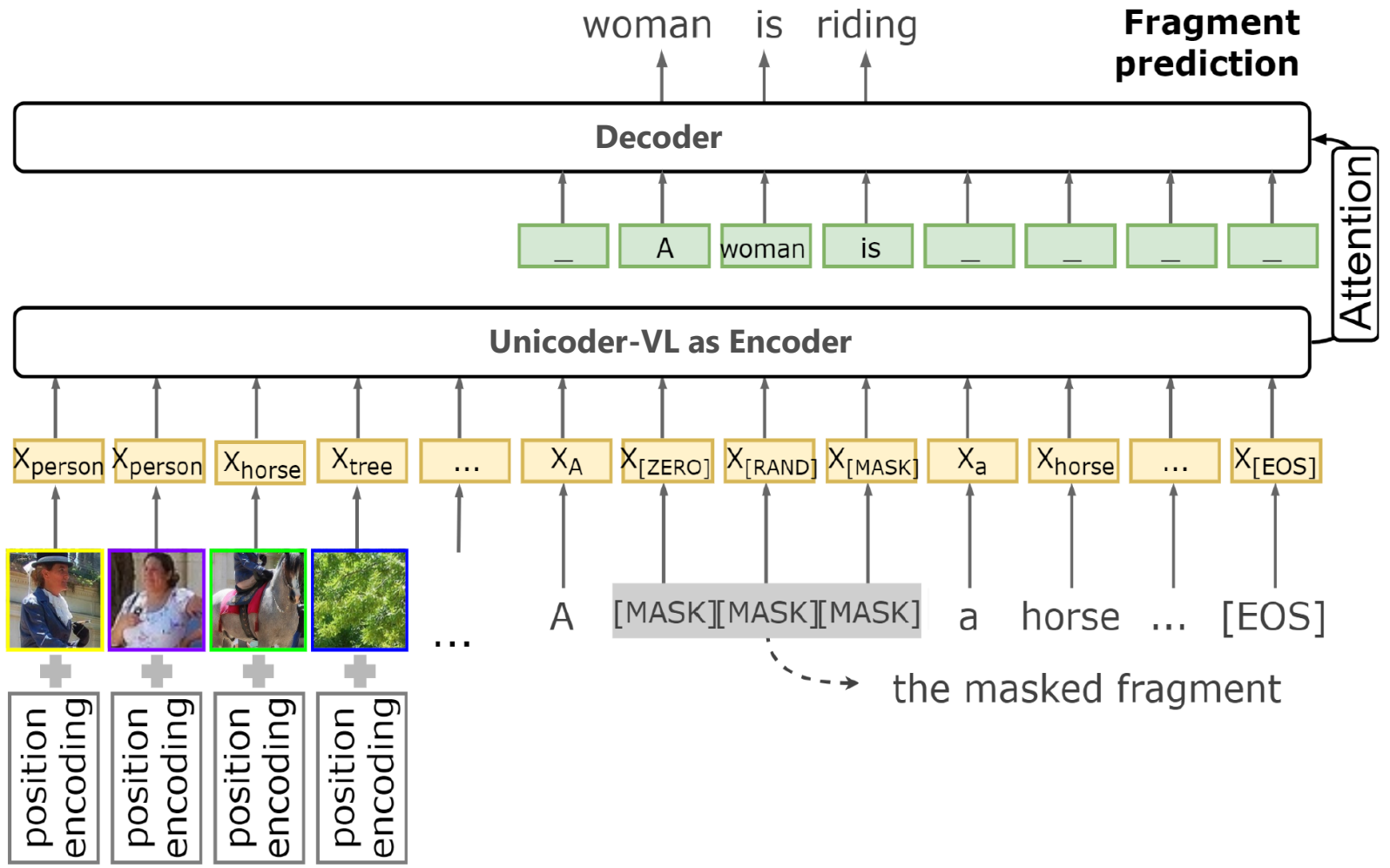
Text-to-Video

Image-to-Video

Image

Video

Image Captioning



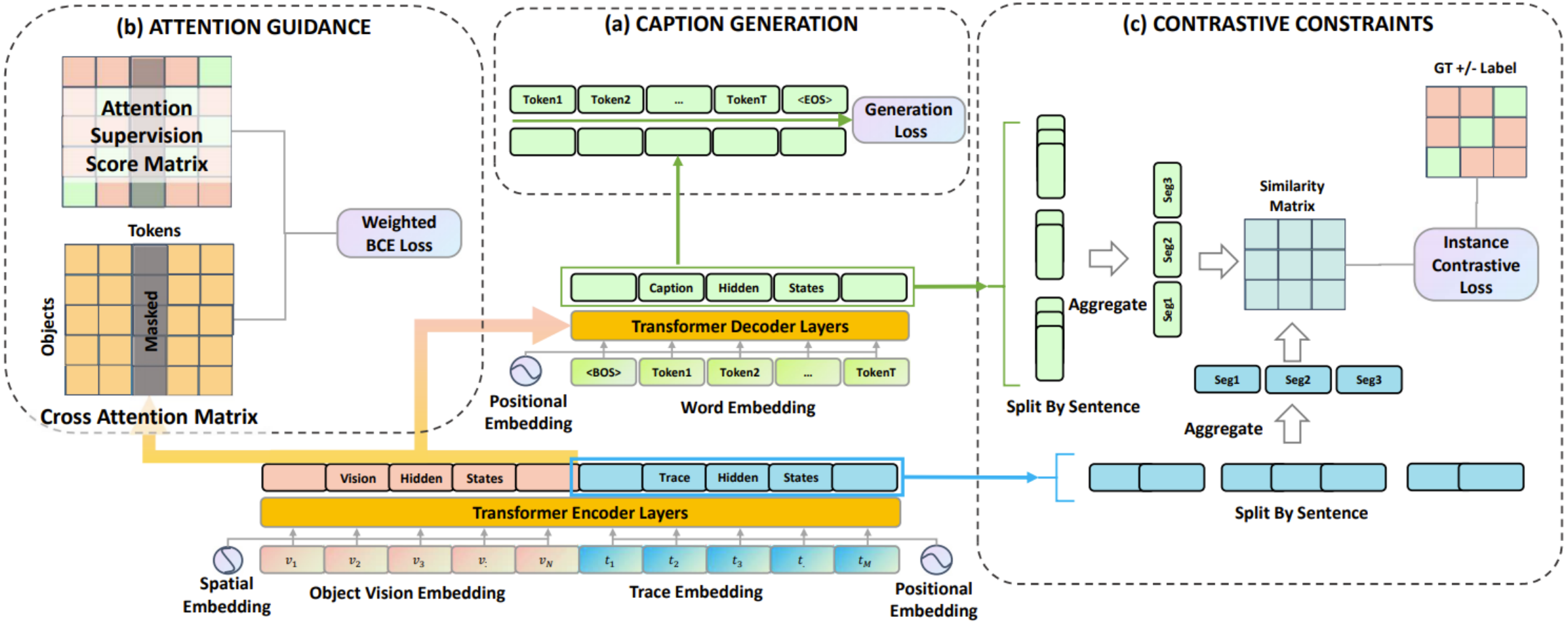
Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, Xin Liu, Ming Zhou. **XGPT: Cross-modal Generative Pre-Training for Image Captioning**. NLPCC 2021.

Localized Narratives: A Trace-based Image Captioning Dataset

In this picture there is a stand on a ground. On the backside there is a person. He is riding on a horse. He is wearing a cap. He is in between the fence. There is a flags on a wall. On the left side there is a score board on a table and flower plants. We can see in the background sky and trees.

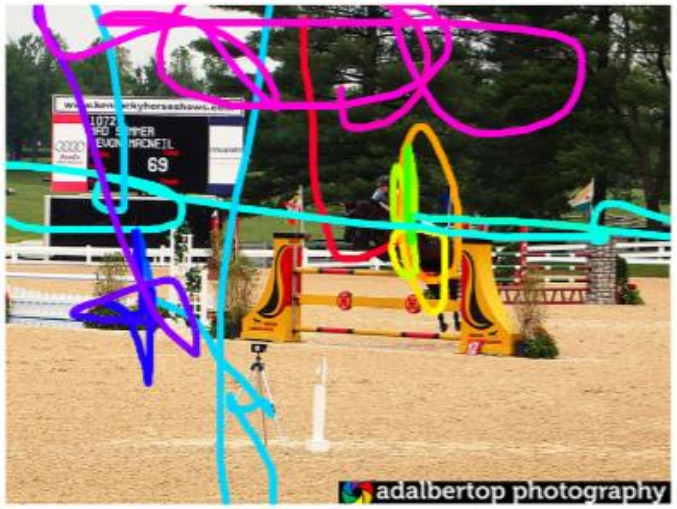
Jordi Pont-Tuset, Jasper Uijlings, Beer Changpinyo, Radu Soricut, Vittorio Ferrari. **Connecting vision and language with localized narratives**. ECCV 2020.

Image Captioning with Trace



Kun Yan, Lei Ji, Huaishao Luo, Ming Zhou, Nan Duan, Shuai Ma. **Control Image Captioning Spatially and Temporally**. ACL 2021.

Evaluation on Localized Narratives



Method	ROUGE-L	ROUGE-1-F1	BLEU-1	BLEU-4	CIDEr-D	METEOR
Baseline(Pont-Tuset et al., 2020)	31.7	47.9	32.2	8.1	29.3	-
+Trace(Pont-Tuset et al., 2020)	48.3	60.7	52.2	24.6	106.5	-
Baseline*	34.1	54.0	36.0	10.3	29.5	16.4
+Trace*	49.0	68.1	55.4	25.0	107.9	25.2
LoopCAG(our)	**50.3**	**69.8**	**57.2**	**27.0**	**114.0**	**26.0**

Table 1: Comparison with baseline methods results: Baseline means an encoder-decoder model without taking trace as input. +Trace means concatenating encoded trace feature to the encoder input, i.e., trace controlled caption performance. LoopCAG is our complete model. The results with * are the baseline performance re-implemented by ourselves

Ground Truth: In this picture there is a stand on a ground. On the backside there is a person. He is riding on a horse. He is wearing a cap. He is in between the fence. There is a flags on a wall. On the left side there is a score board on a table and flower plants. We can see in the background sky. trees.

Baseline: In this image I can see a horse which is in white color, at left there is a person sitting on the horse, at the background there are some people standing, in the background there are few buildings, trees and sky.

LoopCAG: There is a person sitting on a horse. He is holding a horse thread and he is wearing a cap. There are flags, board on the left side. We can see in the background sky, trees.

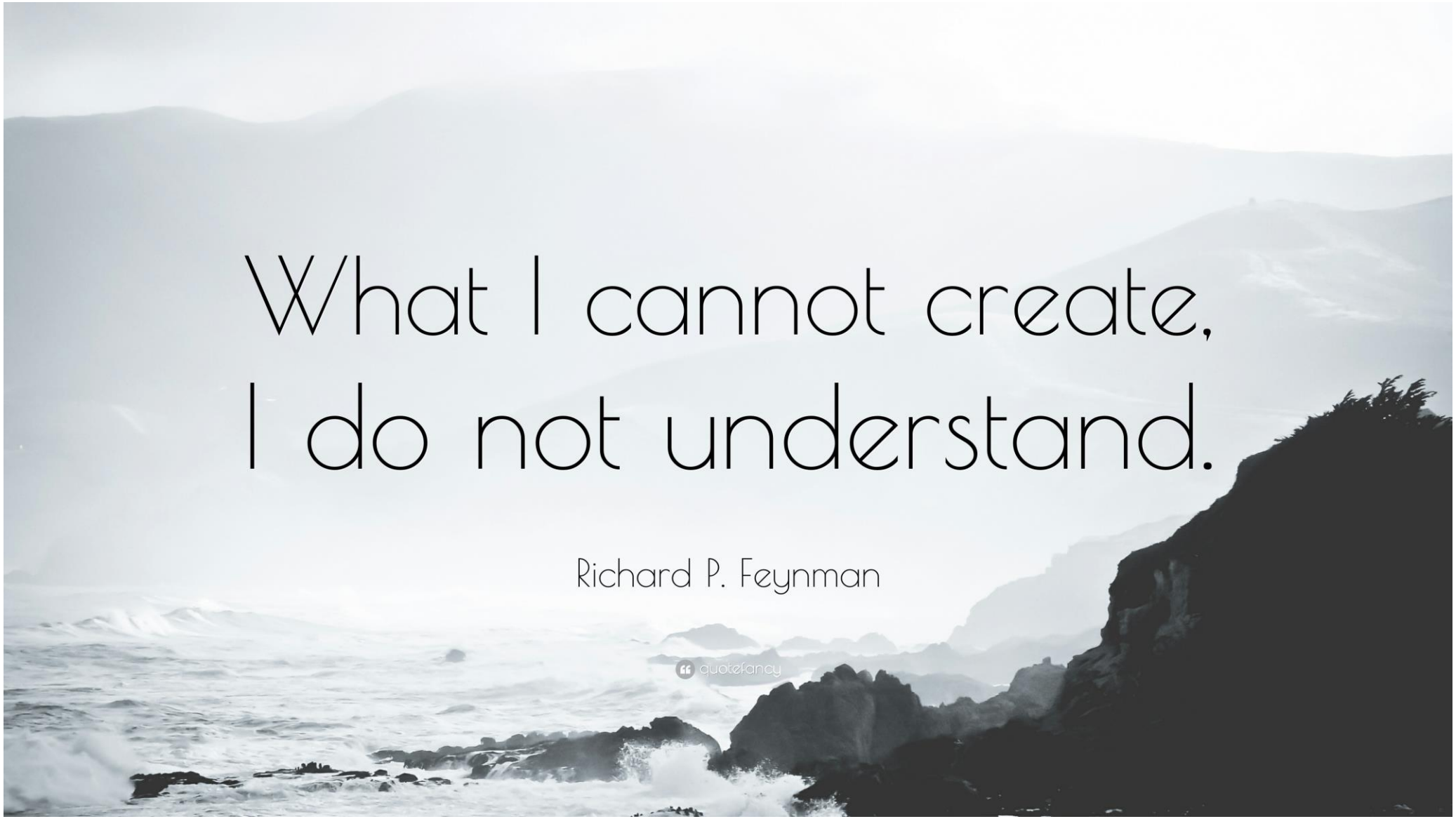
Summary

- **NLP and CV trend to be unified.**
  - Similar backbones (Transformer)
  - Similar representation formats (textual tokens and discrete visual tokens)
  - Similar pre-training tasks (auto-regressive decoding, denoising auto-encoding, contrastive learning)

- **Multimodal AI becomes the innovation frontier.**
  - Learn universal vision-language representations (pre-training)
  - Retrieve image/video from text, and vice versa (understanding)
  - Generate image/video from text, and vice versa (generation)
  - Learn commonsense knowledge from visual corpus (knowledge acquisition)
  - Design new benchmarks and metrics for multimodal AI tasks (evaluation)
  - Design new model visualization and interpretation mechanisms (interpretability)

- **Visual content creation opens a new door to AI applications.**
  - Multimodal search engine/question answering/dialogue system
  - Visual advertisement/ppt/news creation
  - AI-assisted image/video editing
  - AR/VR/Metaverse/etc.

Thank you!


What I cannot create,
I do not understand.
Richard P. Feynman
quotefancy