LLM *Meets* Diffusion

LLM-Empowered Text-to-Vision Diffusion Models

大语言模型赋能的文本到视觉扩散模型研究

NExT++ Research Center, NUS

Hao Fei (费豪), Research Fellow

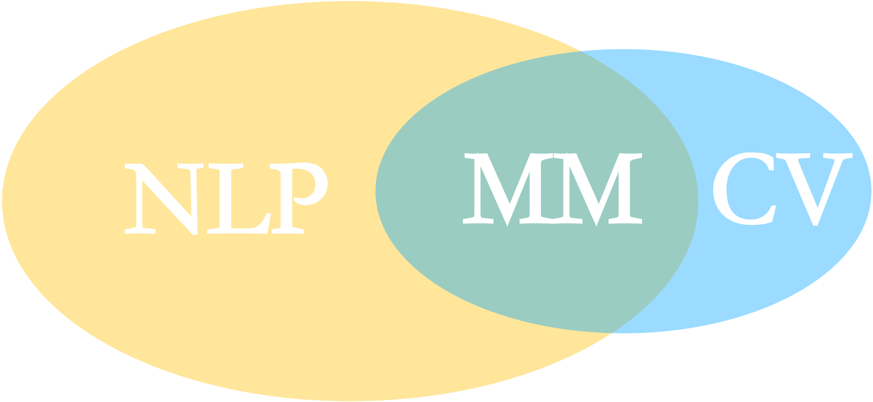
Aug. 25th 2023

https://haofei.vip/

Self-introduction

Research Directions

- Areas



- Angle of interests
  - **Structure-aware** Intelligence Learning (SAIL)

Self-introduction

Research Directions

- Structure-aware NLP
  - Low-level Syntax/Discourse Parsing
    - Dependency/Constituency parsing, Document/Dialogue discourse parsing, etc.
  - High-level Semantic Structure Parsing
    - Semantic analysis, Information extraction, Structured sentiment analysis, etc.
  - Structure-based language modeling/understanding

- Structure-aware MM
  - Multimodal Grammar Induction/Scene Graph (SG) Parsing
  - SG-based Multimodal (Visual-Language) Learning

- Structure-aware LLM
  - Structure-based (World Modeling) Language Modeling
  - Structure-based Reasoning with LLM

CONTENT

1 LLM Meets Diffusion

2 LLM-Empowered Text-to-Image Diffusion

3 LLM-Empowered Text-to-Vision Diffusion

4 Summary and Outlook

CONTENT

1 LLM Meets Diffusion

2 LLM-Empowered Text-to-Image Diffusion

3 LLM-Empowered Text-to-Vision Diffusion

4 Summary and Outlook

Large Language Models

(Near) Human-level Language Understanding via LLMs

- LLMs → **Artificial General Intelligence (AGI) ?**

OpenAI

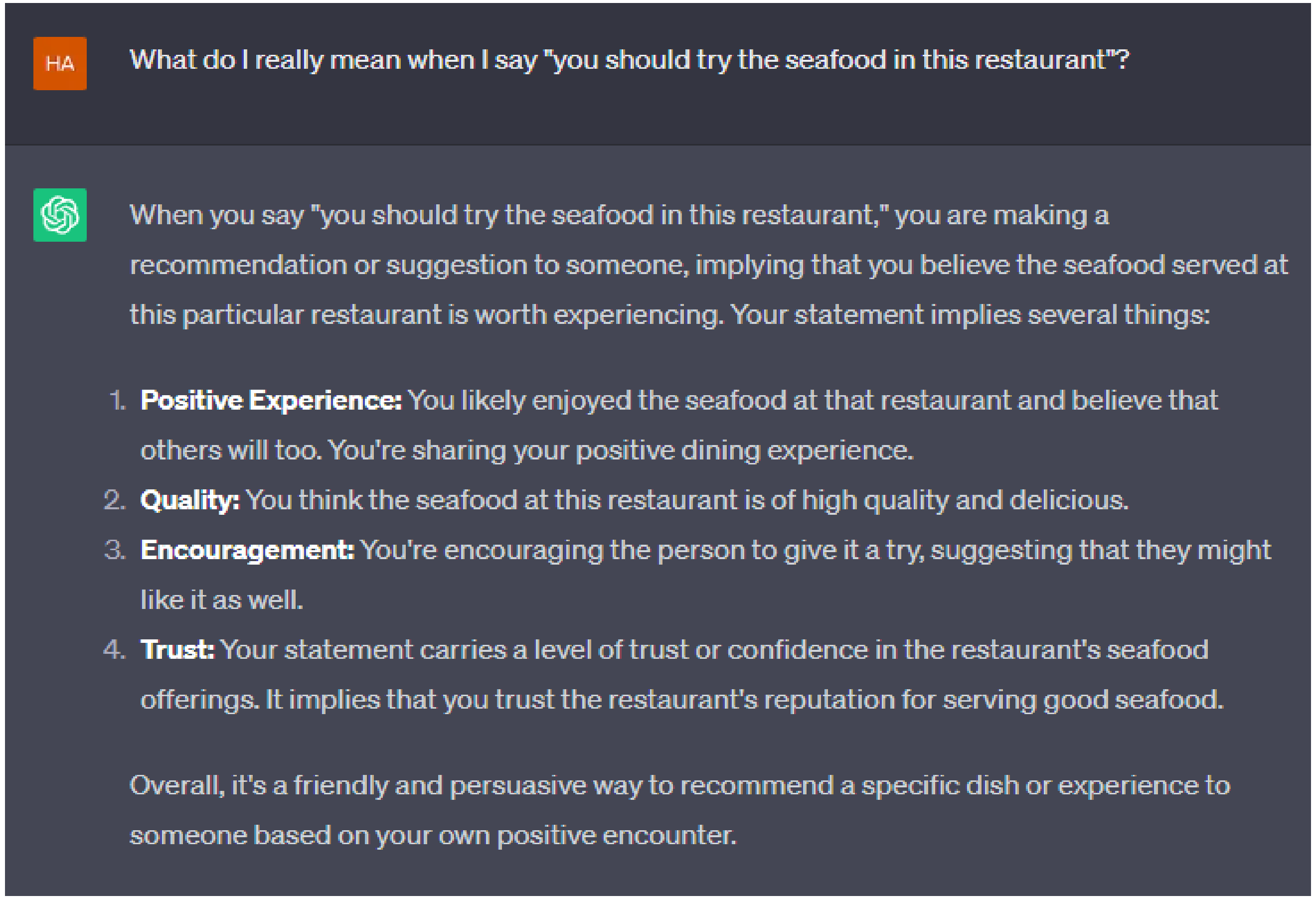
CHAT-GPT

The entire world is witnessing the impact of LLMs for leading to AGI!

Large Language Models

(Near) Human-level Language Understanding via LLMs

Understanding the underlying semantic intent

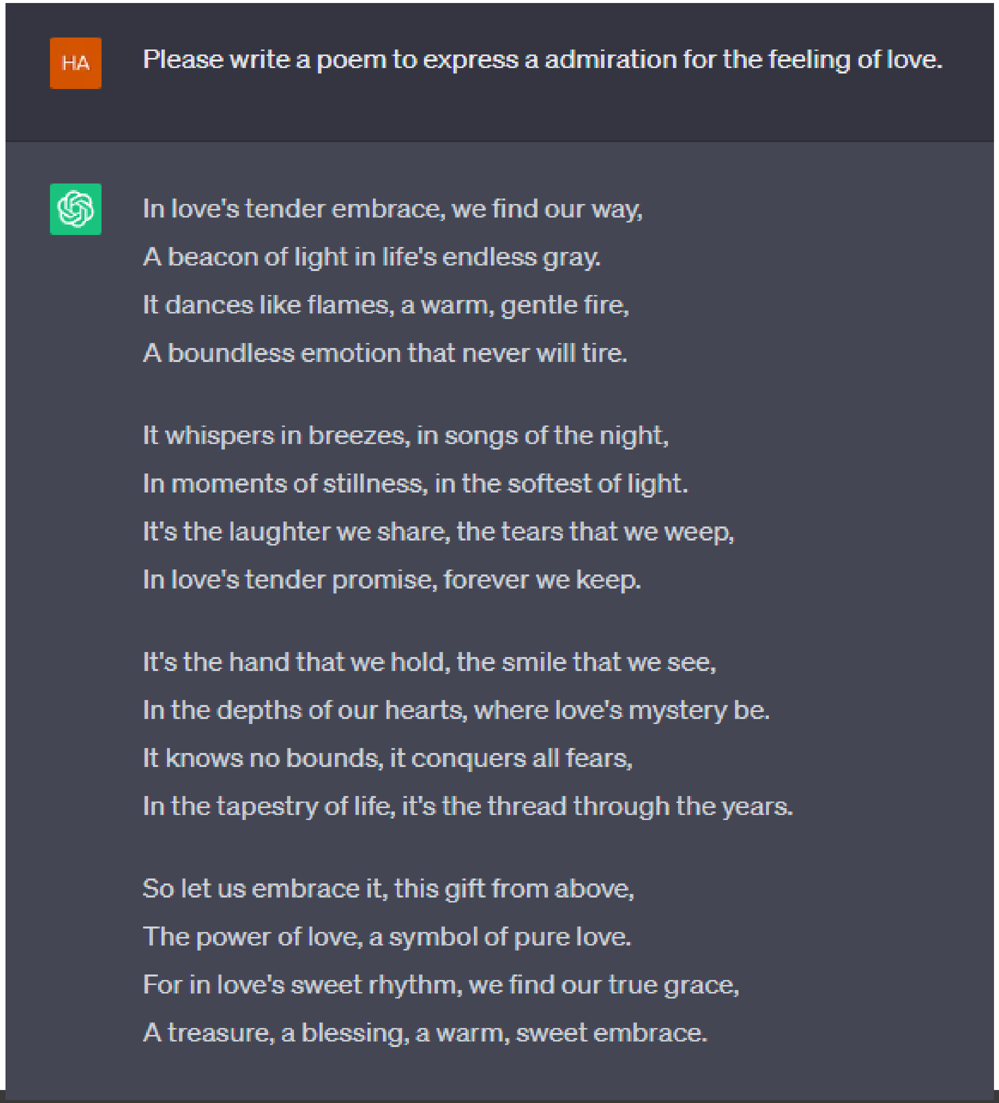


HA What do I really mean when I say "you should try the seafood in this restaurant"?

When you say "you should try the seafood in this restaurant," you are making a recommendation or suggestion to someone, implying that you believe the seafood served at this particular restaurant is worth experiencing. Your statement implies several things:

1. **Positive Experience:** You likely enjoyed the seafood at that restaurant and believe that others will too. You're sharing your positive dining experience.
2. **Quality:** You think the seafood at this restaurant is of high quality and delicious.
3. **Encouragement:** You're encouraging the person to give it a try, suggesting that they might like it as well.
4. **Trust:** Your statement carries a level of trust or confidence in the restaurant's seafood offerings. It implies that you trust the restaurant's reputation for serving good seafood.

Overall, it's a friendly and persuasive way to recommend a specific dish or experience to someone based on your own positive encounter.



HA Please write a poem to express a admiration for the feeling of love.

In love's tender embrace, we find our way,
A beacon of light in life's endless gray.
It dances like flames, a warm, gentle fire,
A boundless emotion that never will tire.

It whispers in breezes, in songs of the night,
In moments of stillness, in the softest of light.
It's the laughter we share, the tears that we weep,
In love's tender promise, forever we keep.

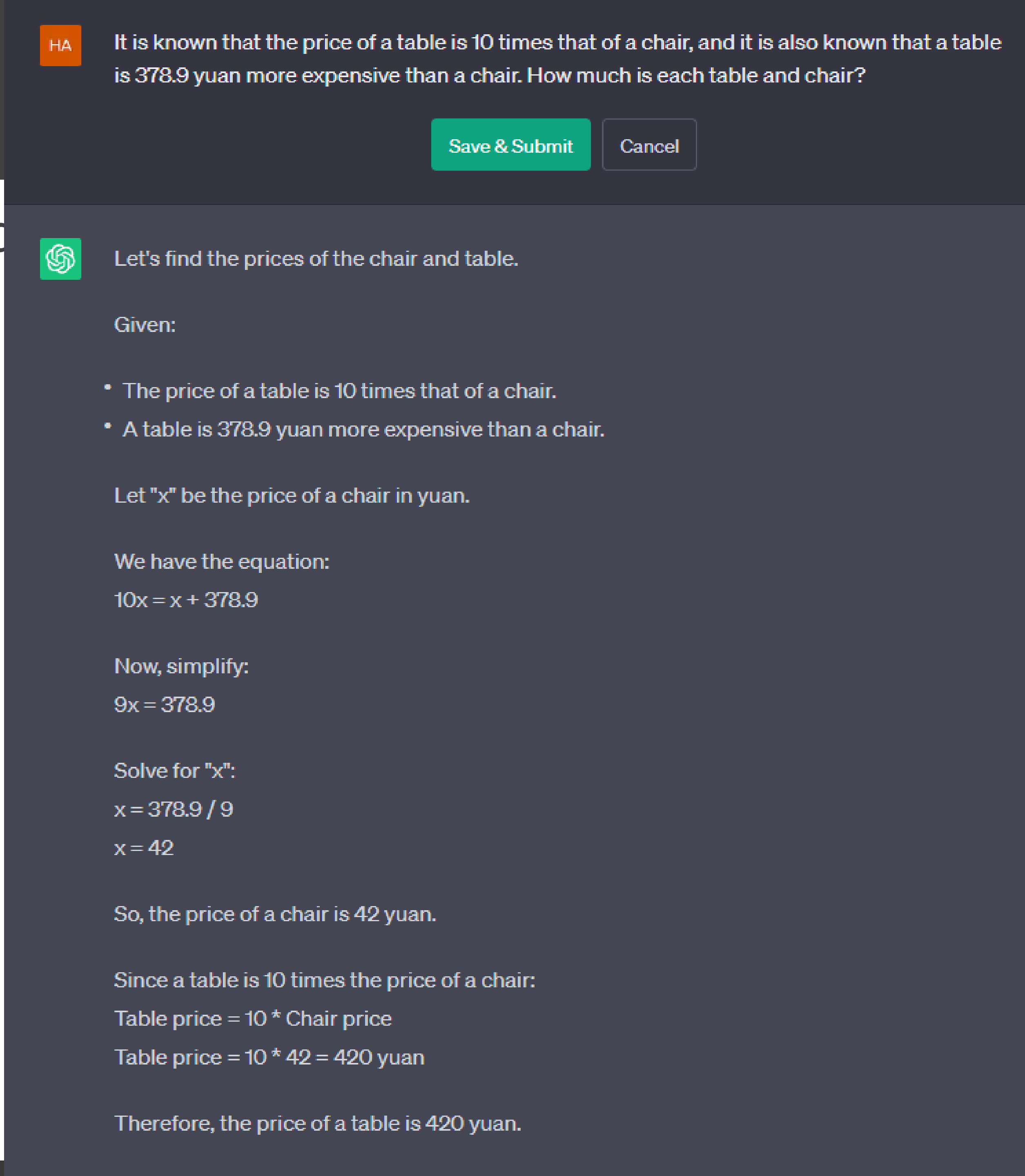
It's the hand that we hold, the smile that we see,
In the depths of our hearts, where love's mystery be.
It knows no bounds, it conquers all fears,
In the tapestry of life, it's the thread through the years.

So let us embrace it, this gift from above,
The power of love, a symbol of pure love.
For in love's sweet rhythm, we find our true grace,
A treasure, a blessing, a warm, sweet embrace.

Large Language Models

- (Near) Human-level Language Unc

☞ Complex (Mathematic) Reasoning Ability



It is known that the price of a table is 10 times that of a chair, and it is also known that a table is 378.9 yuan more expensive than a chair. How much is each table and chair?

Let's find the prices of the chair and table.

Given:

- The price of a table is 10 times that of a chair.
- A table is 378.9 yuan more expensive than a chair.

Let "x" be the price of a chair in yuan.

We have the equation:
10x = x + 378.9

Now, simplify:
9x = 378.9

Solve for "x":
x = 378.9 / 9
x = 42

So, the price of a chair is 42 yuan.

Since a table is 10 times the price of a chair:
Table price = 10 * Chair price
Table price = 10 * 42 = 420 yuan

Therefore, the price of a table is 420 yuan.

Large Language Models

LLMs with In-context Learning

- In-context Learning (ICL), aka., *few-shot prompting*
  - Enhanced understanding of context:
    empowers LLMs with necessary contextual information specific to the problem&domain.
  - Tailored problem-solving capabilities:
    guides LLMs with example inputs&outputs, enables to integrate task-specific knowledge.
  - Improved reasoning and inference:
    empowers models to perform reasoning under more specific manner as user indication.

[1] Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, etc. A survey for in-context learning. 2022.

Large Language Models

LLMs with In-context Learning

- In-context Learning (ICL)

① **Task Instructions**

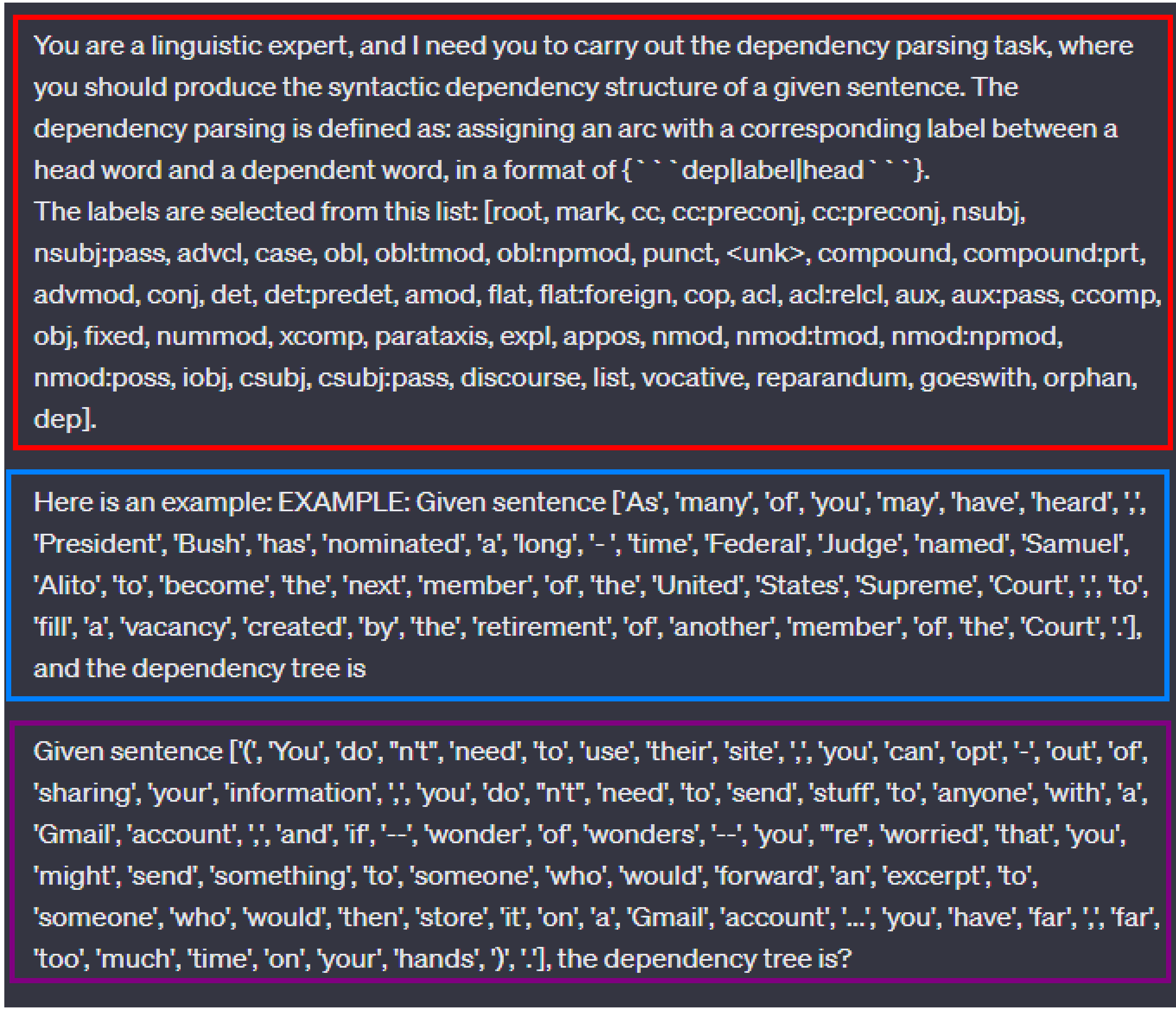
Informing LLM about the task

② **Example Input&Output:**

Showcasing example I/O format

③ **Testing Instance:**

The data to test for LLM



You are a linguistic expert, and I need you to carry out the dependency parsing task, where you should produce the syntactic dependency structure of a given sentence. The dependency parsing is defined as: assigning an arc with a corresponding label between a head word and a dependent word, in a format of {` ` ` dep|label|head ` ` `}.
The labels are selected from this list: [root, mark, cc, cc:preconj, cc:preconj, nsubj, nsubj:pass, advcl, case, obl, obl:tmod, obl:npmod, punct, <unk>, compound, compound:prt, advmod, conj, det, det:predet, amod, flat, flat:foreign, cop, acl, acl:relcl, aux, aux:pass, ccomp, obj, fixed, nummod, xcomp, parataxis, expl, appos, nmod, nmod:tmod, nmod:npmod, nmod:poss, iobj, csubj, csubj:pass, discourse, list, vocative, reparandum, goeswith, orphan, dep].

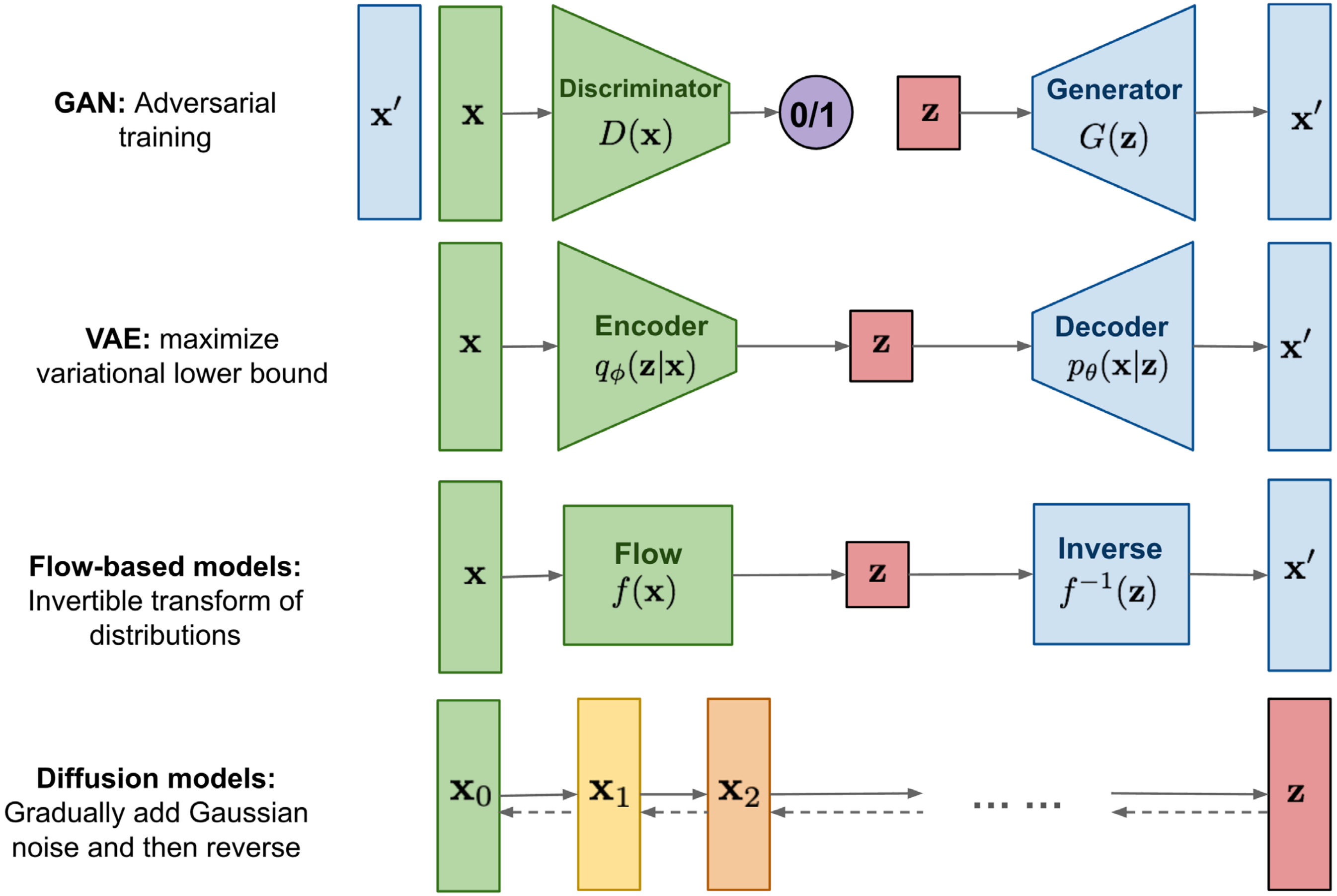
Here is an example: EXAMPLE: Given sentence ['As', 'many', 'of', 'you', 'may', 'have', 'heard', ',', 'President', 'Bush', 'has', 'nominated', 'a', 'long', '- ', 'time', 'Federal', 'Judge', 'named', 'Samuel', 'Alito', 'to', 'become', 'the', 'next', 'member', 'of', 'the', 'United', 'States', 'Supreme', 'Court', ',', 'to', 'fill', 'a', 'vacancy', 'created', 'by', 'the', 'retirement', 'of', 'another', 'member', 'of', 'the', 'Court', '.'], and the dependency tree is

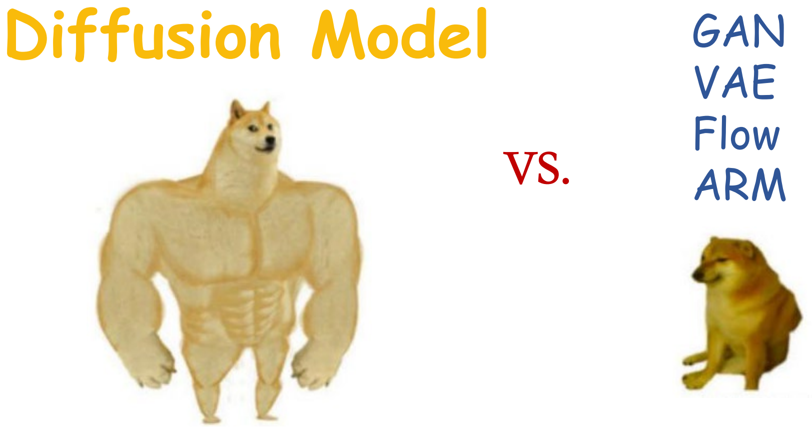
Given sentence ['(', 'You', 'do', "n't", 'need', 'to', 'use', 'their', 'site', ',', 'you', 'can', 'opt', '-', 'out', 'of', 'sharing', 'your', 'information', ',', 'you', 'do', "n't", 'need', 'to', 'send', 'stuff', 'to', 'anyone', 'with', 'a', 'Gmail', 'account', ',', 'and', 'if', '--', 'wonder', 'of', 'wonders', '--', 'you', "'re", 'worried', 'that', 'you', 'might', 'send', 'something', 'to', 'someone', 'who', 'would', 'forward', 'an', 'excerpt', 'to', 'someone', 'who', 'would', 'then', 'store', 'it', 'on', 'a', 'Gmail', 'account', '...', 'you', 'have', 'far', ',', 'far', 'too', 'much', 'time', 'on', 'your', 'hands', ')', '.'], the dependency tree is?

Diffusion Models

Denoising Diffusion Probabilistic Models

- Advantages over prior Visual Generative Models





Diffusion Models

Denoising Diffusion Probabilistic Models

- Advantages over Prior Generative Models
  - *Diverse & Higher-quality Samples*
  - *No Mode Collapse*
  - *Explicit Likelihood Estimation*
  - *Stable Training*
  - …

SoTA generative method

Diffusion Model

vs.

GAN
VAE
Flow
ARM

ARTSSON

Diffusion Models

Denoising Diffusion Probabilistic Models

- Representative Diffusion-based Products

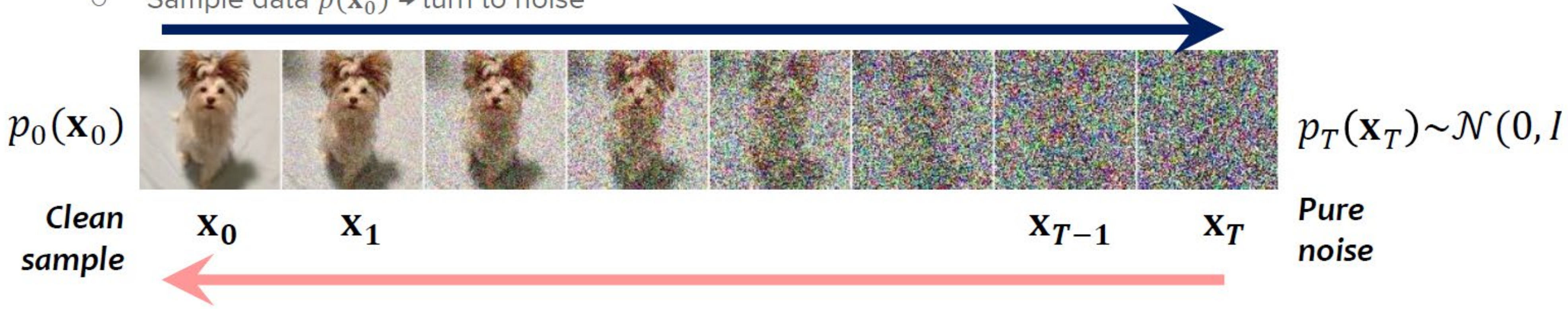
Diffusion Model

SoTA generative method

Diffusion Models

Denoising Diffusion Probabilistic Models

- **Forward / noising process**
  - Sample data $p(\mathbf{x}_0)$ → turn to noise



$p_0(\mathbf{x}_0)$

Clean sample

$\mathbf{x}_0$ $\mathbf{x}_1$ $\mathbf{x}_{T-1}$ $\mathbf{x}_T$

$p_T(\mathbf{x}_T) \sim \mathcal{N}(0, I$

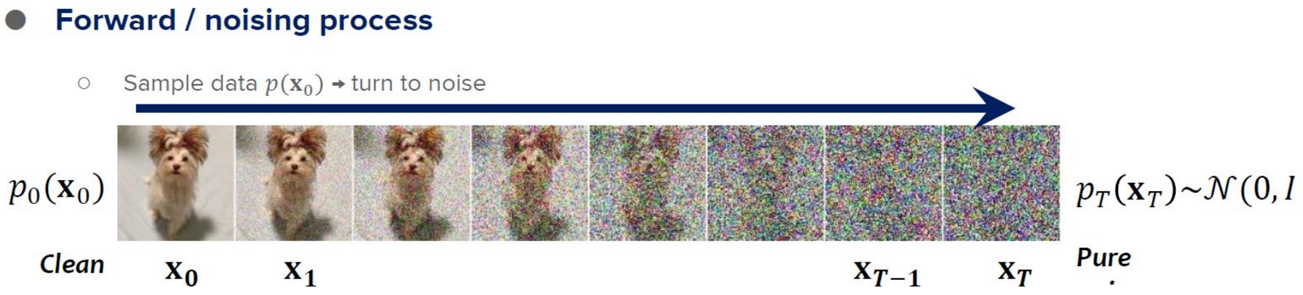
Pure noise

- **Reverse / denoising process**
  - Sample noise $p_T(\mathbf{x}_T)$ → turn into data

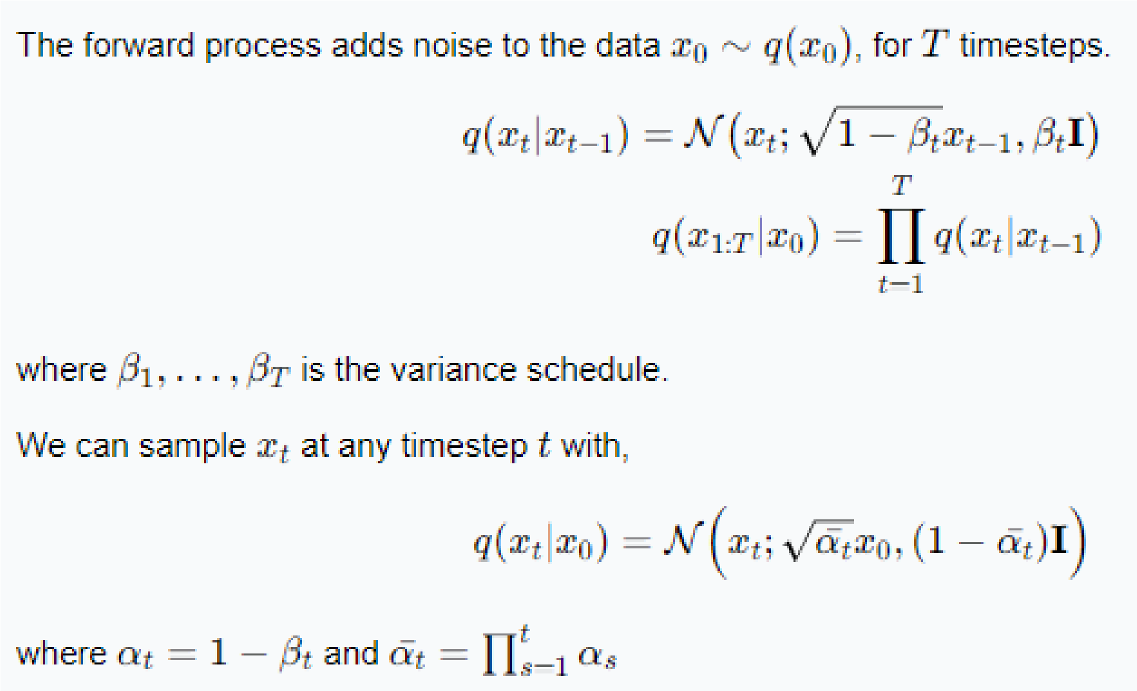
Diffusion Models

Denoising Diffusion Probabilistic Models

- Forward/Noising Pass



- **Forward / noising process**
  - ○ Sample data $p(\mathbf{x}_0)$ → turn to noise



The forward process adds noise to the data $x_0 \sim q(x_0)$, for T timesteps.

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}\right)$$

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$$

where $\beta_1, \ldots, \beta_T$ is the variance schedule.

We can sample x_t at any timestep t with,

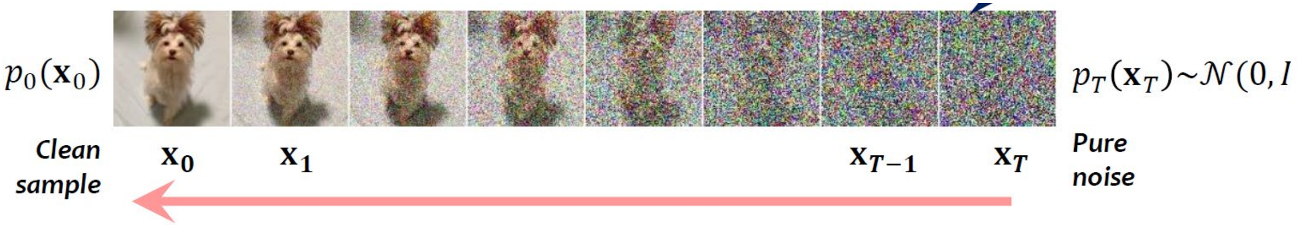
$$q(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}\right)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$

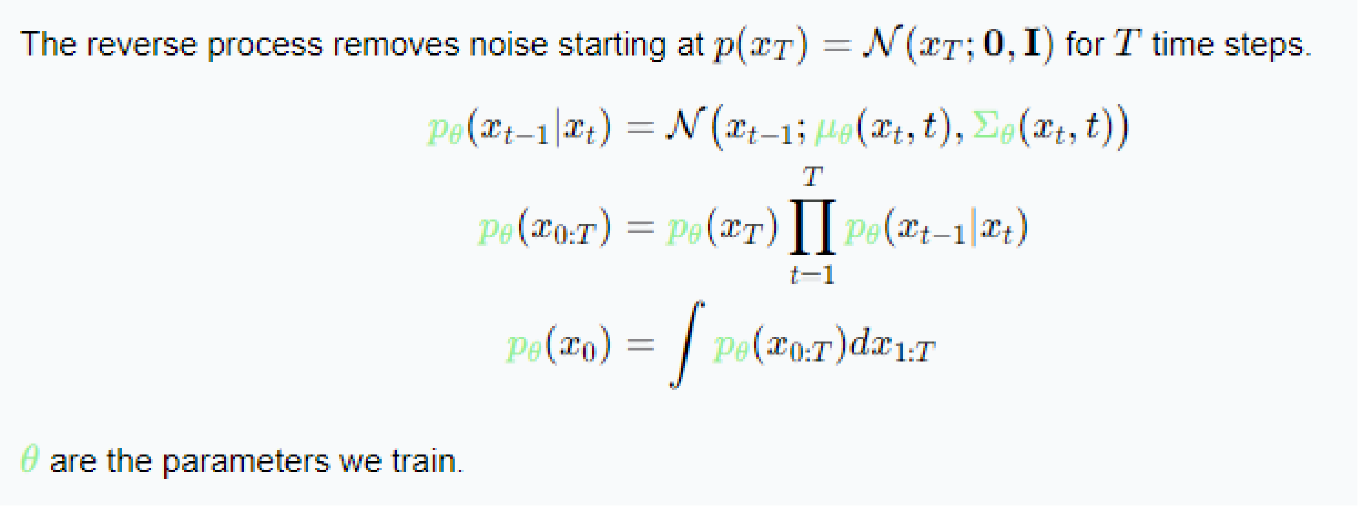
Diffusion Models

Denoising Diffusion Probabilistic Models

- Reverse/Denoising Pass



- **Reverse / denoising process**
  - Sample noise $p_T(\mathbf{x}_T)$ → turn into data



The reverse process removes noise starting at $p(x_T) = \mathcal{N}(x_T; \mathbf{0}, \mathbf{I})$ for T time steps.

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\right)$$

$$p_\theta(x_{0:T}) = p_\theta(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$$

$$p_\theta(x_0) = \int p_\theta(x_{0:T}) dx_{1:T}$$

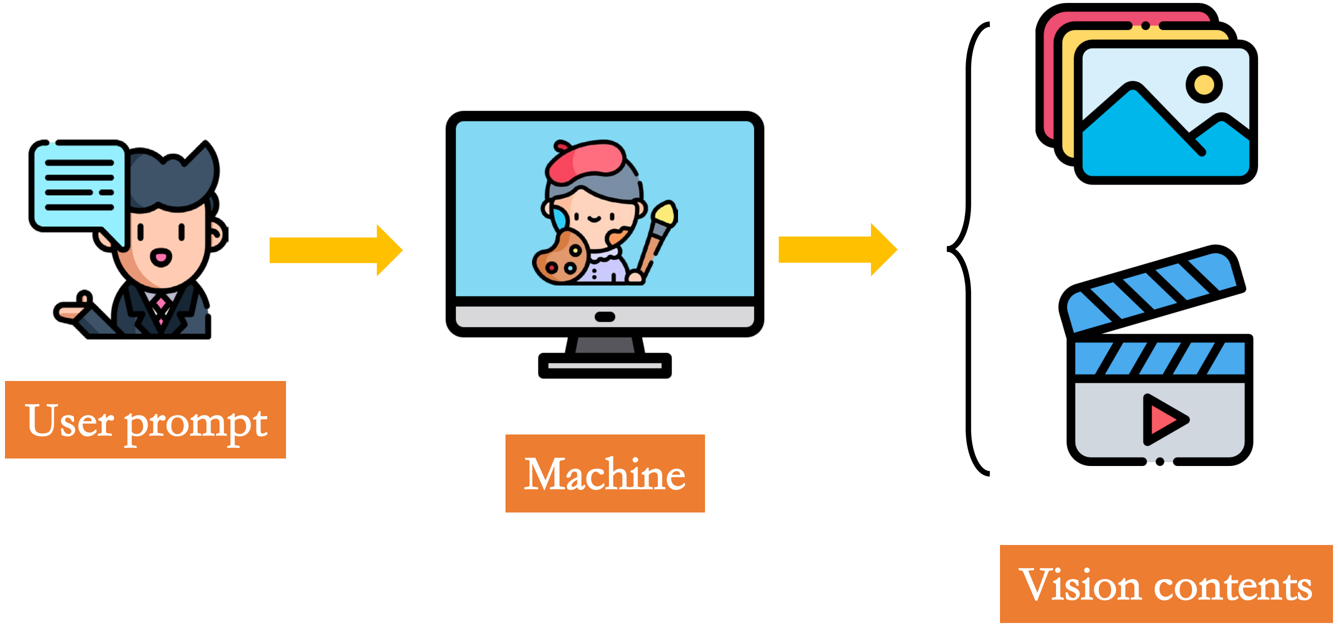
θ are the parameters we train.

Diffusion Models

Text-to-Vision Learning

➢ Text-to-vision synthesis has wide range of practical demands

- Content Creation
- Accessibility
- E-commerce
- Educational Materials
- Art and Design
- Data Visualization
- …



User prompt

Machine

Vision contents

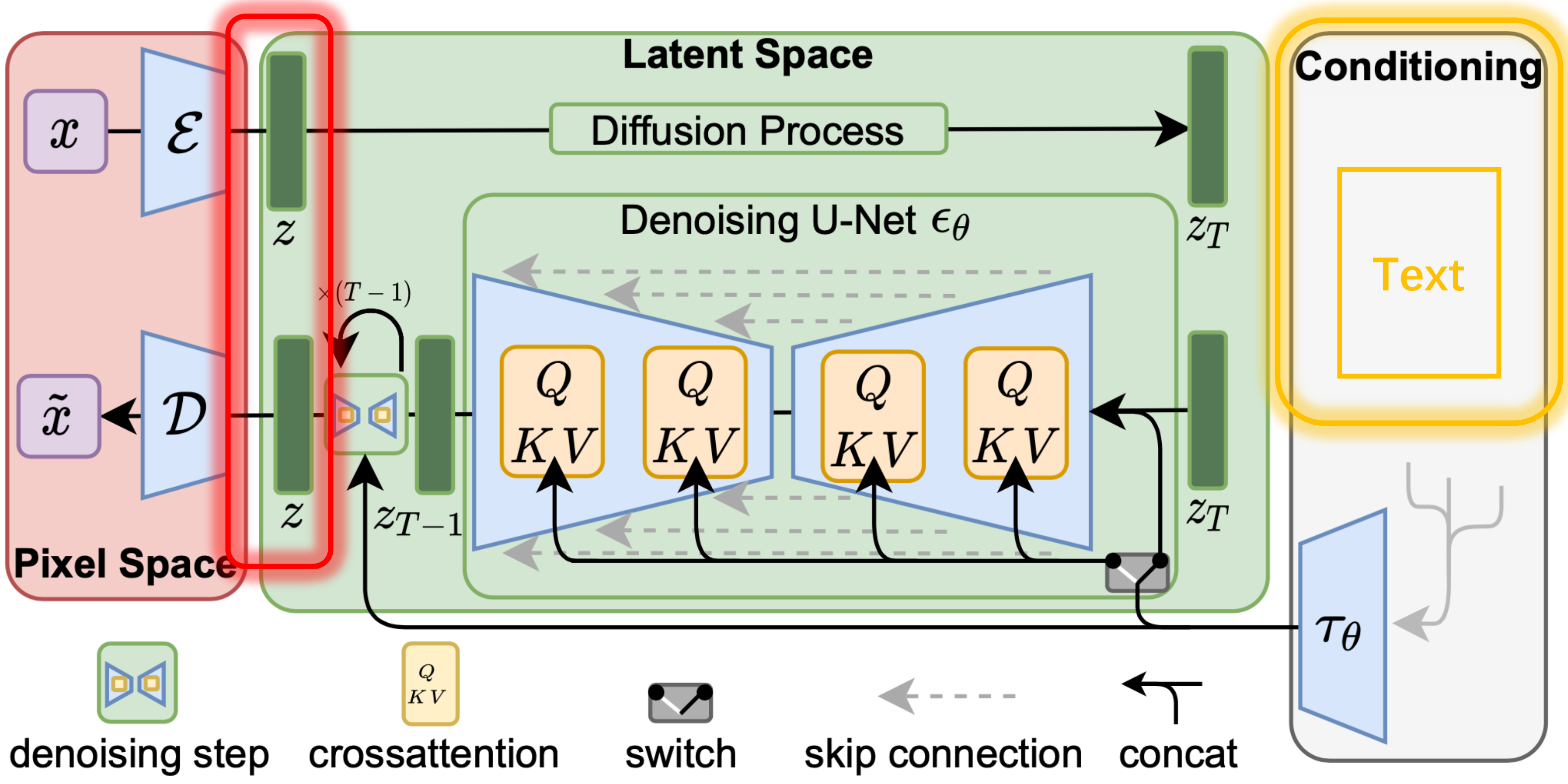
Diffusion Models

Diffusion-based **Text-to-Image** Synthesis

- Latent Diffusion Model (LDM) based T2I

Transform into Latent space:

- *Lower dimension representation*
- *Faster message passing*
- *Lower memory requirement*

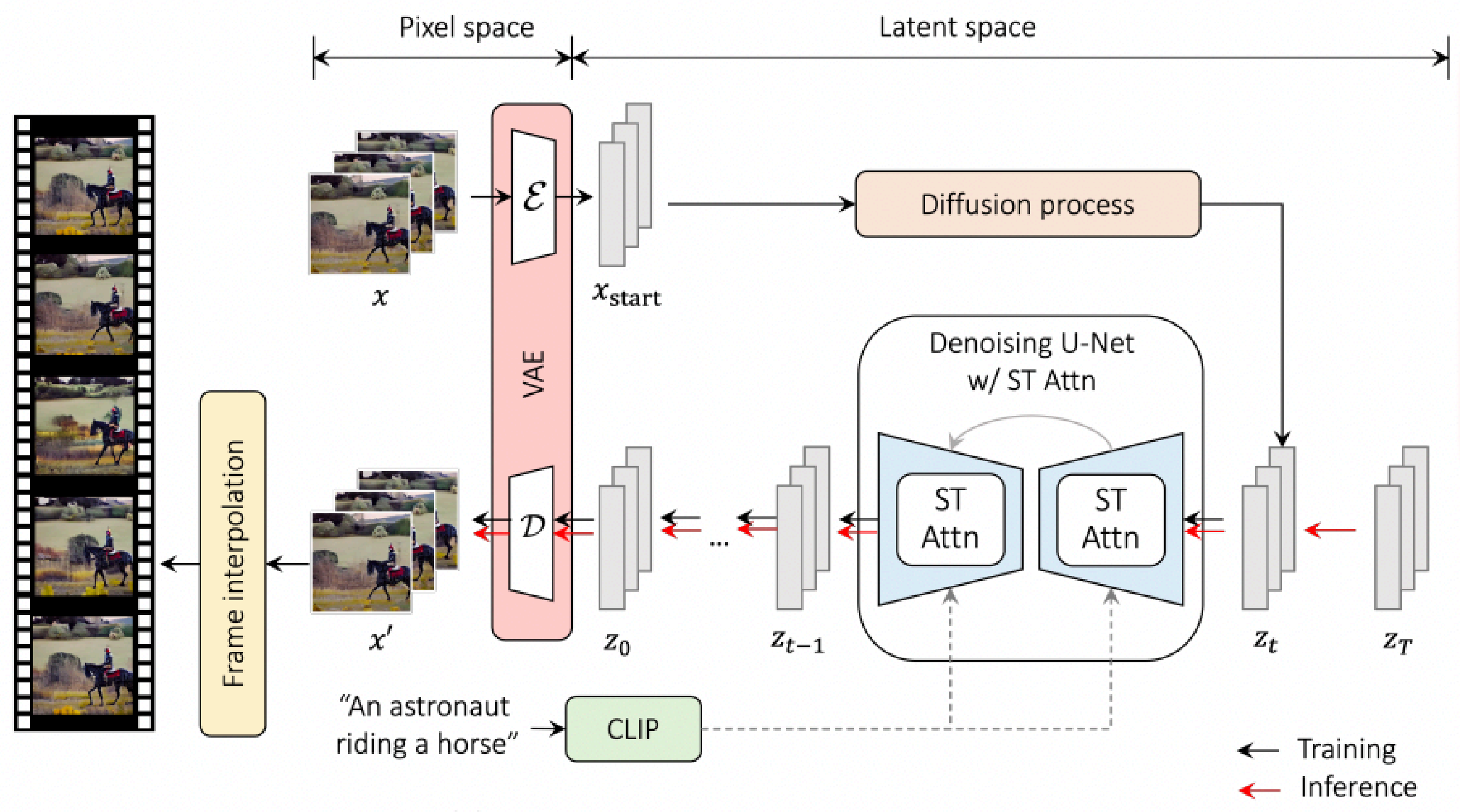


[1] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. High-resolution image synthesis with latent diffusion models. CVPR. 2022.

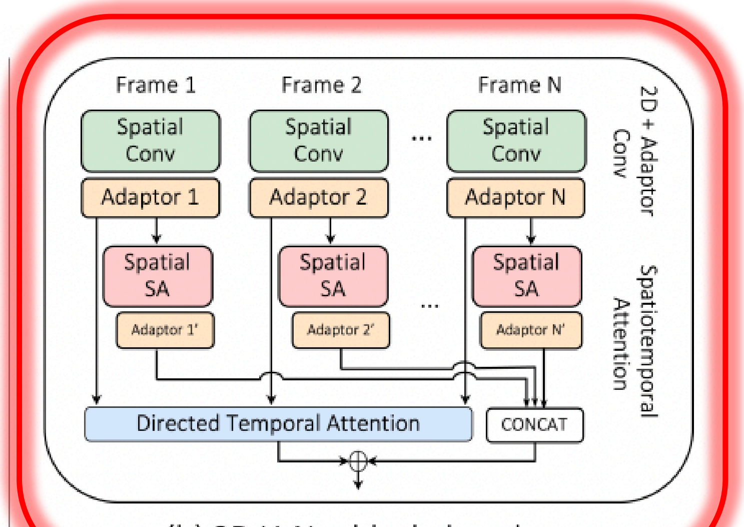
Diffusion Models

Diffusion-based **Text-to-Video** Synthesis

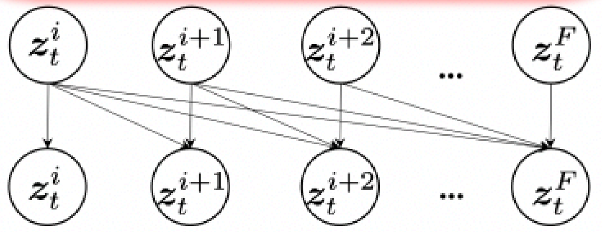
- Latent Diffusion Model (LDM) based T2V



(a) MagicVideo pipeline



(b) 3D U-Net block decoder



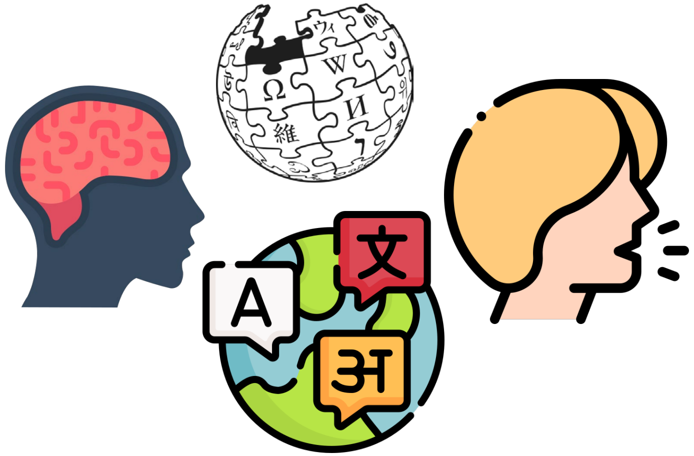
(c) Directed Temporal Attention

[1] Zhou, D., Wang, W., etc. Magicvideo: Efficient video generation with latent diffusion models. 2022.

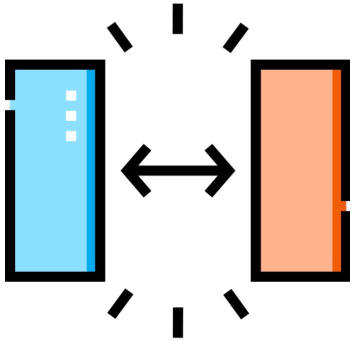
How Can LLM Aid Text-to-Vision Diffusion?

The Gap between Language and Vision

Language



abstract & concise



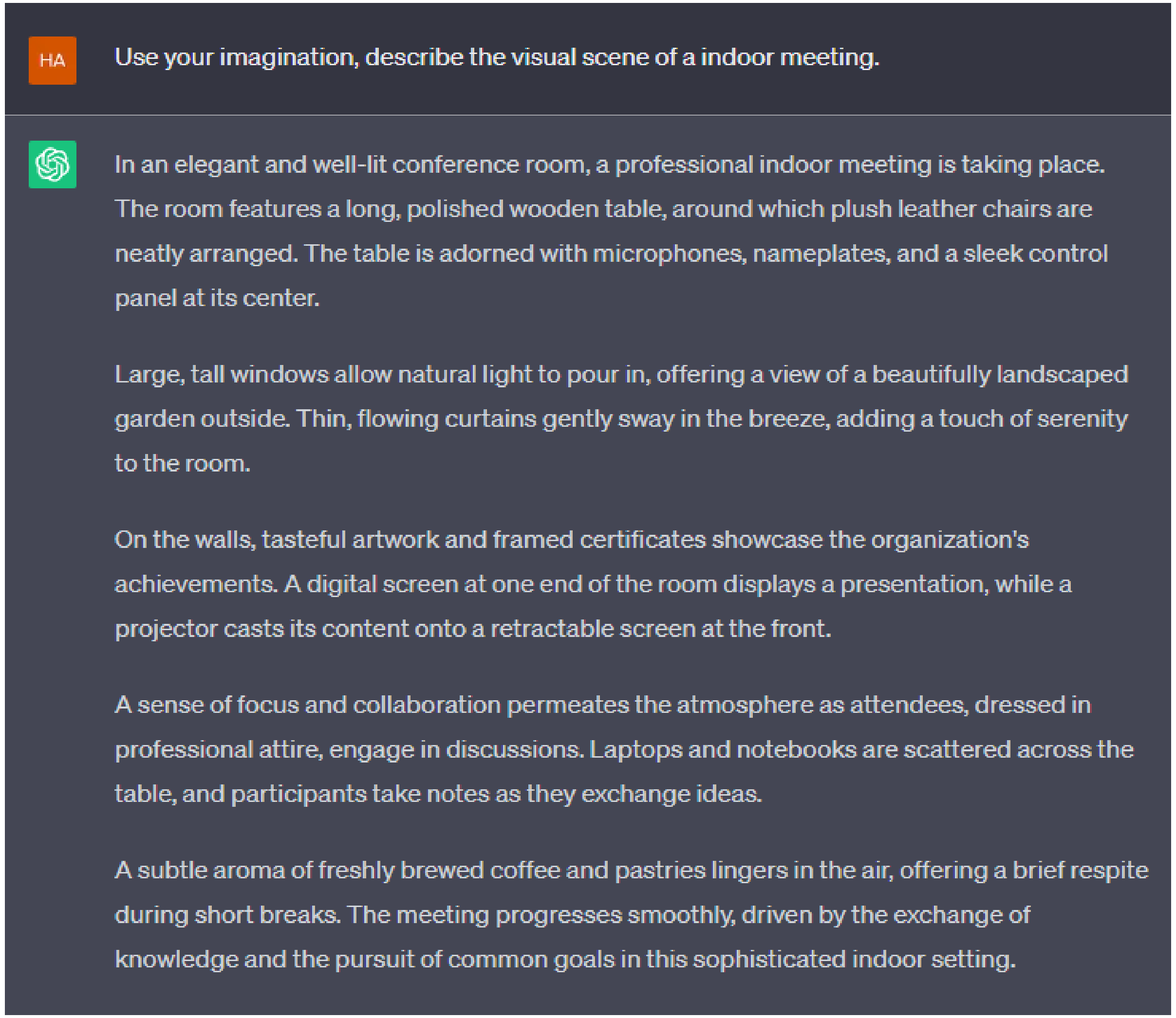
Gap to bridge

Vision

intricate & specific

How Can LLM Aid Text-to-Vision Diffusion?

Are LLMs Able to Understand Vision?



Use your imagination, describe the visual scene of a indoor meeting.

In an elegant and well-lit conference room, a professional indoor meeting is taking place. The room features a long, polished wooden table, around which plush leather chairs are neatly arranged. The table is adorned with microphones, nameplates, and a sleek control panel at its center.

Large, tall windows allow natural light to pour in, offering a view of a beautifully landscaped garden outside. Thin, flowing curtains gently sway in the breeze, adding a touch of serenity to the room.

On the walls, tasteful artwork and framed certificates showcase the organization's achievements. A digital screen at one end of the room displays a presentation, while a projector casts its content onto a retractable screen at the front.

A sense of focus and collaboration permeates the atmosphere as attendees, dressed in professional attire, engage in discussions. Laptops and notebooks are scattered across the table, and participants take notes as they exchange ideas.

A subtle aroma of freshly brewed coffee and pastries lingers in the air, offering a brief respite during short breaks. The meeting progresses smoothly, driven by the exchange of knowledge and the pursuit of common goals in this sophisticated indoor setting.

*LLMs **DO** have visual understanding capability!*

How Can LLM Aid Text-to-Vision Diffusion?

Workaround for LLMs to Aid Diffusion Process

Enriching the raw textual prompt with more details of visual descriptions?

- Raw input prompt

  Two young men give a presentation in the office.

- Prompt enriching-I

  Two middle-aged, nice, enthusiastic, confident, man with polished shoes and sleek hair give a professional presentation in the spacious and modern conference room of the corporate blue office, room with chairs and tables.

  Issue: Vision Distraction

- Prompt enriching-II

  Two young man give a presentation in the office, old, nice, confident, enthusiastic, laughing man with polished hair, seek hair, room with chairs and tables, speaking to each other.

  Issue: Wrong Binding

Language
abstract & concise

Gap still not closed!

Vision
intricate & specific

How Can LLM Aid Text-to-Vision Diffusion?

Workaround for LLMs to Aid Diffusion Process

Representing the visual scenes induced from LLMs into structured feature representations?

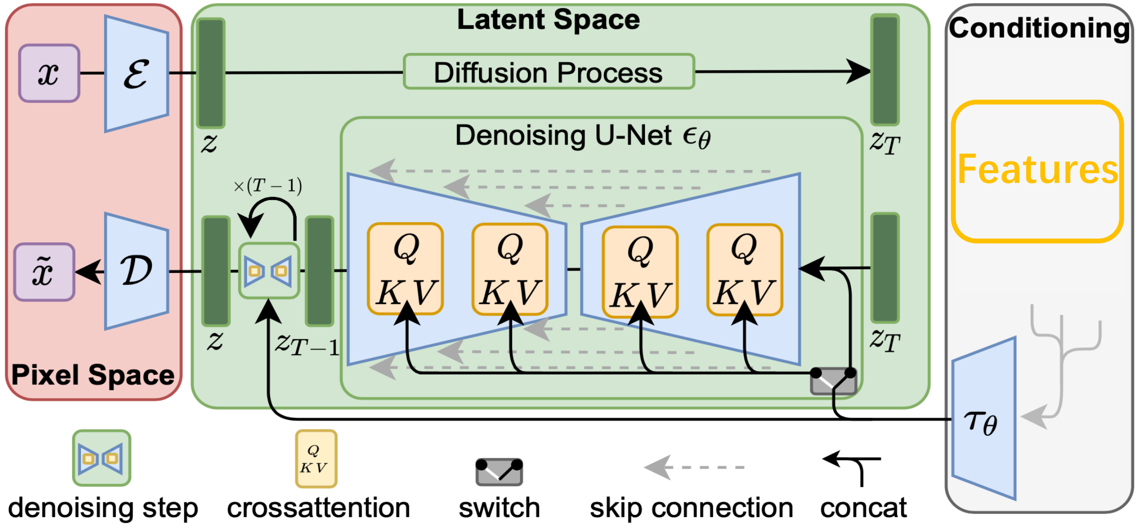
- Spatial Understanding
  - *overall layout*
  - *dimension*
  - *sketches*
  - *scene structure*

- Temporal Understanding
  - *visual coherence*
  - *action dynamics*

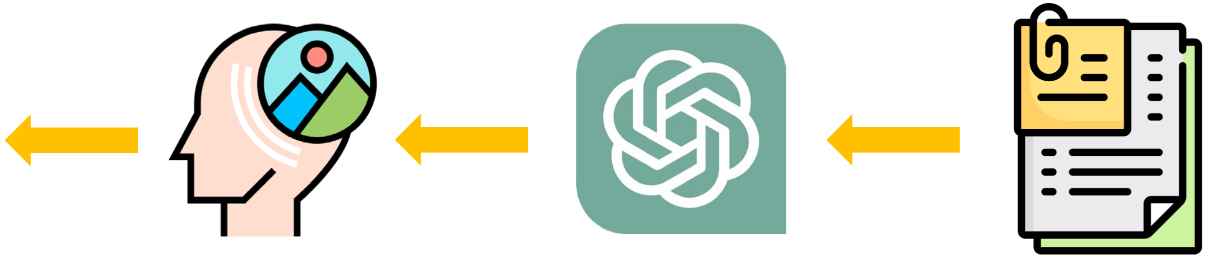
Structured Visual Representation

How Can LLM Aid Text-to-Vision Diffusion?

- Visual scenes into structured feature representations



Diffusion model



Diffusion understandable ***visual representation***

LLM

Text prompt

CONTENT

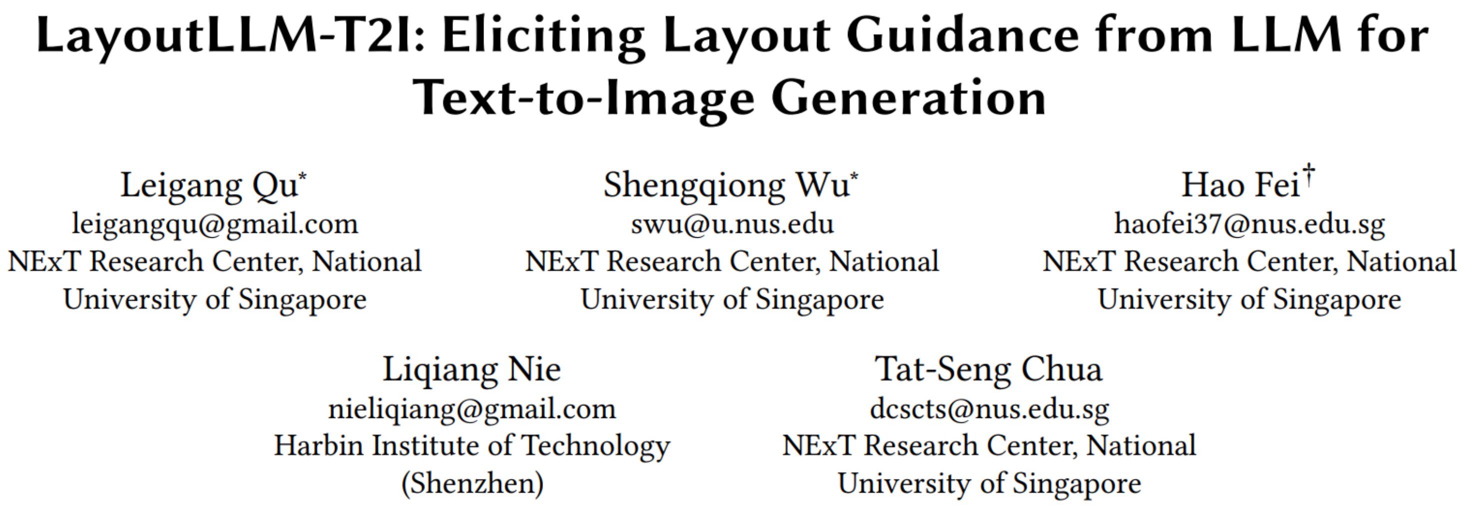
1 LLM Meets Diffusion

2 LLM-Empowered Text-to-Image Diffusion

3 LLM-Empowered Text-to-Vision Diffusion

4 Summary and Outlook

LLM-Empowered Text-to-Image Diffusion



LayoutLLM-T2I: Eliciting Layout Guidance from LLM for Text-to-Image Generation

Leigang Qu*
leigangqu@gmail.com
NExT Research Center, National University of Singapore

Shengqiong Wu*
swu@u.nus.edu
NExT Research Center, National University of Singapore

Hao Fei†
haofei37@nus.edu.sg
NExT Research Center, National University of Singapore

Liqiang Nie
nieliqiang@gmail.com
Harbin Institute of Technology (Shenzhen)

Tat-Seng Chua
dcscts@nus.edu.sg
NExT Research Center, National University of Singapore

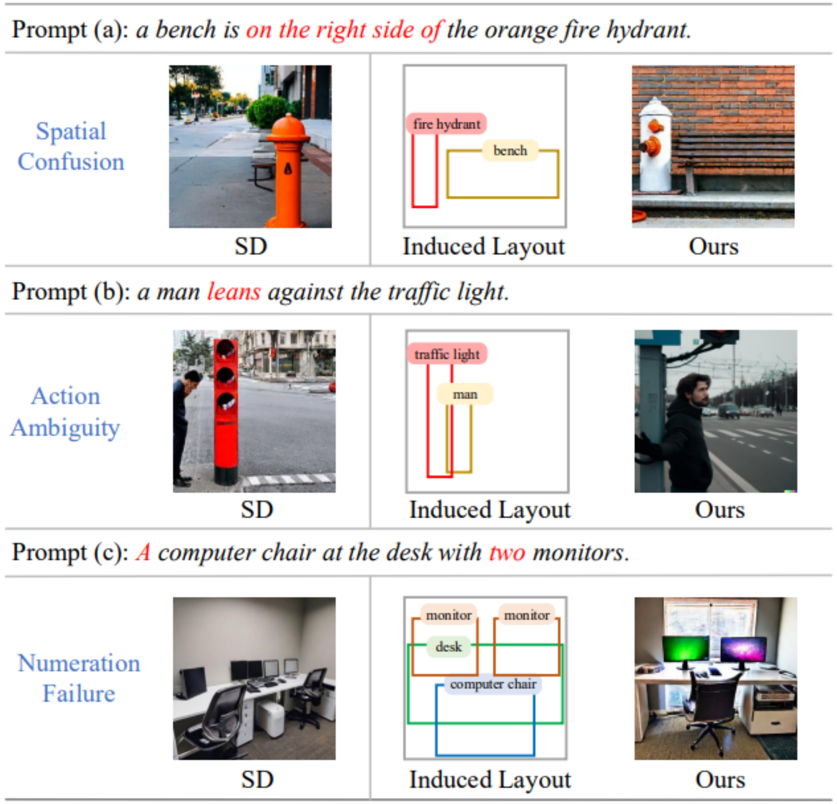
https://layoutllm-t2i.github.io/

[1] Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, Tat-Seng Chua. ***LayoutLLM-T2I: Eliciting Layout Guidance from LLM for Text-to-Image Generation****. ACM MM. 2023.*

LLM-Empowered Text-to-Image Diffusion

Motivation

- Diffusion-based Text-to-Image Generation
  - *Spatial Confusion*
  - *Action Ambiguity*
  - *Numeration Failure*



Prompt (a): *a bench is on the right side of the orange fire hydrant.*

Prompt (b): *a man leans against the traffic light.*

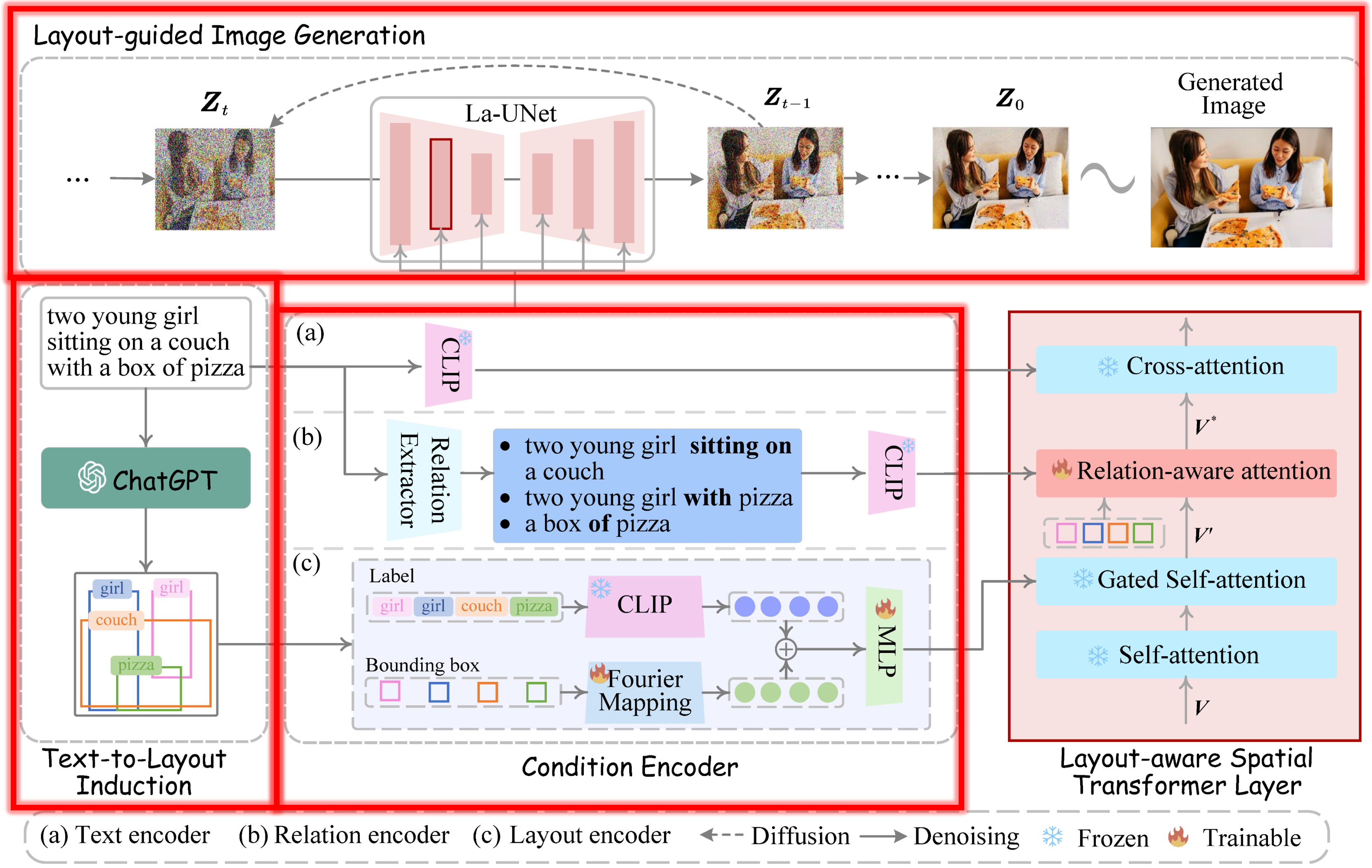
Prompt (c): *A computer chair at the desk with two monitors.*

Figure 1: Illustration of T2I task. Given the prompt, Stable Diffusion (SD) is subject to certain issues such as *spatial confusion, action ambiguilty* and *numeration failure*. Our proposed model is able to synthesize high-faithfulness images by leveraging the automatically generated layouts. Numeration and relation terms in prompts are marked with red.

LLM-Empowered Text-to-Image Diffusion

Method

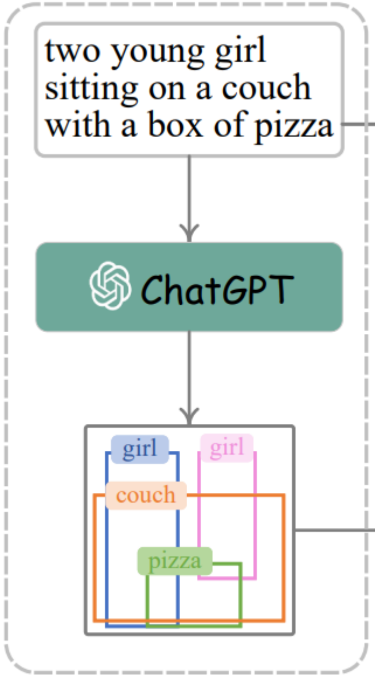
- Framework

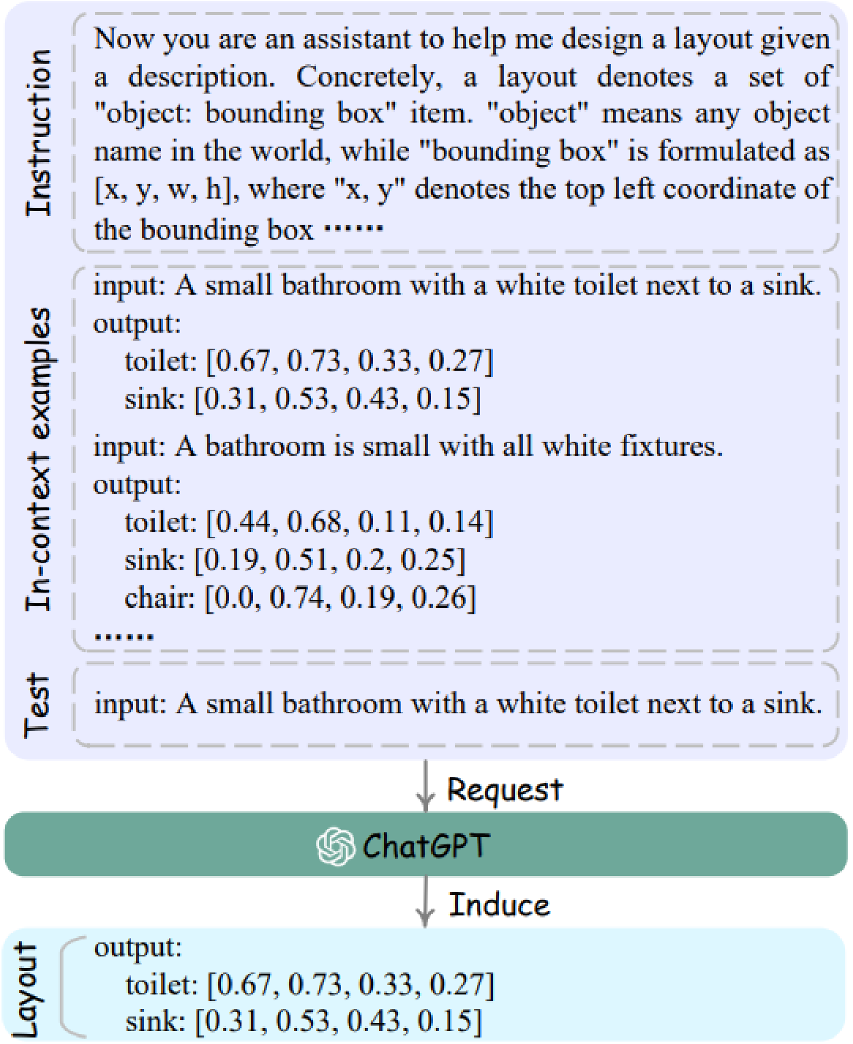


LLM-Empowered Text-to-Image Diffusion

Method

- Text-to-Layout Induction via LLM





Now you are an assistant to help me design a layout given a description. Concretely, a layout denotes a set of "object: bounding box" item. "object" means any object name in the world, while "bounding box" is formulated as [x, y, w, h], where "x, y" denotes the top left coordinate of the bounding box ……

input: A small bathroom with a white toilet next to a sink.
output:
toilet: [0.67, 0.73, 0.33, 0.27]
sink: [0.31, 0.53, 0.43, 0.15]

input: A bathroom is small with all white fixtures.
output:
toilet: [0.44, 0.68, 0.11, 0.14]
sink: [0.19, 0.51, 0.2, 0.25]
chair: [0.0, 0.74, 0.19, 0.26]
……

input: A small bathroom with a white toilet next to a sink.

output:
toilet: [0.67, 0.73, 0.33, 0.27]
sink: [0.31, 0.53, 0.43, 0.15]

Figure 3: Schematic illustration of layout generation.

LLM-Empowered Text-to-Image Diffusion

Method

- Reinforce-optimized ICL demonstration selection
  - Policy Network

$$c_i^k \sim \pi_\psi(c_i|y_i),$$

$$\pi_\psi(c_i|y_i) = \frac{\exp(f(y[c_i]) \cdot f(y_i))}{\sum_{c' \in C} \exp(f(y[c']) \cdot f(y_i))},$$

  - Reward

$$R(\hat{b}_i|y_i) = R_i^B + R_i^I,$$

$$R_i^B = \mathrm{mIoU}(\hat{b}_i, b_i),$$

$$R_i^I = \mathrm{Sim}(\hat{x}_i, x_i, y_i) + \mathrm{Aes}(\hat{x}_i),$$

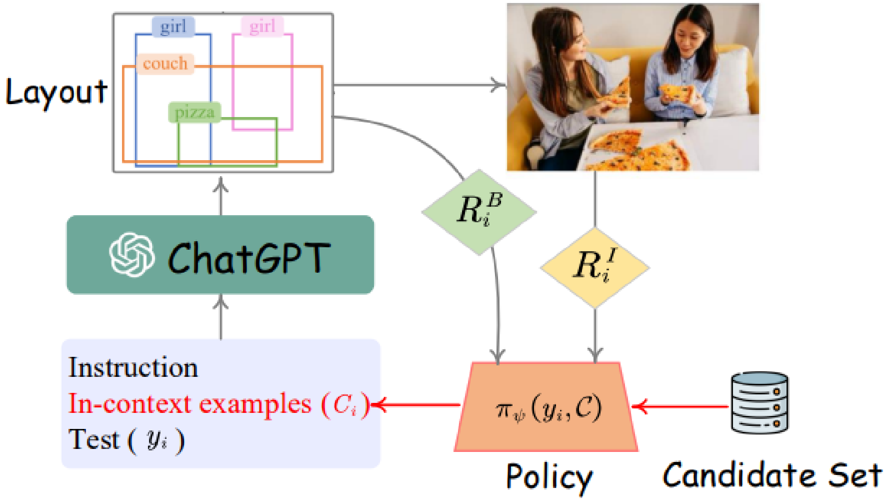


Figure 4: The layout-image feedback module consists of a policy network $\pi_\psi(y_i, C)$ and two rewards R_i^I and R_i^B. Guided by these two rewards, the policy learns to sample informative training data instances as the context fed into LLMs to activate the layout planning abilities.

LLM-Empowered Text-to-Image Diffusion

Experiment

- How does the proposed method perform in the layout planning and high-faithfulness image synthesis compared with state-of-the-art baselines?

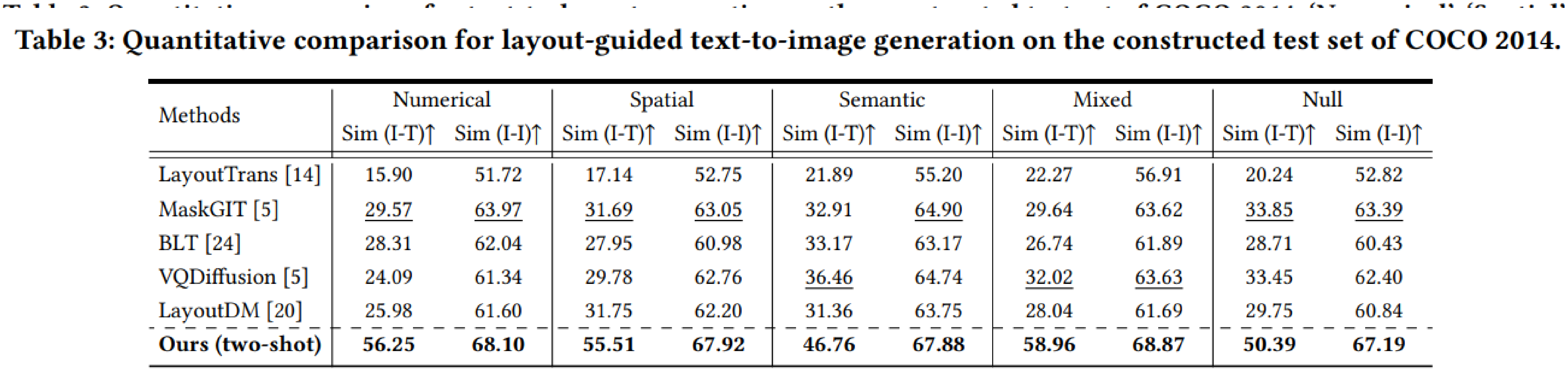
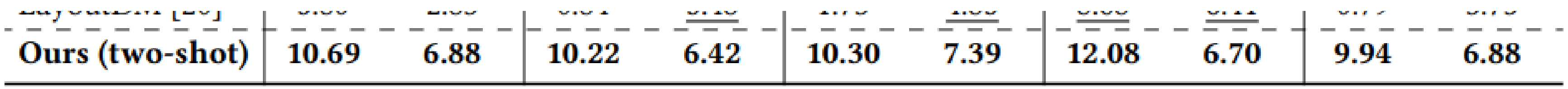


Table 3: Quantitative comparison for layout-guided text-to-image generation on the constructed test set of COCO 2014.

Methods	Numerical		Spatial		Semantic		Mixed		Null	
	Sim (I-T)↑	Sim (I-I)↑	Sim (I-T)↑	Sim (I-I)↑	Sim (I-T)↑	Sim (I-I)↑	Sim (I-T)↑	Sim (I-I)↑	Sim (I-T)↑	Sim (I-I)↑
LayoutTrans [14]	15.90	51.72	17.14	52.75	21.89	55.20	22.27	56.91	20.24	52.82
MaskGIT [5]	29.57	63.97	31.69	63.05	32.91	64.90	29.64	63.62	33.85	63.39
BLT [24]	28.31	62.04	27.95	60.98	33.17	63.17	26.74	61.89	28.71	60.43
VQDiffusion [5]	24.09	61.34	29.78	62.76	36.46	64.74	32.02	63.63	33.45	62.40
LayoutDM [20]	25.98	61.60	31.75	62.20	31.36	63.75	28.04	61.69	29.75	60.84
Ours (two-shot)	**56.25**	**68.10**	**55.51**	**67.92**	**46.76**	**67.88**	**58.96**	**68.87**	**50.39**	**67.19**

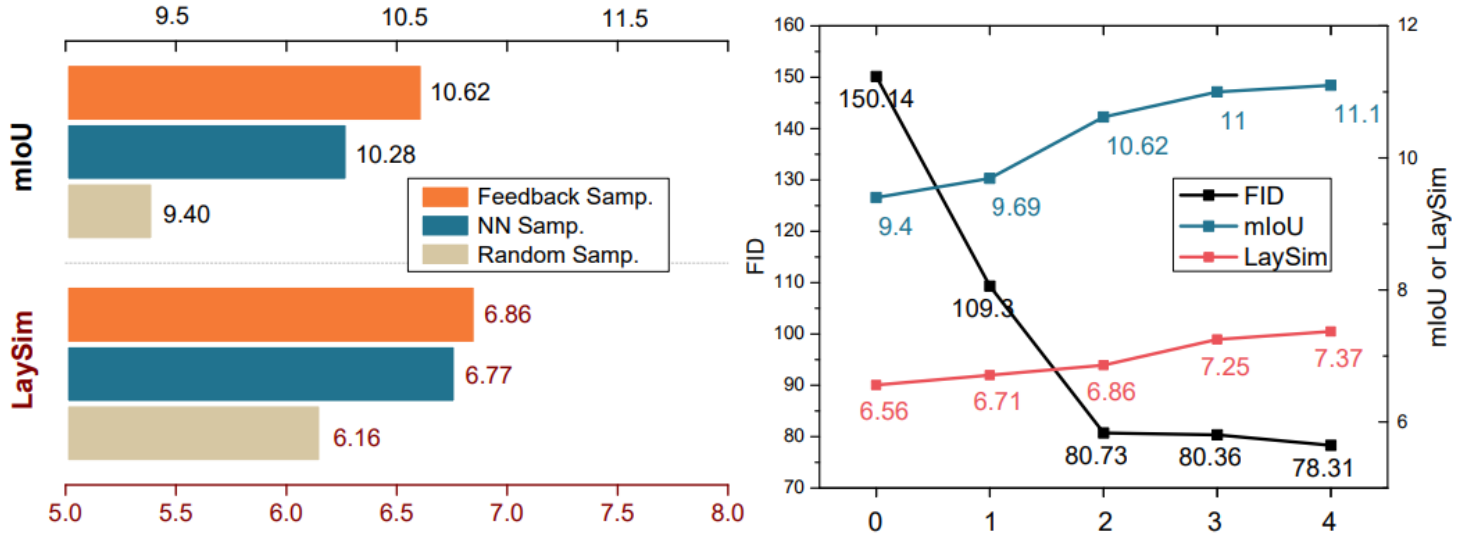


Ours (two-shot)	**10.69**	**6.88**	**10.22**	**6.42**	**10.30**	**7.39**	**12.08**	**6.70**	**9.94**	**6.88**

LLM-Empowered Text-to-Image Diffusion

Experiment

- ICL demonstration selection



(a) In-context Example Sampling

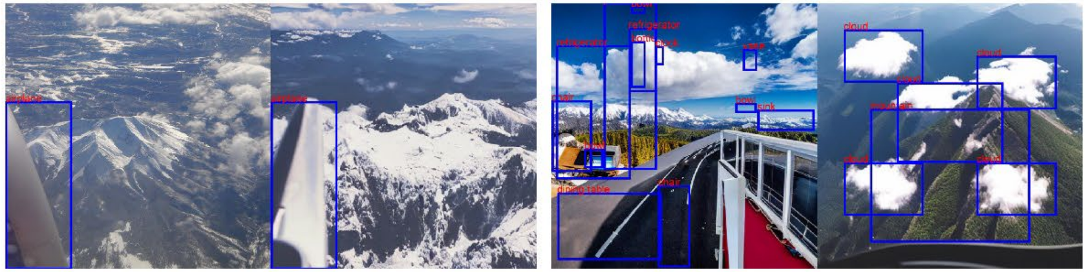
(b) Shot number

Figure 5: Comparison of the (a) in-context example sampling strategies and (b) shot numbers for layout performance. Random and NN Samp. denote random sampling and the nearest neighbor sampling, respectively.

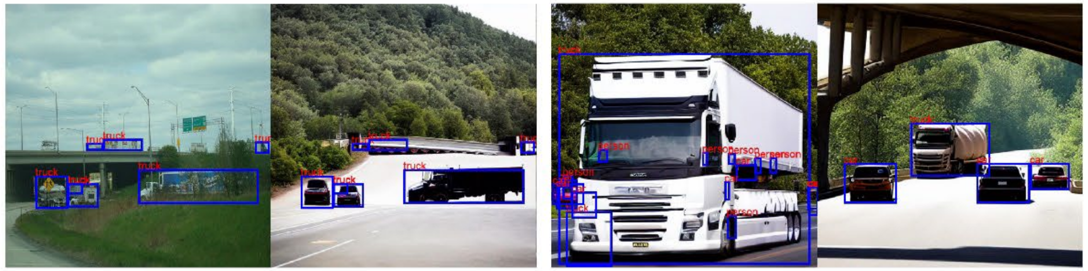
LLM-Empowered Text-to-Image Diffusion

Spatial

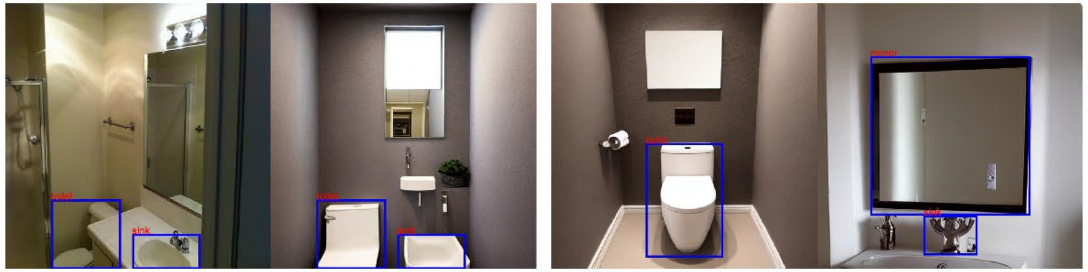
GT **GT*** **LayoutDM** **Ours**



View of a mountain top and clouds from a very high angle

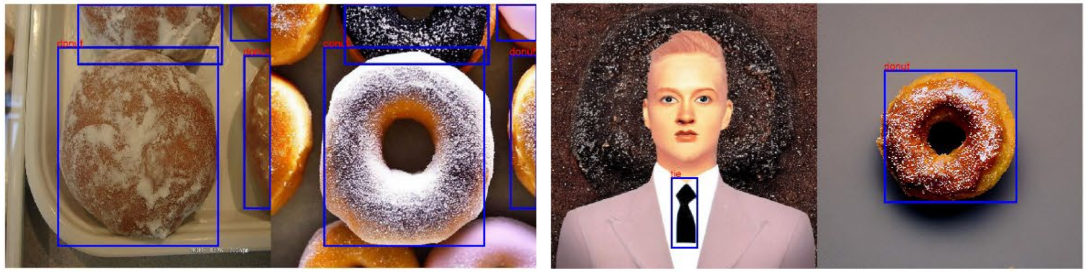


A truck and other cars under a bridge

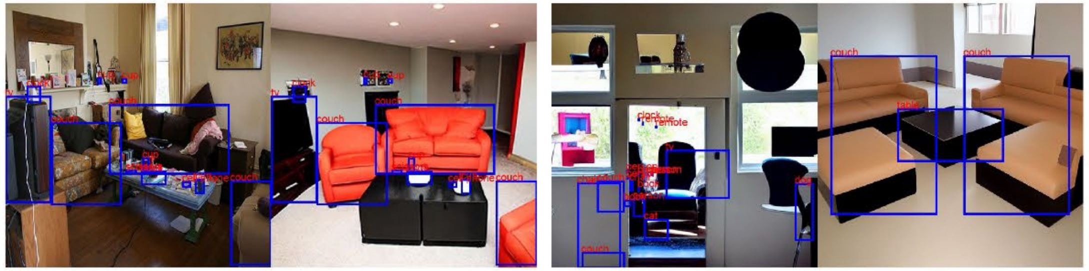


A bathroom with a big mirror above the sink

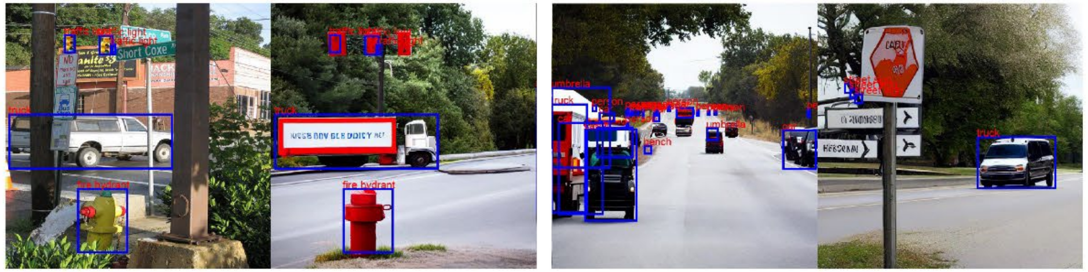
GT **GT*** **LayoutDM** **Ours**



There is a close up picture of a sugar donut



A room with some couches and a table inside of it



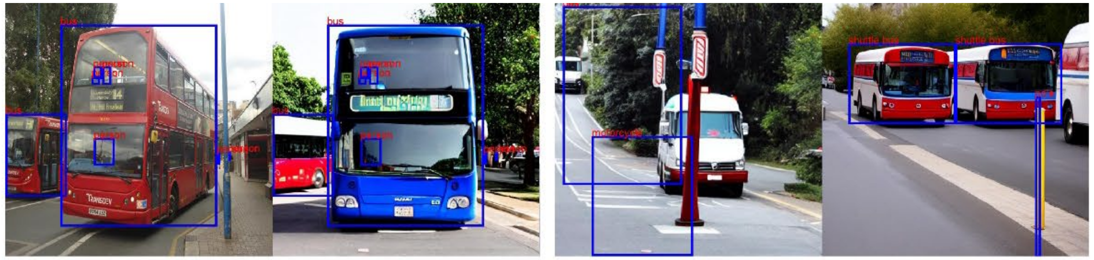
Some street signs near a road with a truck.

LLM-Empowered Text-to-Image Diffusion

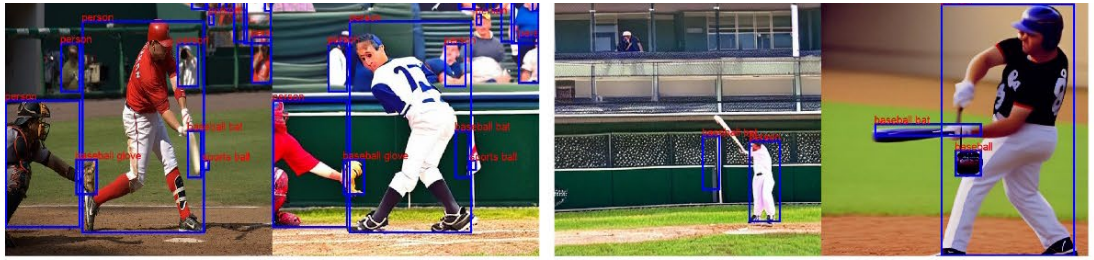
Semantic

GT **GT*** **LayoutDM** **Ours**

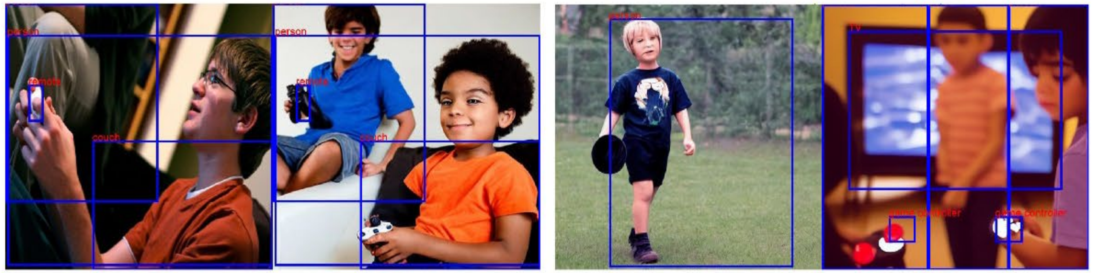
GT **GT*** **LayoutDM** **Ours**



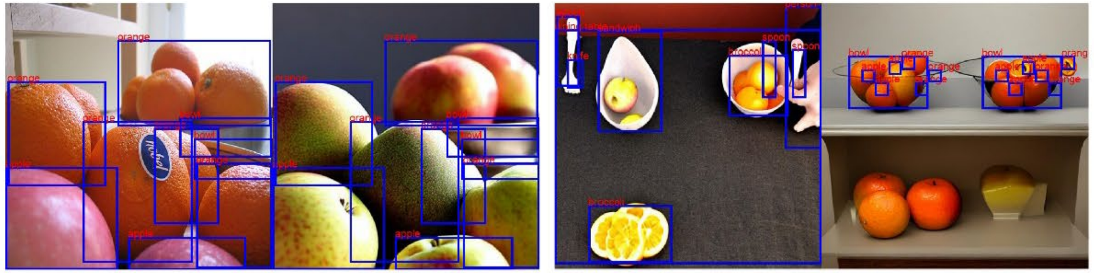
Two red shuttle buses riding down a street next to a blue pole.



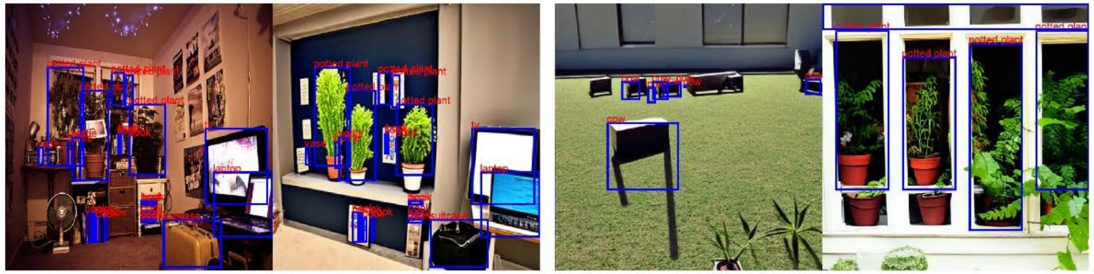
The batter taking a low swing at the ball



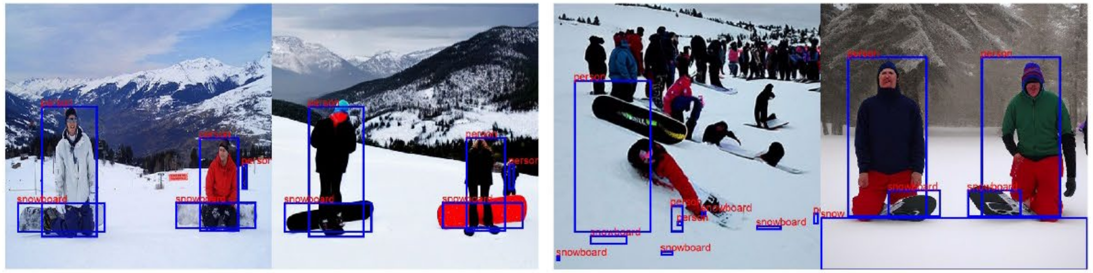
A young boy playing video games with her friends



Two metal bowls filled with apples and oranges.

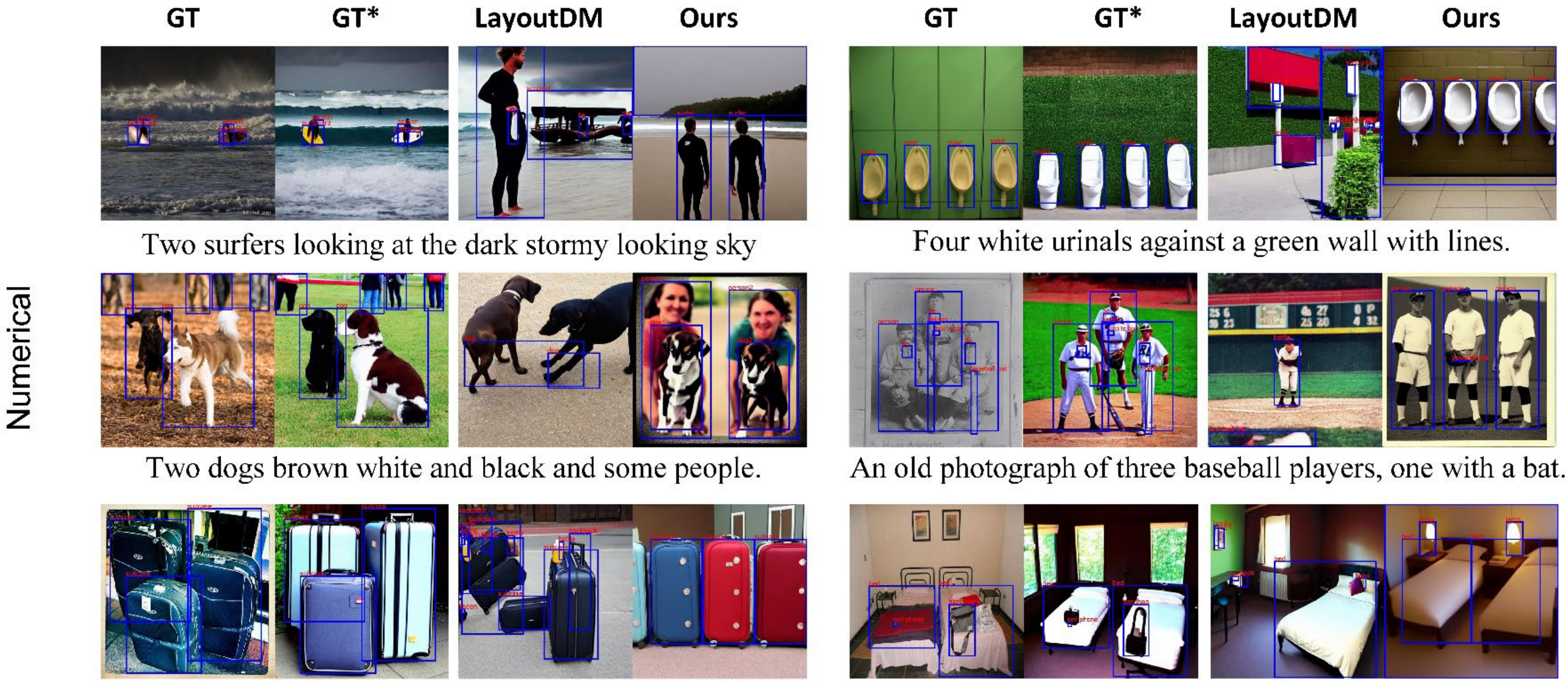


The room has many potted plants in the window.



People kneeling on their knees in the snow with snowboards.

LLM-Empowered Text-to-Image Diffusion



Two surfers looking at the dark stormy looking sky

Four white urinals against a green wall with lines.

Two dogs brown white and black and some people.

An old photograph of three baseball players, one with a bat.

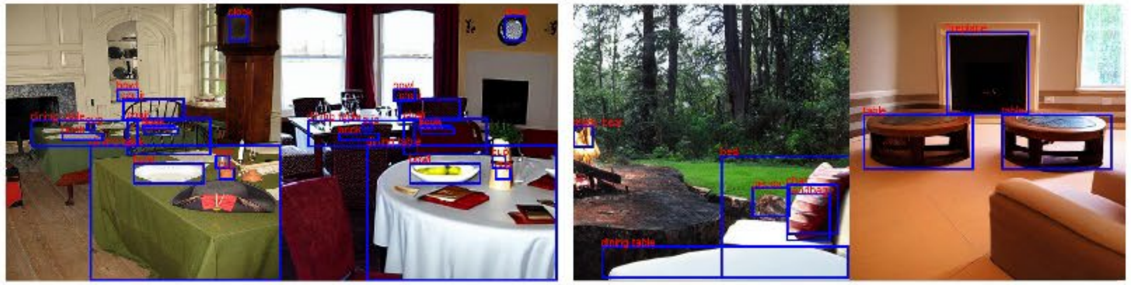
These suitcases are three piece matching set.

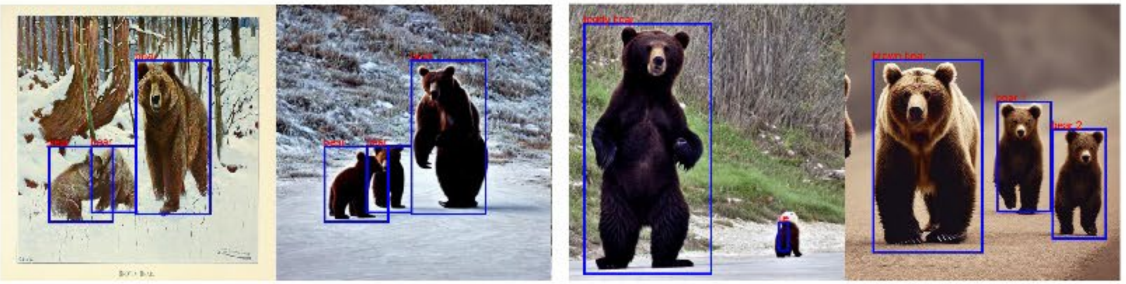
A bed room with two bends and two lamps.

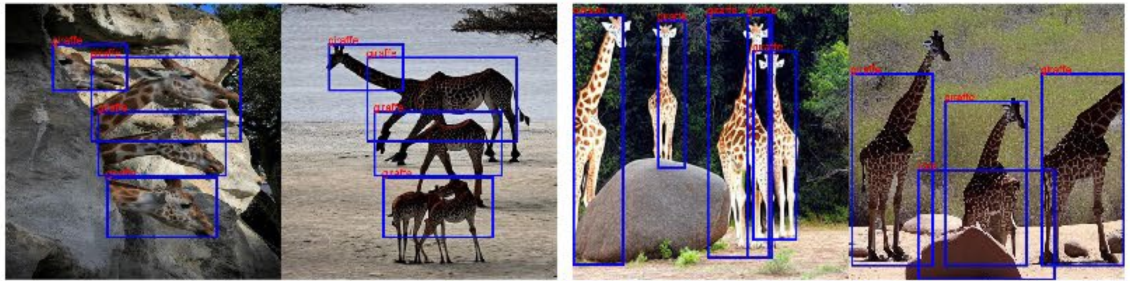
LLM-Empowered Text-to-Image Diffusion

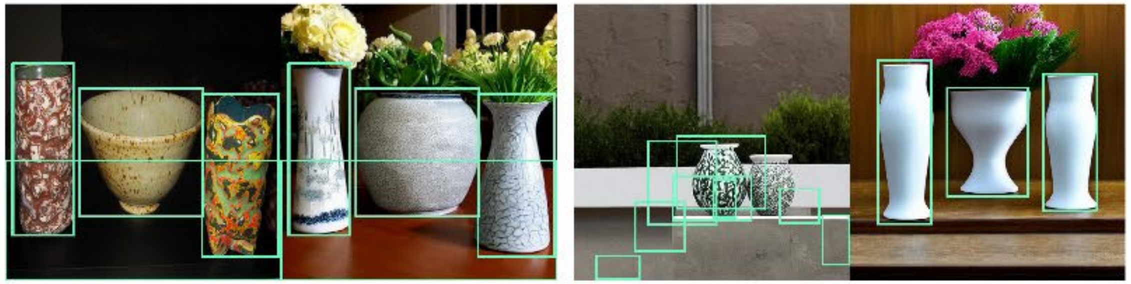
GT **GT*** **LayoutDM** **Ours** **GT** **GT*** **LayoutDM** **Ours**

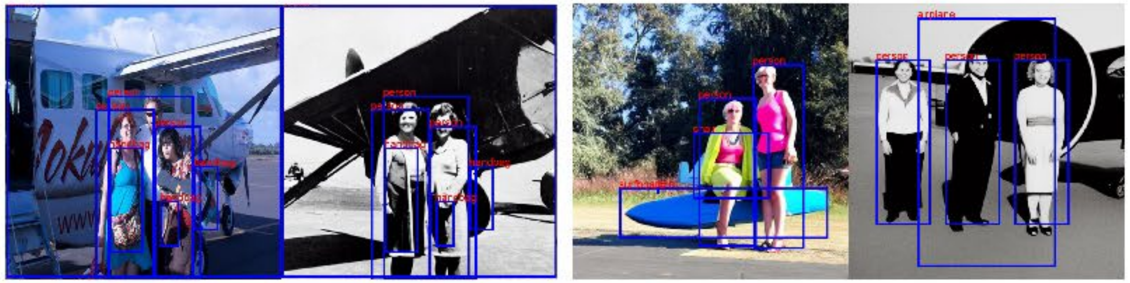
Mixed

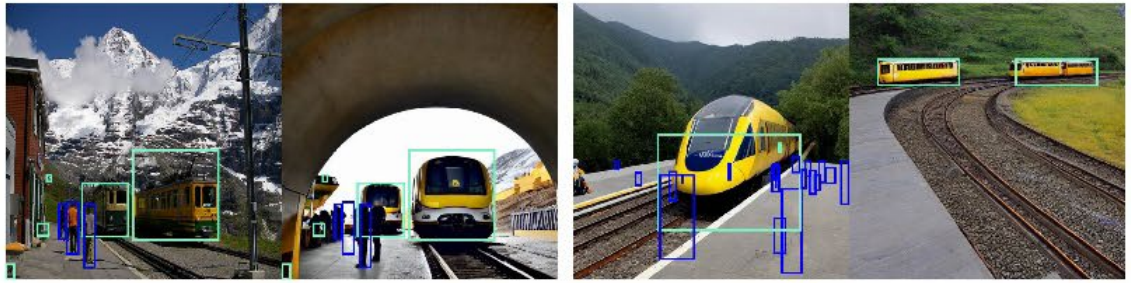

A room with two tables sitting around a fire place.


A brown bear walking next to two small bears.


A group of three giraffe standing next to each other behind a rock.


A group of three vases sitting next to each other.

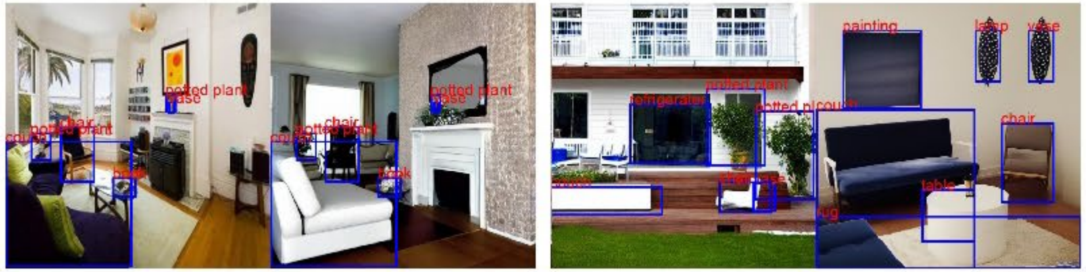

A man and two women are standing beside an airplane.


The two yellow trains are coming down from the mountain.

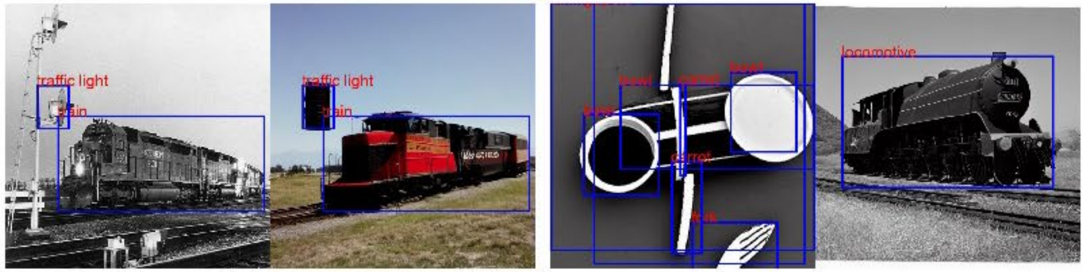
LLM-Empowered Text-to-Image Diffusion

GT **GT*** **LayoutDM** **Ours** **GT** **GT*** **LayoutDM** **Ours**

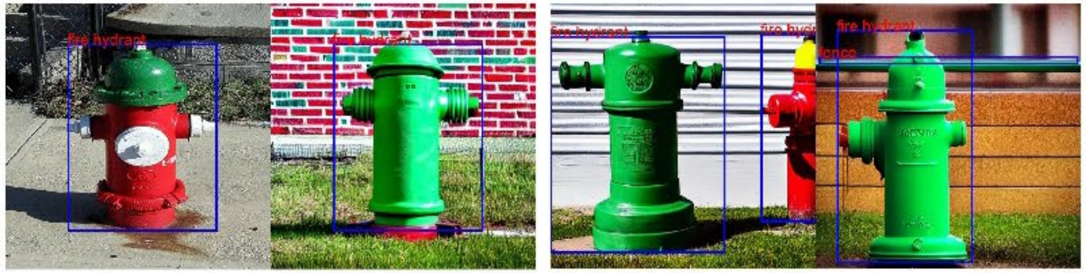
Null



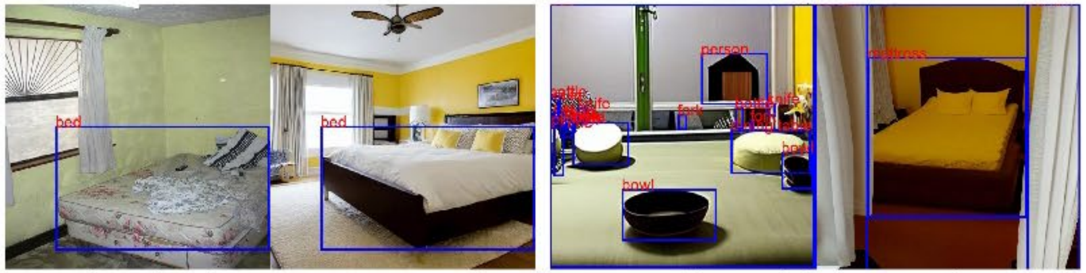
A white living room with blue furniture and decorations.



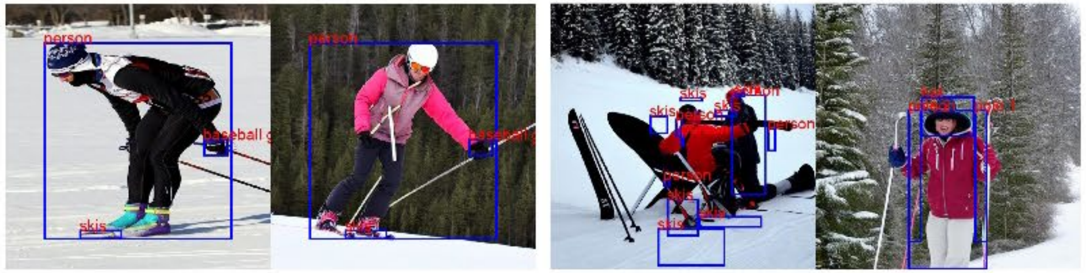
Black and white picture of of a Southern Pacific locomotive.



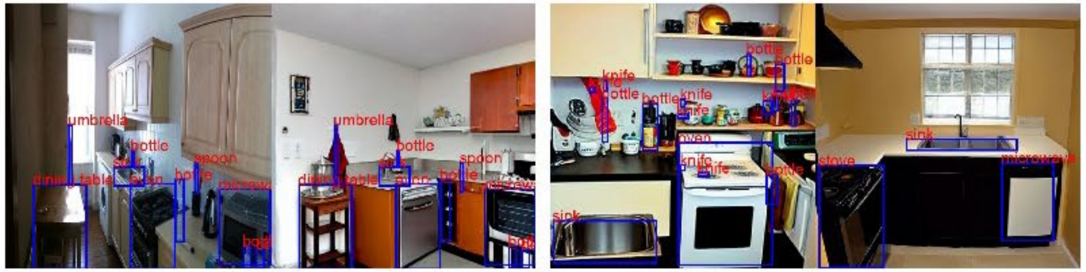
A red white and green fire hydrant and a fence.



Interior bedroom scene of yellow walls, white curtains and a mattress



A person with a hat and some ski poles.



A kitchen with a sink a stove and a microwave.

CONTENT

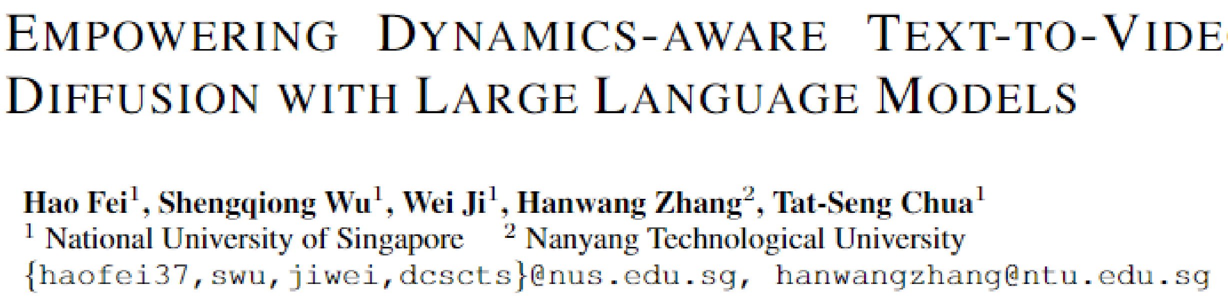
1 LLM Meets Diffusion

2 LLM-Empowered Text-to-Image Diffusion

3 LLM-Empowered Text-to-Video Diffusion

4 Summary and Outlook

LLM-Empowered Text-to-Video Diffusion



EMPOWERING DYNAMICS-AWARE TEXT-TO-VIDEO DIFFUSION WITH LARGE LANGUAGE MODELS

Hao Fei[1], **Shengqiong Wu**[1], **Wei Ji**[1], **Hanwang Zhang**[2], **Tat-Seng Chua**[1]
[1] National University of Singapore [2] Nanyang Technological University
`{haofei37,swu,jiwei,dcscts}@nus.edu.sg, hanwangzhang@ntu.edu.sg`

https://haofei.vip/Dysen-VDM

[1] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Tat-Seng Chua. ***Empowering Dynamics-aware Text-to-Video Diffusion with Large Language Models****. Preprint. 2023.*

LLM-Empowered Text-to-Video Diffusion

Motivation

➢ Diffusion-based Text-to-Video Generation

- *Unsmooth video transition*
- *Crude video motion*
- *Action occurrence disorder*

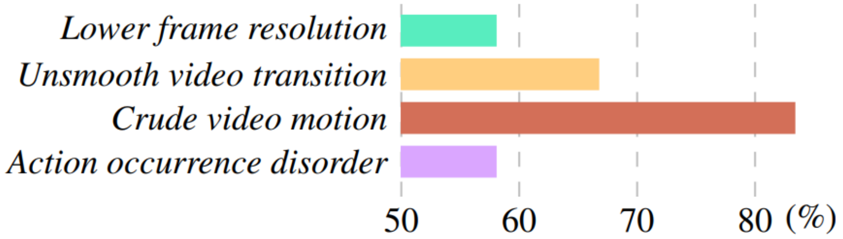


Figure 1: Common issues in the existing text-to-video (T2V) synthesis. We run the video diffusion model (VDM) [21] with random 100 prompts, and ask different users to summarize the problems.

✓ Real crux of high-quality video synthesis: modeling the intricate **video temporal dynamics**

LLM-Empowered Text-to-Video Diffusion

Method

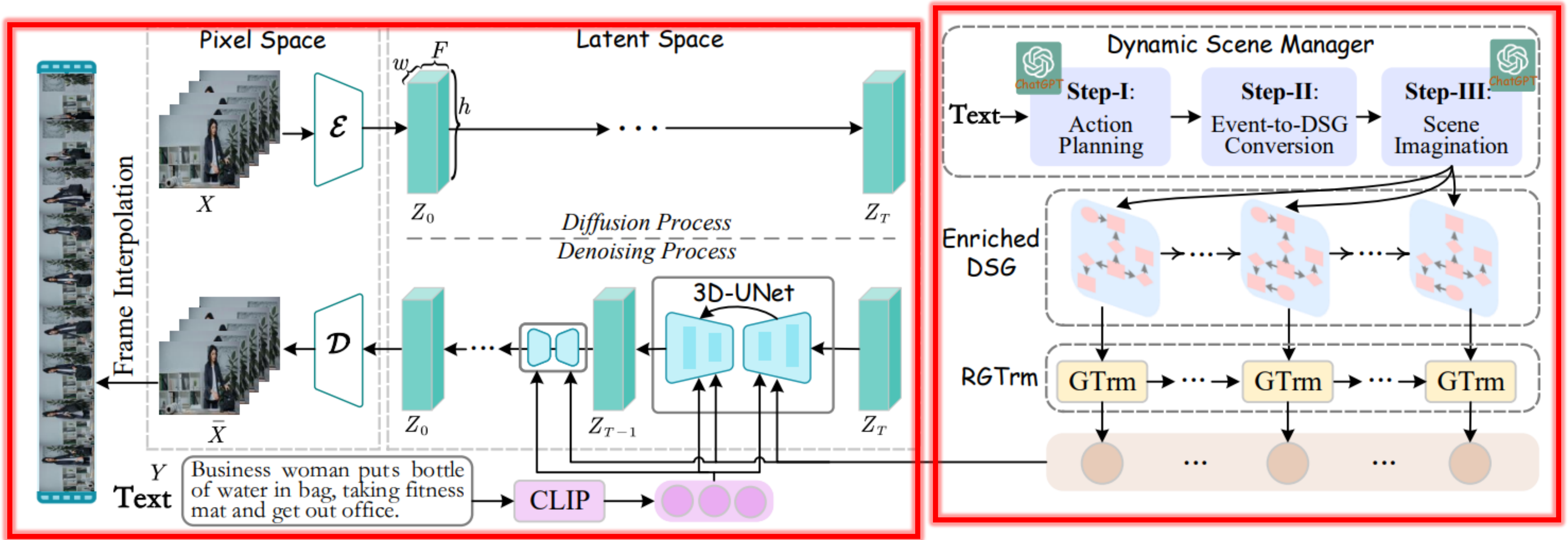


Figure 2: Our dynamics-aware T2V diffusion framework. The dynamic scene manager (Dysen) module operates over the input text prompt and produces the enriched dynamic scene graph (DSG), which is encoded by the recurrent graph Transformer (RGTrm), and the resulting fine-grained spatio-temporal scene features are integrated into the video generation (denoising) process.

LLM-Empowered Text-to-Video Diffusion

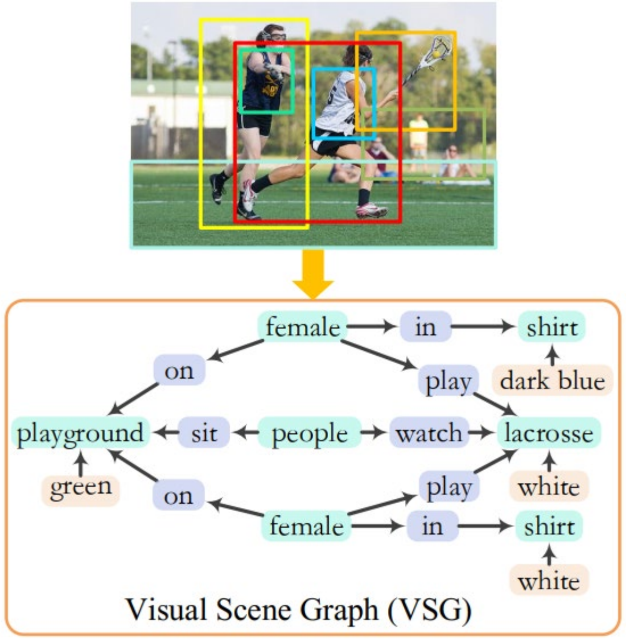
Method

- Dynamic Scene Graph (DSG) Representation
  - Visual Scene Graph (VSG): *Representing visual content into semantic structured representation*
    - **Object Nodes:**

      Visually-seen entity objects
    - **Relation Nodes:**

      describing the semantic relations between objects
    - **Attribute Nodes**

      depicting the objects



Visual Scene Graph (VSG)

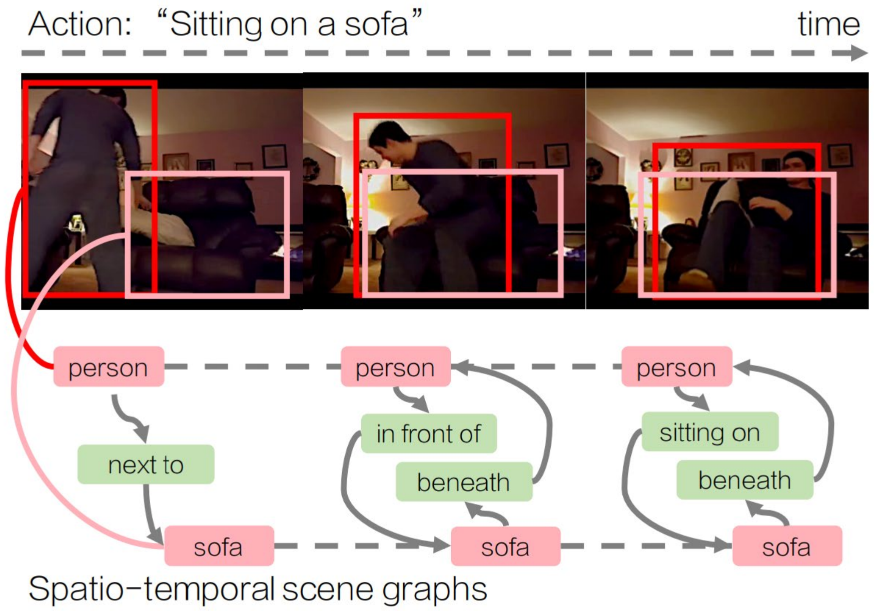
[1] Justin Johnson, etc, and Li Fei-Fei. Image retrieval using scene graphs. CVPR. 2015.

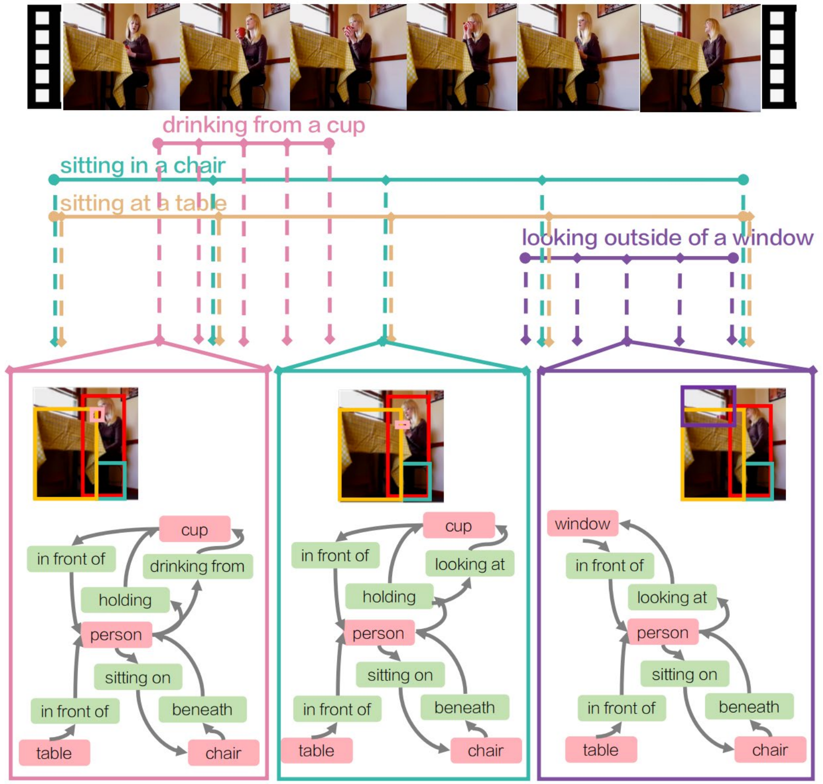
LLM-Empowered Text-to-Video Diffusion

Method

- Dynamic Scene Graph (DSG) Representation

A sequence of VSG along time frames.



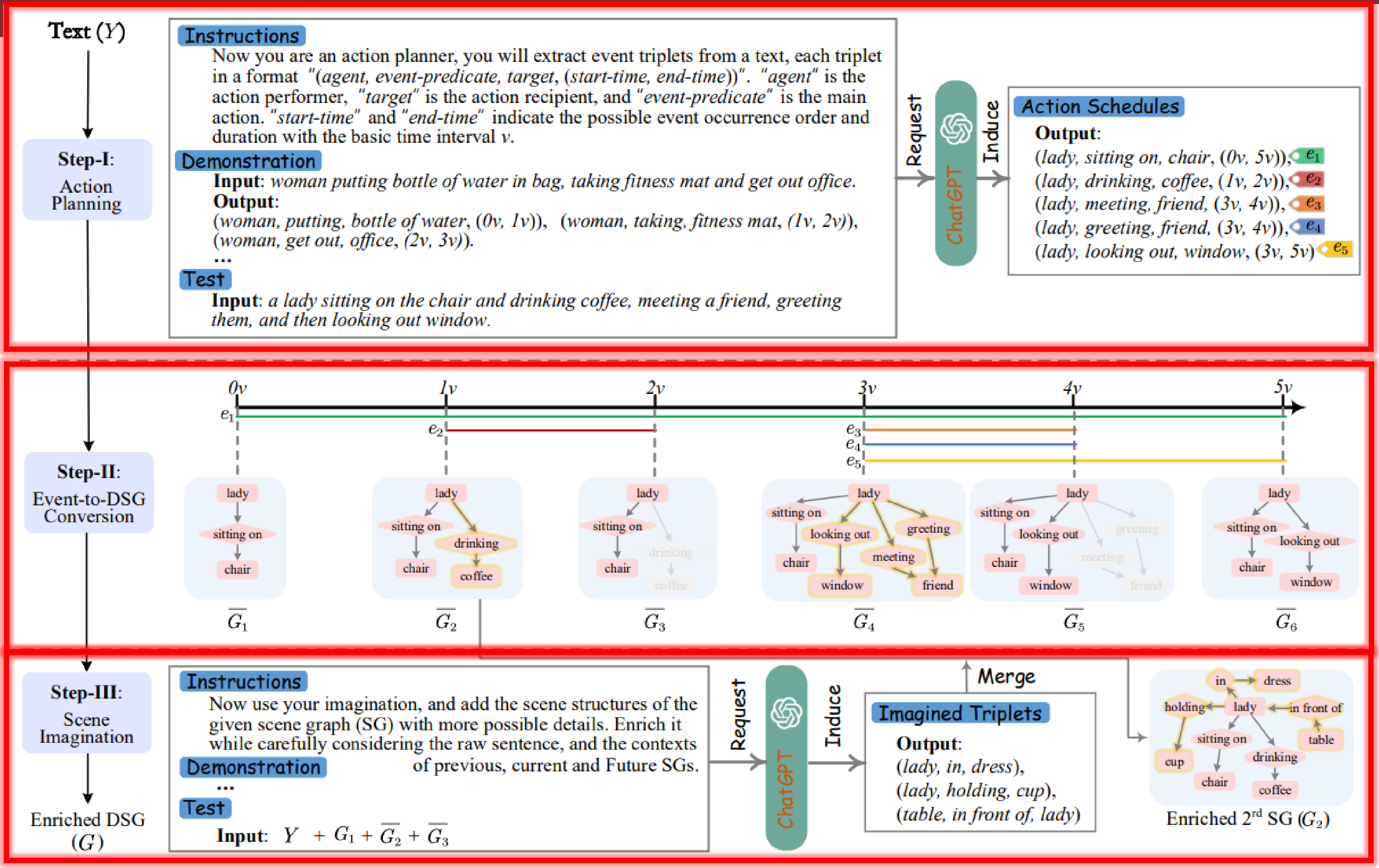


[1] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. CVPR, 2020.

LLM-Empowered Text-to-Video Diffusion

Method

- Dysen



Instructions

Now you are an action planner, you will extract event triplets from a text, each triplet in a format "(*agent, event-predicate, target, (start-time, end-time)*)". "*agent*" is the action performer, "*target*" is the action recipient, and "*event-predicate*" is the main action. "*start-time*" and "*end-time*" indicate the possible event occurrence order and duration with the basic time interval *v*.

Demonstration

Input: *woman putting bottle of water in bag, taking fitness mat and get out office.*
Output:
(*woman, putting, bottle of water,* (0*v*, 1*v*)), (*woman, taking, fitness mat,* (1*v*, 2*v*)), (*woman, get out, office,* (2*v*, 3*v*)).
...

Test

Input: *a lady sitting on the chair and drinking coffee, meeting a friend, greeting them, and then looking out window.*

Action Schedules

Output:
(*lady, sitting on, chair,* (0*v*, 5*v*)), e_1
(*lady, drinking, coffee,* (1*v*, 2*v*)), e_2
(*lady, meeting, friend,* (3*v*, 4*v*)), e_3
(*lady, greeting, friend,* (3*v*, 4*v*)), e_4
(*lady, looking out, window,* (3*v*, 5*v*)) e_5

Instructions

Now use your imagination, and add the scene structures of the given scene graph (SG) with more possible details. Enrich it while carefully considering the raw sentence, and the contexts of previous, current and Future SGs.

Demonstration

...

Test

Input: $Y + G_1 + \overline{G}_2 + \overline{G}_3$

Imagined Triplets

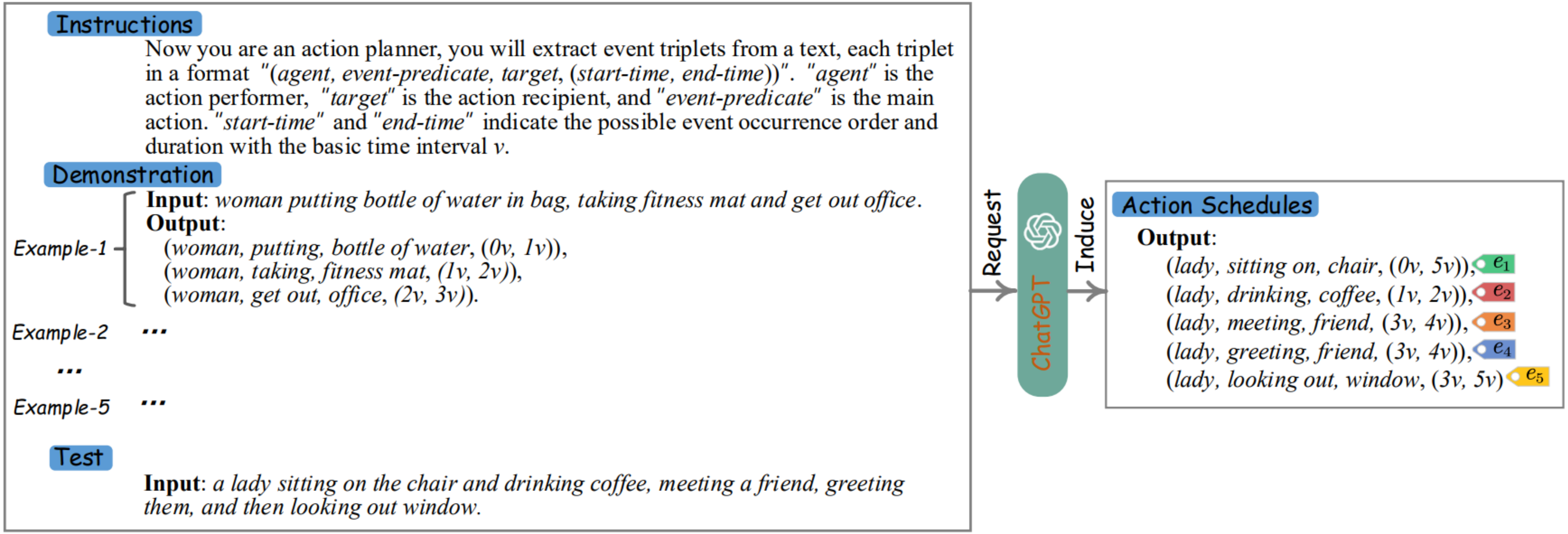
Output:
(*lady, in, dress*),
(*lady, holding, cup*),
(*table, in front of, lady*)

Figure 3: Based on the given text, Dysen module carries out three steps of operations to obtain the enriched DSG: 1) action planning, 2) event-to-DSG conversion, and 3) scene imagination, where we take advantage of the ChatGPT with in-context learning. Best viewed by zooming in.

LLM-Empowered Text-to-Video Diffusion

Method

- Step-I, ICL for action planning



Instructions

Now you are an action planner, you will extract event triplets from a text, each triplet in a format "(*agent, event-predicate, target*, (*start-time, end-time*))". "*agent*" is the action performer, "*target*" is the action recipient, and "*event-predicate*" is the main action. "*start-time*" and "*end-time*" indicate the possible event occurrence order and duration with the basic time interval v.

Demonstration

Example-1

Input: *woman putting bottle of water in bag, taking fitness mat and get out office*.
Output:
(*woman, putting, bottle of water*, $(0v, 1v)$),
(*woman, taking, fitness mat*, $(1v, 2v)$),
(*woman, get out, office*, $(2v, 3v)$).

Example-2 ...

...

Example-5 ...

Test

Input: *a lady sitting on the chair and drinking coffee, meeting a friend, greeting them, and then looking out window.*

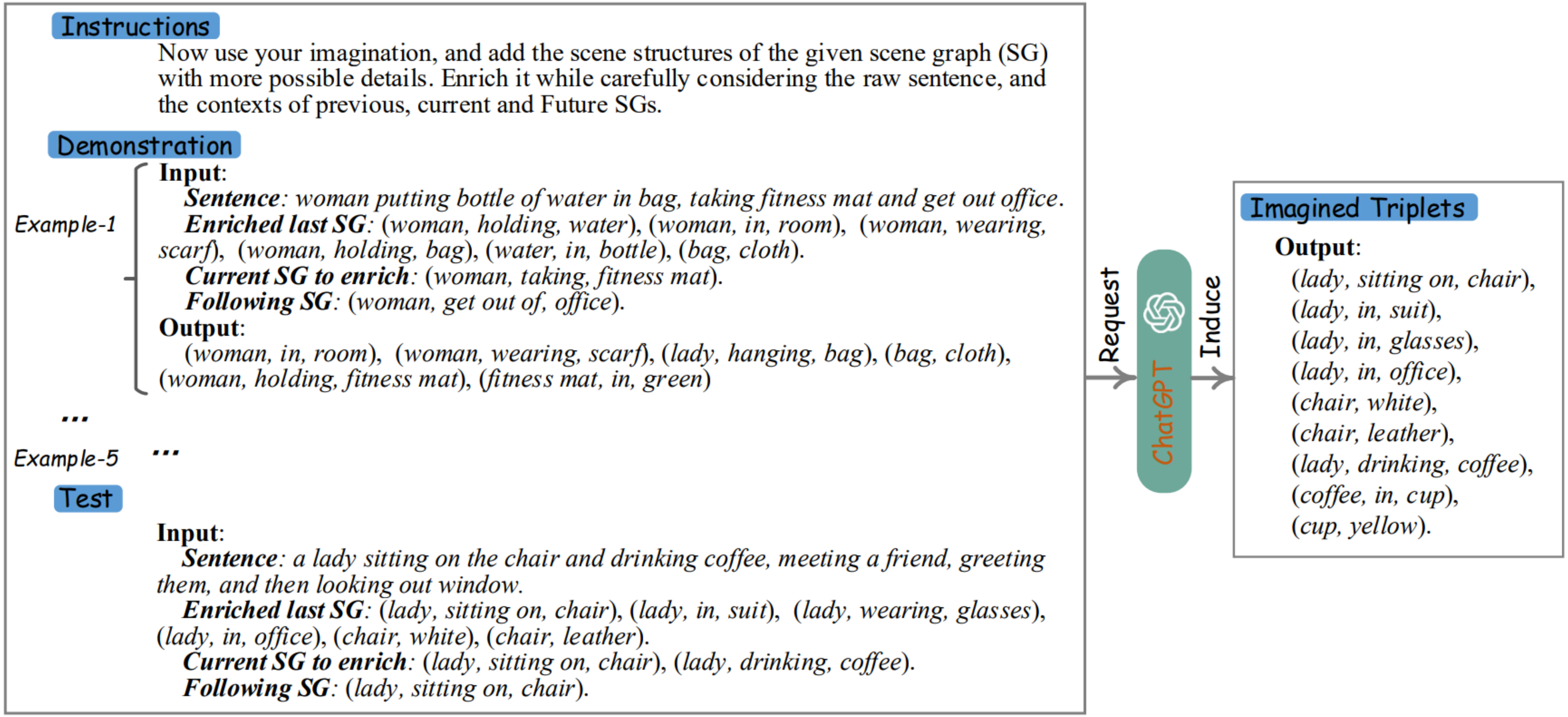
Action Schedules

Output:
(*lady, sitting on, chair*, $(0v, 5v)$), e_1
(*lady, drinking, coffee*, $(1v, 2v)$), e_2
(*lady, meeting, friend*, $(3v, 4v)$), e_3
(*lady, greeting, friend*, $(3v, 4v)$), e_4
(*lady, looking out, window*, $(3v, 5v)$) e_5

LLM-Empowered Text-to-Video Diffusion

Method

- Step-II, ICL for step-wise scene imagination



Instructions

Now use your imagination, and add the scene structures of the given scene graph (SG) with more possible details. Enrich it while carefully considering the raw sentence, and the contexts of previous, current and Future SGs.

Demonstration

Example-1

Input:

Sentence: *woman putting bottle of water in bag, taking fitness mat and get out office*.

Enriched last SG: (*woman, holding, water*), (*woman, in, room*), (*woman, wearing, scarf*), (*woman, holding, bag*), (*water, in, bottle*), (*bag, cloth*).

Current SG to enrich: (*woman, taking, fitness mat*).

Following SG: (*woman, get out of, office*).

Output:

(*woman, in, room*), (*woman, wearing, scarf*), (*lady, hanging, bag*), (*bag, cloth*), (*woman, holding, fitness mat*), (*fitness mat, in, green*)

...

Example-5 ...

Test

Input:

Sentence: *a lady sitting on the chair and drinking coffee, meeting a friend, greeting them, and then looking out window.*

Enriched last SG: (*lady, sitting on, chair*), (*lady, in, suit*), (*lady, wearing, glasses*), (*lady, in, office*), (*chair, white*), (*chair, leather*).

Current SG to enrich: (*lady, sitting on, chair*), (*lady, drinking, coffee*).

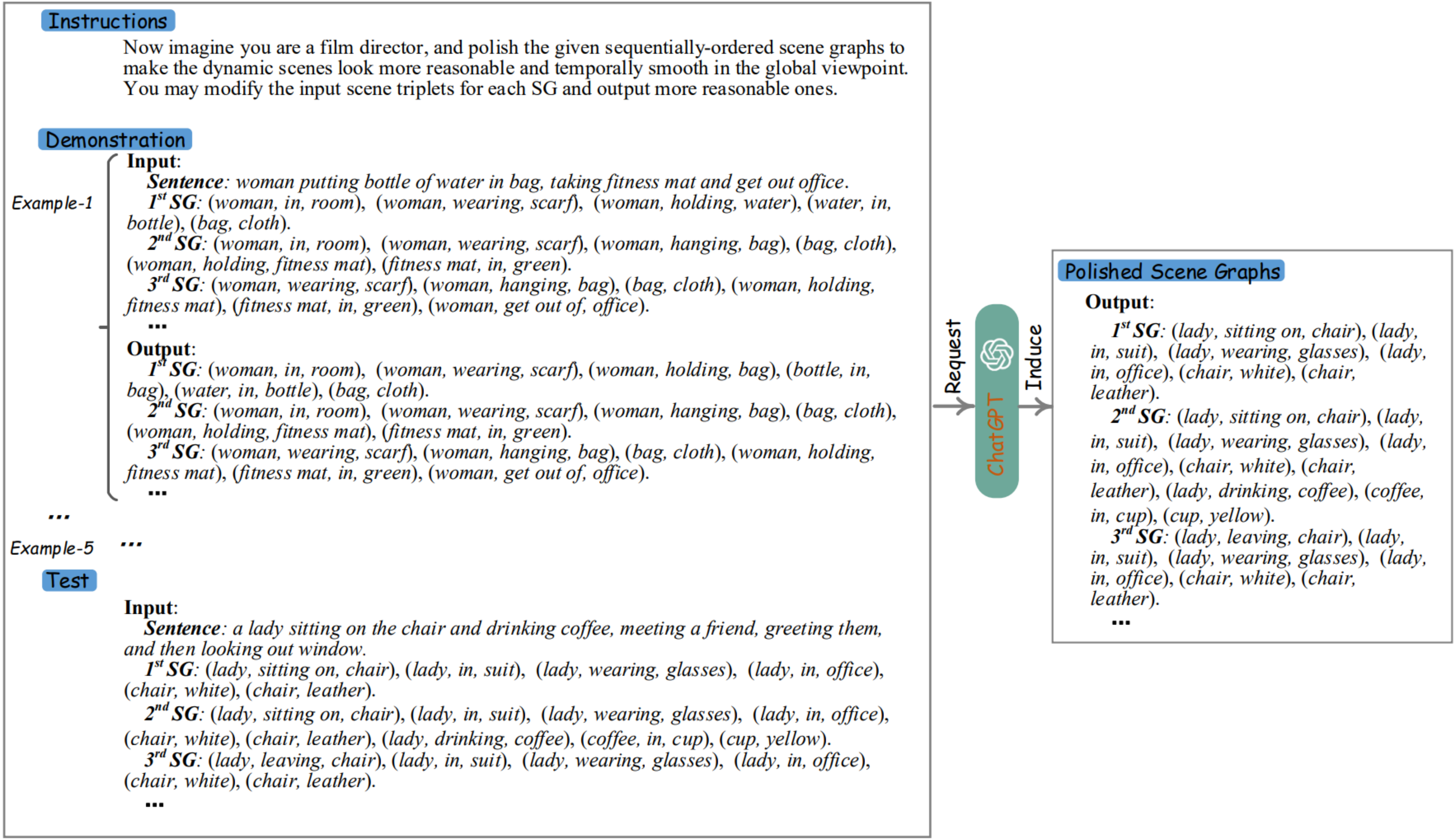
Following SG: (*lady, sitting on, chair*).

Request → ChatGPT → Induce

Imagined Triplets

Output:

(*lady, sitting on, chair*),
(*lady, in, suit*),
(*lady, in, glasses*),
(*lady, in, office*),
(*chair, white*),
(*chair, leather*),
(*lady, drinking, coffee*),
(*coffee, in, cup*),
(*cup, yellow*).



Instructions

Now imagine you are a film director, and polish the given sequentially-ordered scene graphs to make the dynamic scenes look more reasonable and temporally smooth in the global viewpoint. You may modify the input scene triplets for each SG and output more reasonable ones.

Demonstration

Example-1

Input:

Sentence: *woman putting bottle of water in bag, taking fitness mat and get out office.*

1st SG: (*woman, in, room*), (*woman, wearing, scarf*), (*woman, holding, water*), (*water, in, bottle*), (*bag, cloth*).

2nd SG: (*woman, in, room*), (*woman, wearing, scarf*), (*woman, hanging, bag*), (*bag, cloth*), (*woman, holding, fitness mat*), (*fitness mat, in, green*).

3rd SG: (*woman, wearing, scarf*), (*woman, hanging, bag*), (*bag, cloth*), (*woman, holding, fitness mat*), (*fitness mat, in, green*), (*woman, get out of, office*).

...

Output:

1st SG: (*woman, in, room*), (*woman, wearing, scarf*), (*woman, holding, bag*), (*bottle, in, bag*), (*water, in, bottle*), (*bag, cloth*).

2nd SG: (*woman, in, room*), (*woman, wearing, scarf*), (*woman, hanging, bag*), (*bag, cloth*), (*woman, holding, fitness mat*), (*fitness mat, in, green*).

3rd SG: (*woman, wearing, scarf*), (*woman, hanging, bag*), (*bag, cloth*), (*woman, holding, fitness mat*), (*fitness mat, in, green*), (*woman, get out of, office*).

...

...

Example-5 **...**

Test

Input:

Sentence: *a lady sitting on the chair and drinking coffee, meeting a friend, greeting them, and then looking out window.*

1st SG: (*lady, sitting on, chair*), (*lady, in, suit*), (*lady, wearing, glasses*), (*lady, in, office*), (*chair, white*), (*chair, leather*).

2nd SG: (*lady, sitting on, chair*), (*lady, in, suit*), (*lady, wearing, glasses*), (*lady, in, office*), (*chair, white*), (*chair, leather*), (*lady, drinking, coffee*), (*coffee, in, cup*), (*cup, yellow*).

3rd SG: (*lady, leaving, chair*), (*lady, in, suit*), (*lady, wearing, glasses*), (*lady, in, office*), (*chair, white*), (*chair, leather*).

...

Polished Scene Graphs

Output:

1st SG: (*lady, sitting on, chair*), (*lady, in, suit*), (*lady, wearing, glasses*), (*lady, in, office*), (*chair, white*), (*chair, leather*).

2nd SG: (*lady, sitting on, chair*), (*lady, in, suit*), (*lady, wearing, glasses*), (*lady, in, office*), (*chair, white*), (*chair, leather*), (*lady, drinking, coffee*), (*coffee, in, cup*), (*cup, yellow*).

3rd SG: (*lady, leaving, chair*), (*lady, in, suit*), (*lady, wearing, glasses*), (*lady, in, office*), (*chair, white*), (*chair, leather*).

...

LLM-Empowered Text-to-Video Diffusion

Experiment

- Result on zero-shot T2V

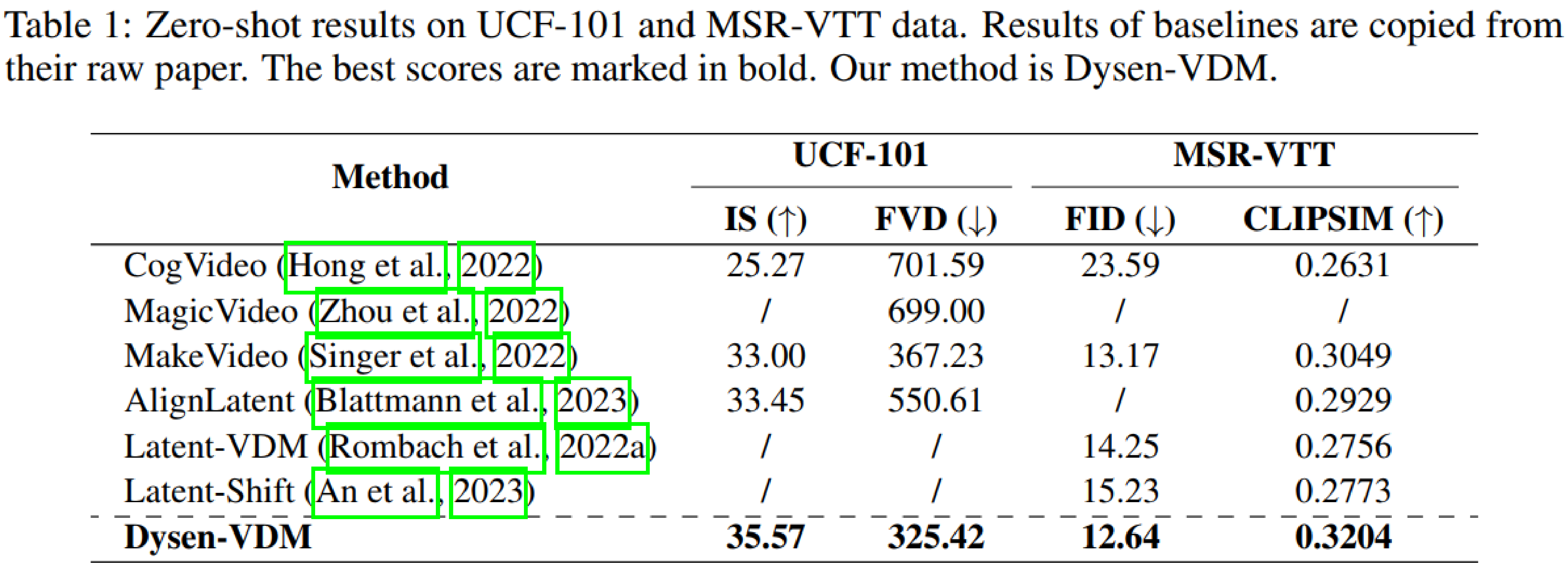


Table 1: Zero-shot results on UCF-101 and MSR-VTT data. Results of baselines are copied from their raw paper. The best scores are marked in bold. Our method is Dysen-VDM.

Method	**UCF-101**		**MSR-VTT**	
	IS (↑)	**FVD (↓)**	**FID (↓)**	**CLIPSIM (↑)**
CogVideo (Hong et al., 2022)	25.27	701.59	23.59	0.2631
MagicVideo (Zhou et al., 2022)	/	699.00	/	/
MakeVideo (Singer et al., 2022)	33.00	367.23	13.17	0.3049
AlignLatent (Blattmann et al., 2023)	33.45	550.61	/	0.2929
Latent-VDM (Rombach et al., 2022a)	/	/	14.25	0.2756
Latent-Shift (An et al., 2023)	/	/	15.23	0.2773
Dysen-VDM	**35.57**	**325.42**	**12.64**	**0.3204**

LLM-Empowered Text-to-Video Diffusion

Experiment

- Result on supervised fine-tuned T2V

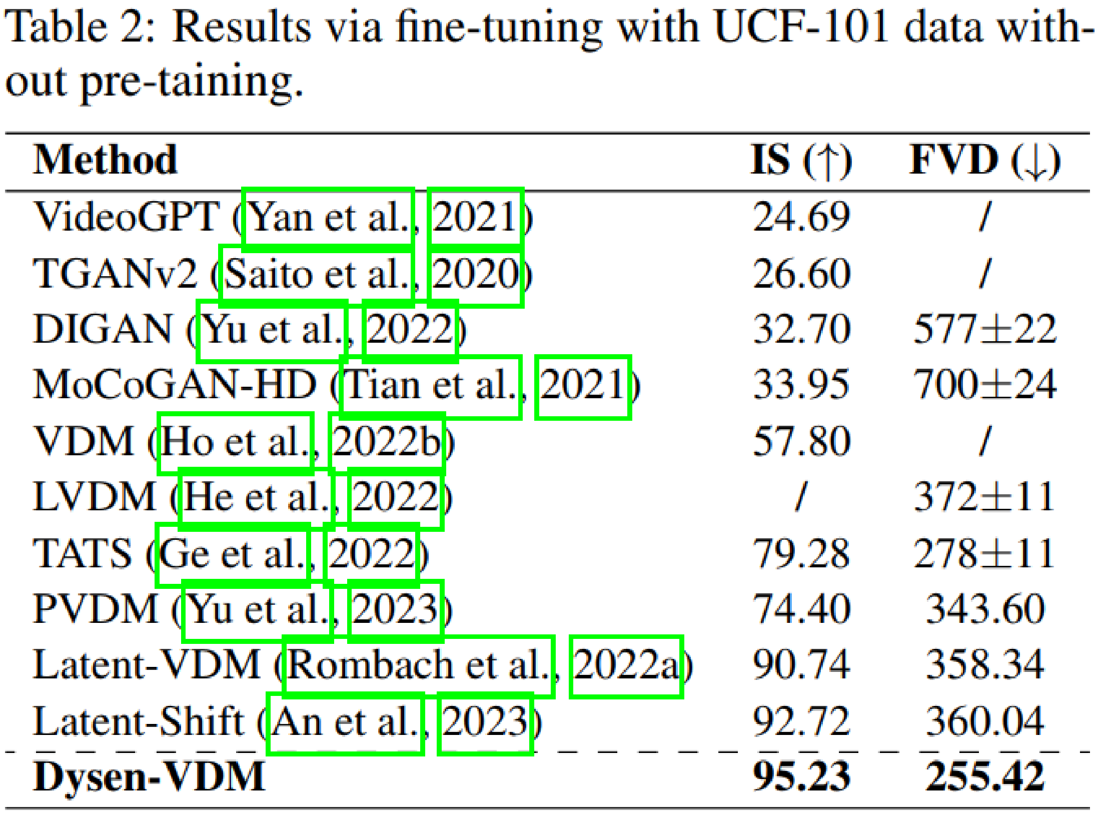


Table 2: Results via fine-tuning with UCF-101 data without pre-taining.

Method	IS (↑)	FVD (↓)
VideoGPT (Yan et al., 2021)	24.69	/
TGANv2 (Saito et al., 2020)	26.60	/
DIGAN (Yu et al., 2022)	32.70	577±22
MoCoGAN-HD (Tian et al., 2021)	33.95	700±24
VDM (Ho et al., 2022b)	57.80	/
LVDM (He et al., 2022)	/	372±11
TATS (Ge et al., 2022)	79.28	278±11
PVDM (Yu et al., 2023)	74.40	343.60
Latent-VDM (Rombach et al., 2022a)	90.74	358.34
Latent-Shift (An et al., 2023)	92.72	360.04
Dysen-VDM	**95.23**	**255.42**

LLM-Empowered Text-to-Video Diffusion

Experiment

- Results on Action-complex T2V Generation

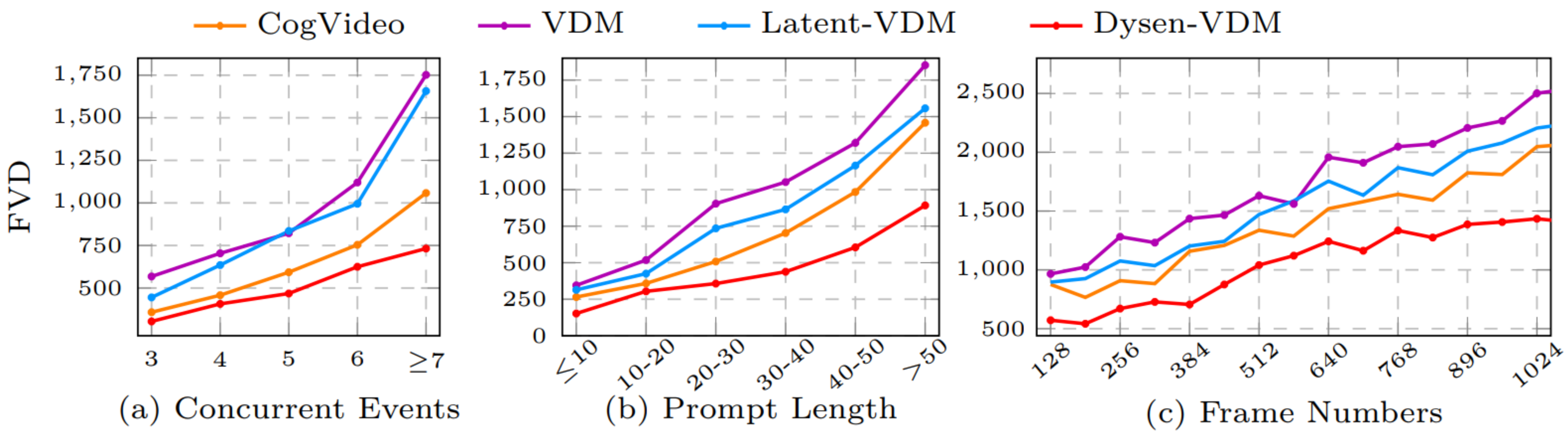


Figure 5: Performance on the action-complex scene video generation of ActivityNet data.

LLM-Empowered Text-to-Video Diffusion

Experiment

- In-depth Analyses
  - *Controllability with DSG*
  - *Change of Scenes*

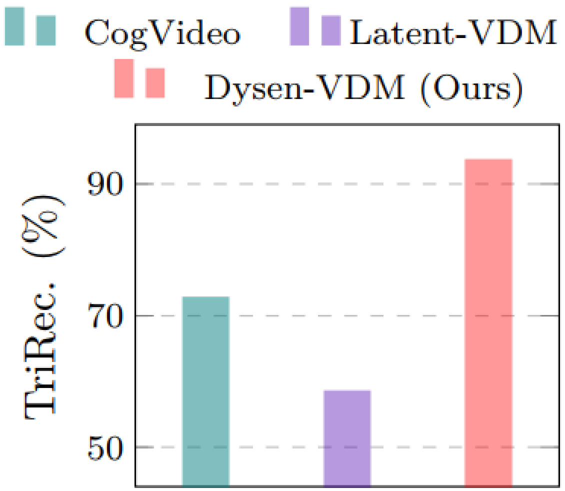


Figure 6: Aligning recall rate (TriRec.) of '*subject-predicate-object*' structures between input text and generated video frames.

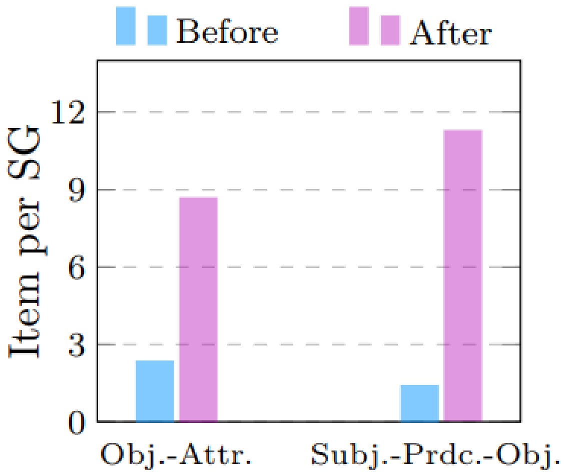
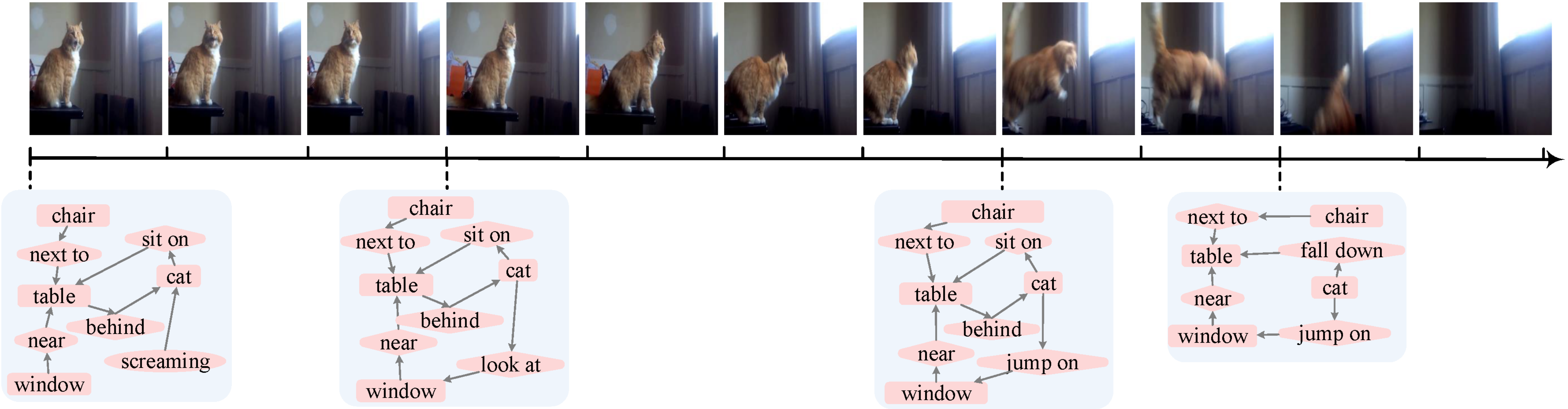
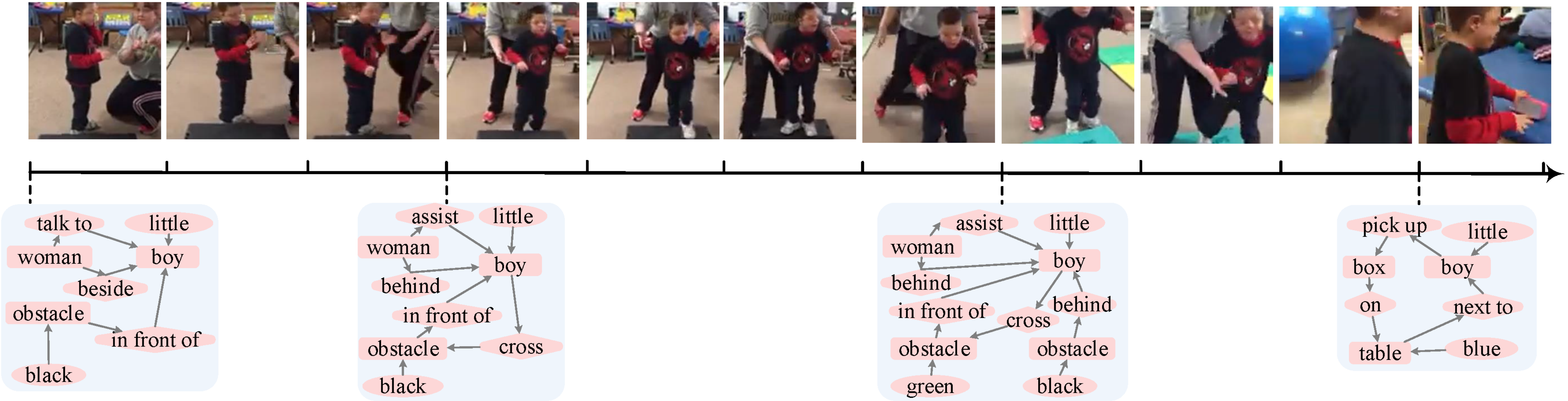


Figure 7: The number of SG structures ('*object-attribute*' and '*subject-predicate-object*') before and after scene imagination.

Text prompt: *A cat is screaming, looks at the window, and wants to jump on it, but falls down the table.*



Text prompt: *A woman told to the little boy, and then she helped the little boy cross two different color of obstacles one by one, and the little boy picked up the pink box on the table.*



LLM-Empowered Text-to-Video Diffusion

Experiment

- Examples

A clownfish swimming with elegance through coral reefs, presenting the beautiful scenery under the sea.

A woman is looking after the plant in her garden, and then she raises her head to observe the weather.

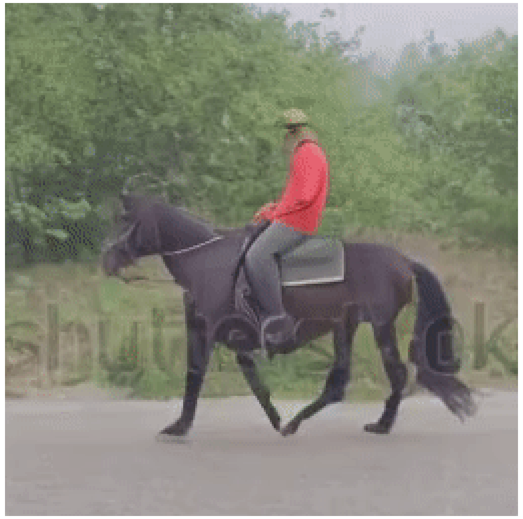
A man dressed as Santa Claus is riding a motorcycle on a big city road.

A horse in a blue cover walks at a fast pace, and then begins to slow down, taking a walk in the paddock.

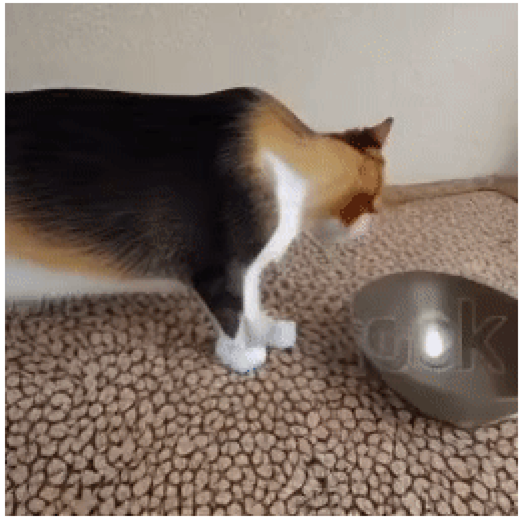
LLM-Empowered Text-to-Video Diffusion

Experiment

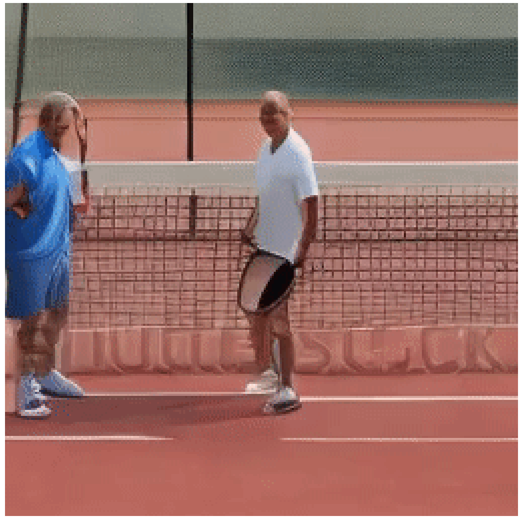
- Examples



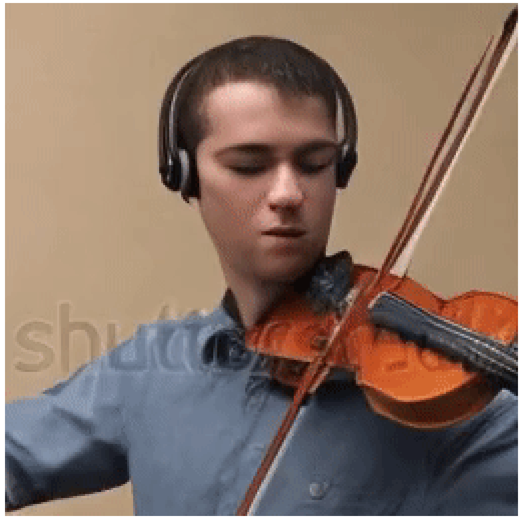
A person in a jacket riding a horse, is walking along the countryside road.



A cat eating food out of a bowl while looking around, then the camera moves away to a scene where another cat eats food.



A man and other man are standing together in the middle of a tennis court, and speaking to the camera.



A young violin player in a neat shirt with a collar, having a headphone on, is playing the violin.

LLM-Empowered Text-to-Video Diffusion

Experiment

- Examples

On a stage, a woman is rotating and waving her arms to show her belly dance.

A band composed of a group of young people is performing live music.

A woman hikes up the green mountain reaches the summit, and takes photos of the breathtaking view.

Two women sit on a park bench, reading books while chatting to each other.

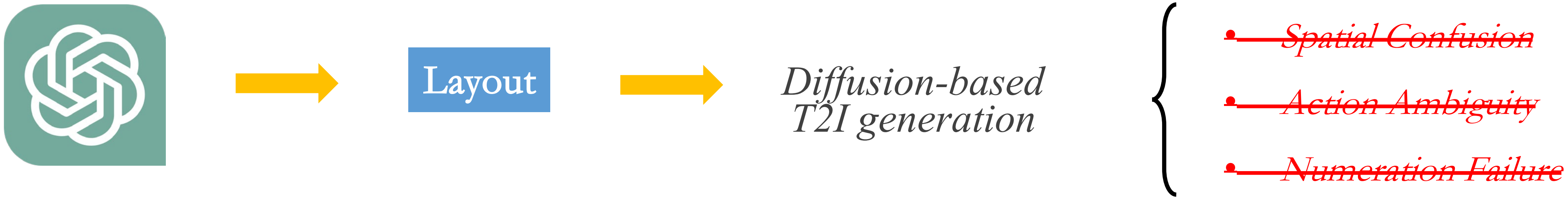
CONTENT

1. LLM Meets Diffusion
2. LLM-Empowered Text-to-Image Diffusion
3. LLM-Empowered Text-to-Vision Diffusion
4. **Summary and Outlook**

Outlook of Future Directions

Summary

1. LLM-Empowered **Text-to-Image** Diffusion

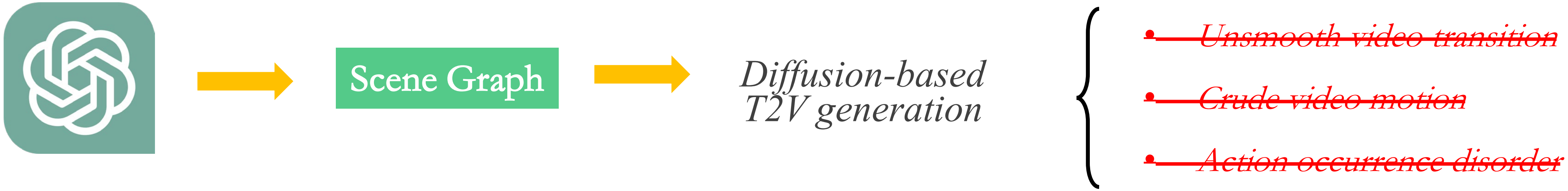


Layout

Diffusion-based T2I generation

- ~~*Spatial Confusion*~~
- ~~*Action Ambiguity*~~
- ~~*Numeration Failure*~~

2. LLM-Empowered **Text-to-Vision** Diffusion



Scene Graph

Diffusion-based T2V generation

- ~~*Unsmooth video transition*~~
- ~~*Crude video motion*~~
- ~~*Action occurrence disorder*~~

Outlook of Future Directions

What Next?

- Inducing other various structured visual representations
  - *overall layout* ✓
  - *dimension*
  - *sketches*
  - *scene structure* ✓
  - *visual coherence*
  - *action dynamics*
  - …

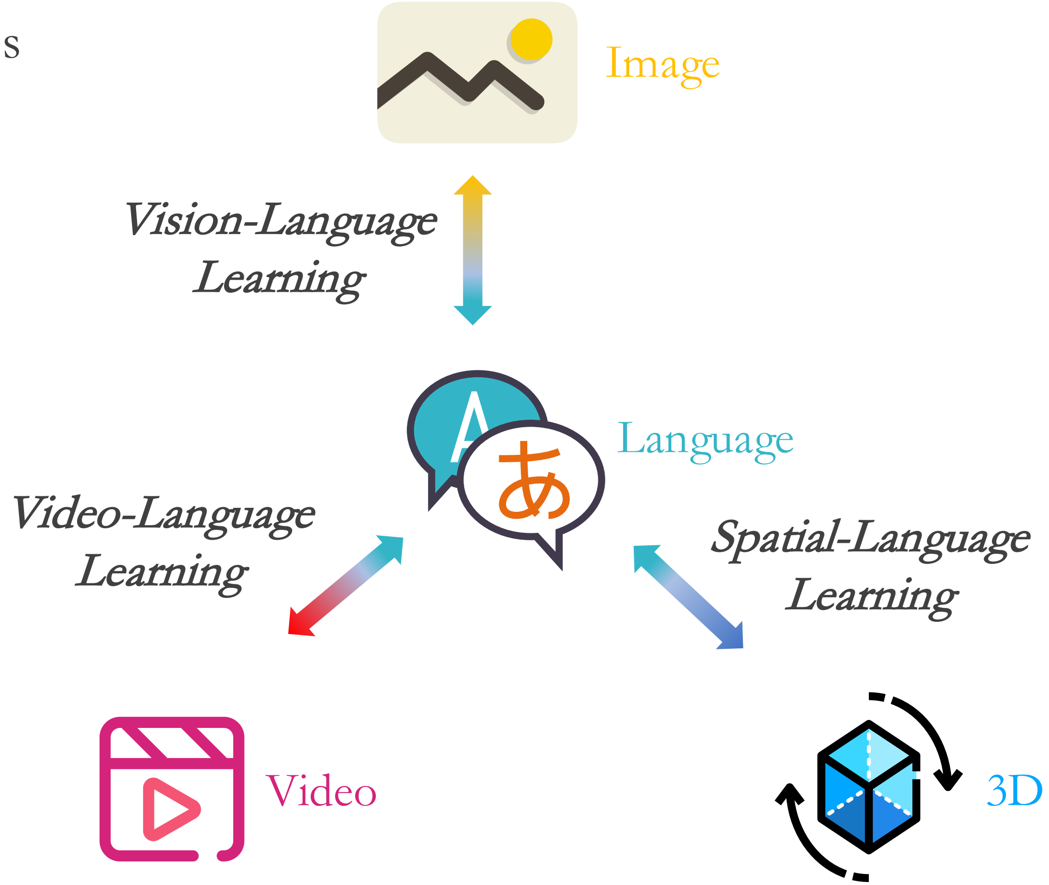
Outlook of Future Directions

What Next?

- Applying the idea to more visual modalities and scenarios
  - Visual modalities

    e.g., 3D
  - More scenarios

    e.g., Editing, In-painting



Image

Vision-Language Learning

Language

Video-Language Learning

Spatial-Language Learning

Video

3D

Outlook of Future Directions

What Next?

- Whether using Multimodal LLM help foster stronger visual understanding?

- Text-based LLM

  ChatGPT
  LLaMA
  Vicuna
  FlanT5
  …

- Multimodal LLM

  Blip-2
  MiniGPT-4
  mPLUG
  MMICL
  …

CONTENT

5 Extra delivery

Universal Structured NLP (XNLP) Demo

XNLP: An Interactive Demonstration System for Universal Structured NLP

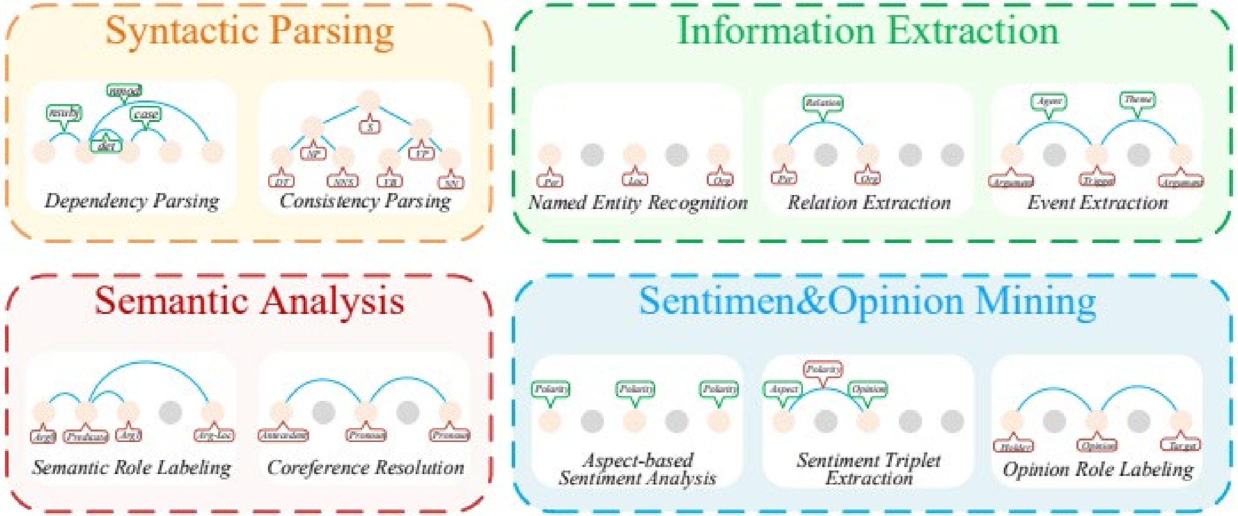
https://xnlp.haofei.vip/

[1] Hao Fei, Meishan Zhang, Min Zhang, Tat-Seng Chua. XNLP: An Interactive Demonstration System for Universal Structured NLP. 2023.

XNLP Demo

Motivation

- ➢ **Structured Natural Language Processing (XNLP)**
  - Many NLP tasks can be reduced into structural predictions
    - 1) textual spans
    - 2) relations between spans



More Emerging XNLP Tasks to Define ···

XNLP Demo

Motivation

- Universal XNLP
  - Unified Sentiment Analysis
  - Universal Information Extraction

☐ a comprehensive and effective approach for unifying all XNLP tasks is not fully established.

- Unification with LLM

?

✓ One model for all

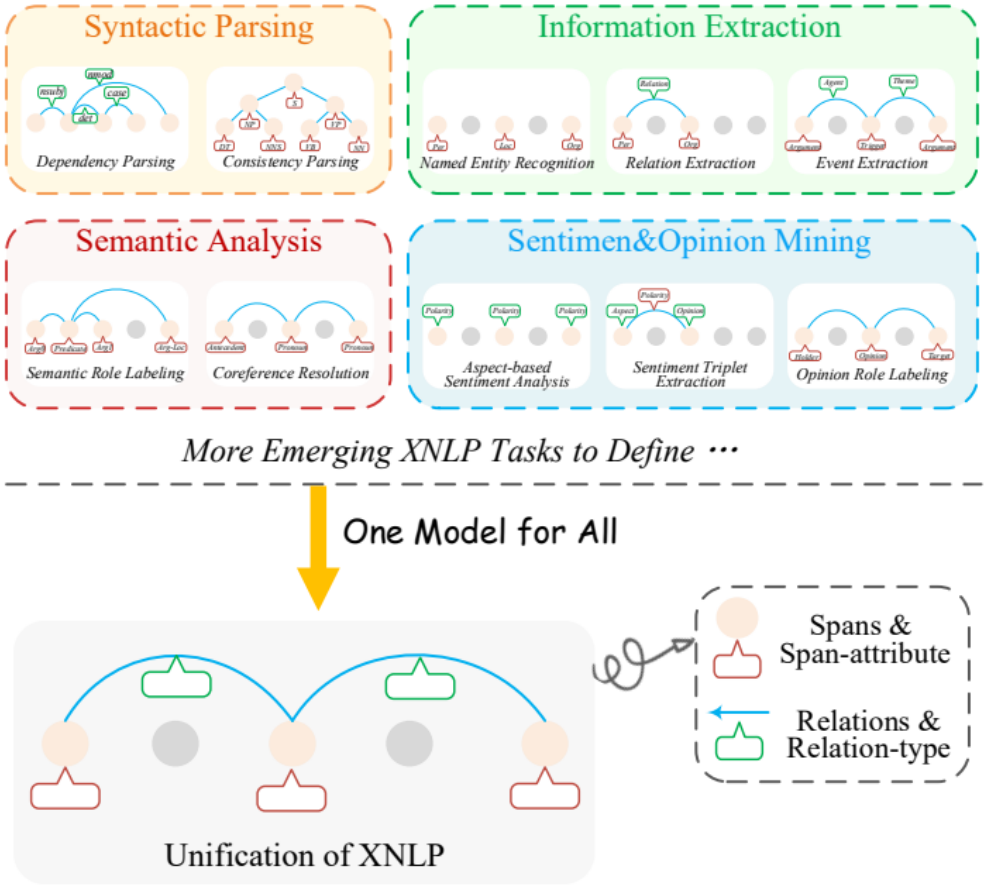
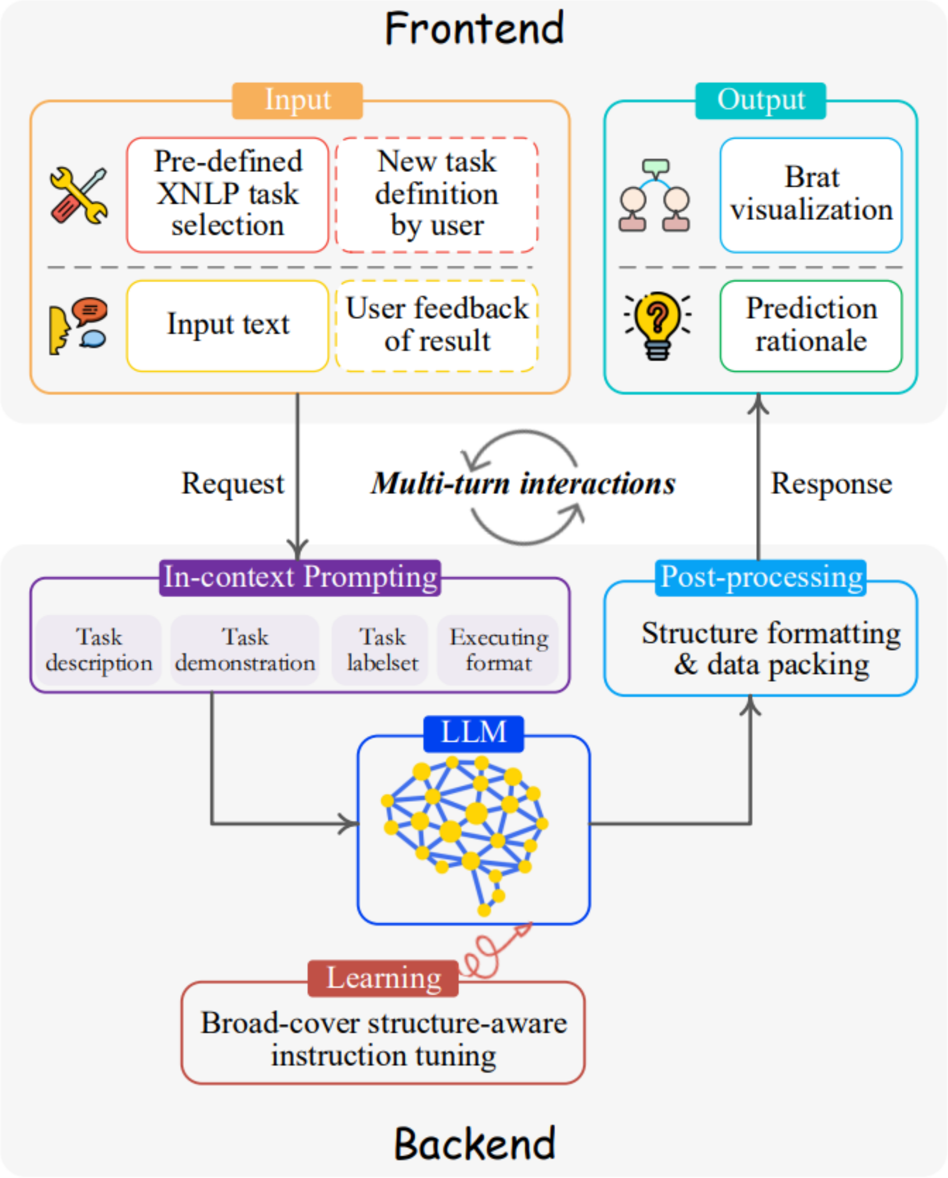


Figure 1: Illustration of the Structured NLP (XNLP) tasks, and the unification of XNLP by decomposing into the predictions of spans and relations.

XNLP Demo

Demo System

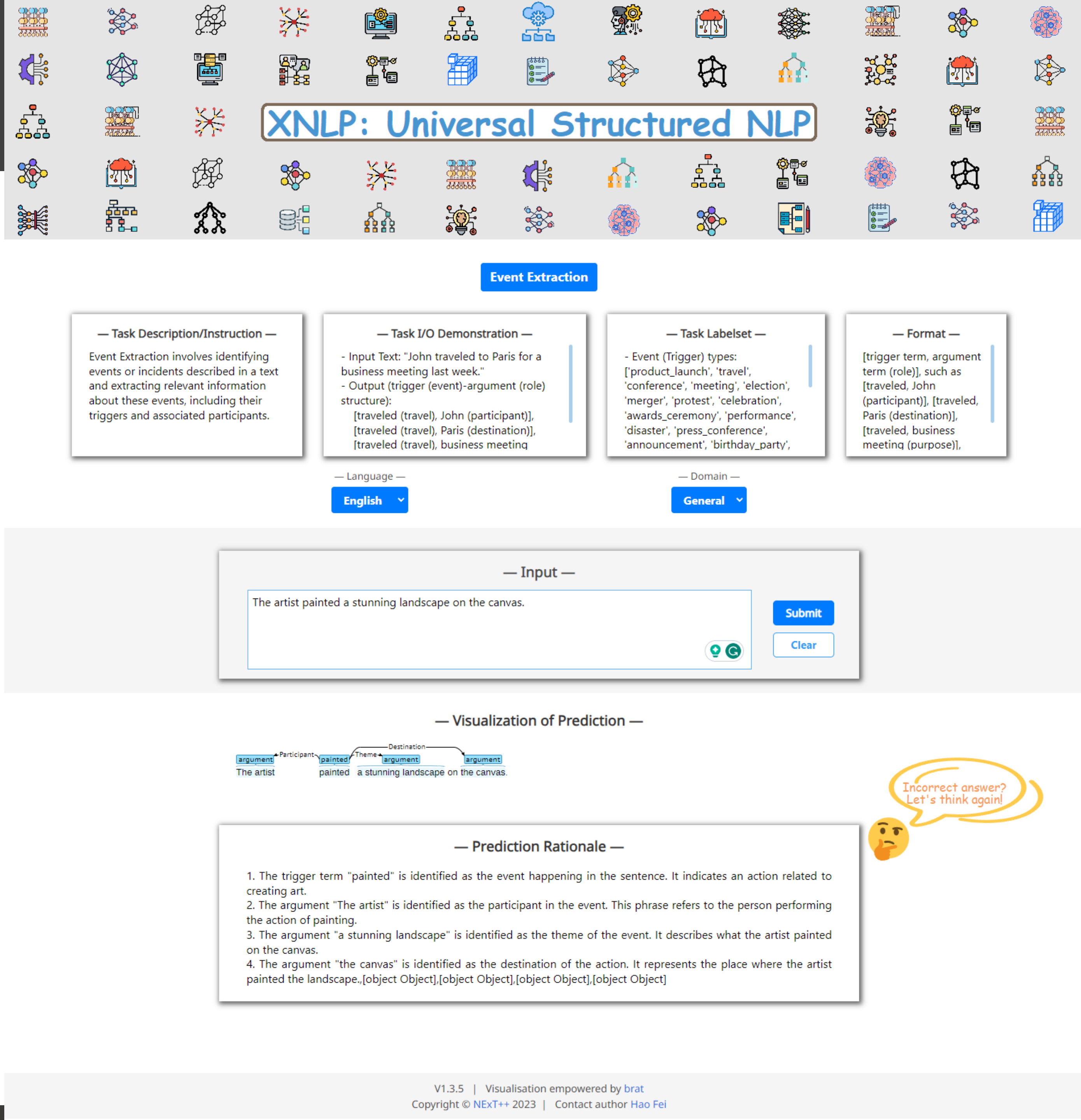
- System Design



XNLP Demo

Demo System

- Screenshot



XNLP: Universal Structured NLP

— Task Description/Instruction —

Event Extraction involves identifying events or incidents described in a text and extracting relevant information about these events, including their triggers and associated participants.

— Task I/O Demonstration —

- Input Text: "John traveled to Paris for a business meeting last week."
- Output (trigger (event)-argument (role) structure):
  [traveled (travel), John (participant)],
  [traveled (travel), Paris (destination)],
  [traveled (travel), business meeting

— Task Labelset —

- Event (Trigger) types: ['product_launch', 'travel', 'conference', 'meeting', 'election', 'merger', 'protest', 'celebration', 'awards_ceremony', 'performance', 'disaster', 'press_conference', 'announcement', 'birthday_party',

— Format —

[trigger term, argument term (role)], such as [traveled, John (participant)], [traveled, Paris (destination)], [traveled, business meeting (purpose)],

— Language —

English

— Domain —

General

— Input —

The artist painted a stunning landscape on the canvas.

— Visualization of Prediction —

— Prediction Rationale —

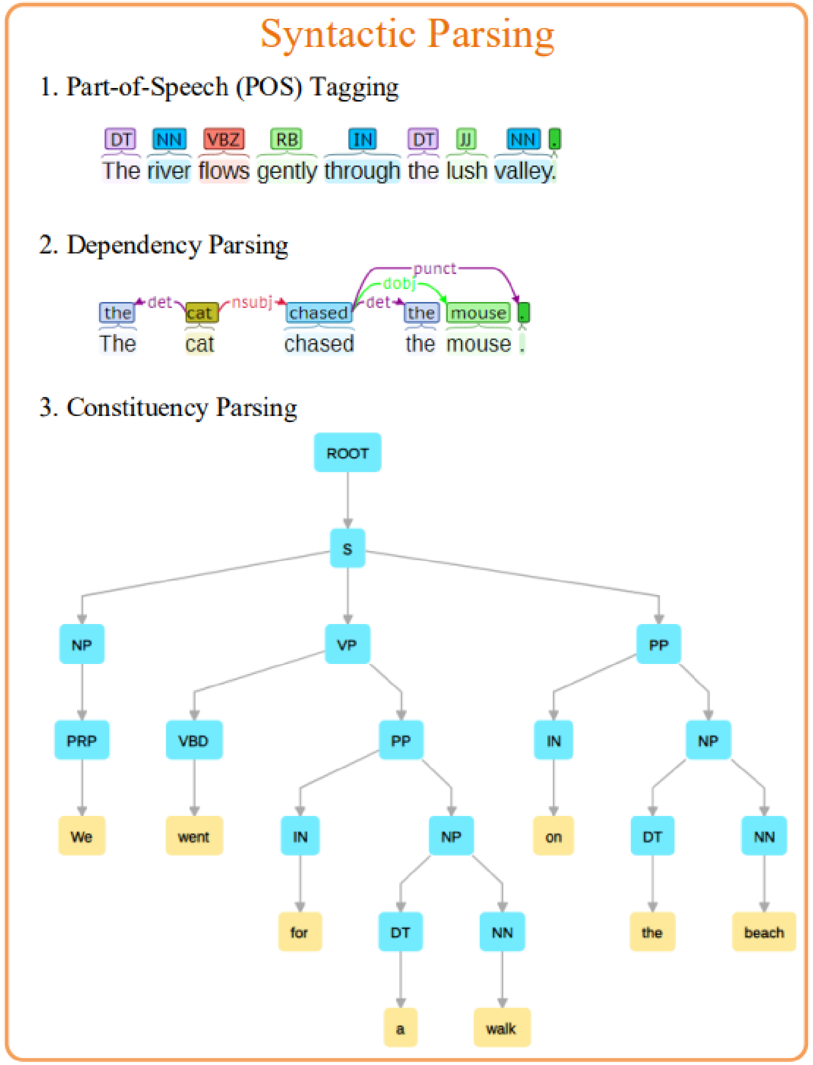
1. The trigger term "painted" is identified as the event happening in the sentence. It indicates an action related to creating art.
2. The argument "The artist" is identified as the participant in the event. This phrase refers to the person performing the action of painting.
3. The argument "a stunning landscape" is identified as the theme of the event. It describes what the artist painted on the canvas.
4. The argument "the canvas" is identified as the destination of the action. It represents the place where the artist painted the landscape.,[object Object],[object Object],[object Object],[object Object]

V1.3.5 | Visualisation empowered by brat
Copyright © NExT++ 2023 | Contact author Hao Fei

XNLP Demo

Demo System

https://xnlp.haofei.vip/



Syntactic Parsing

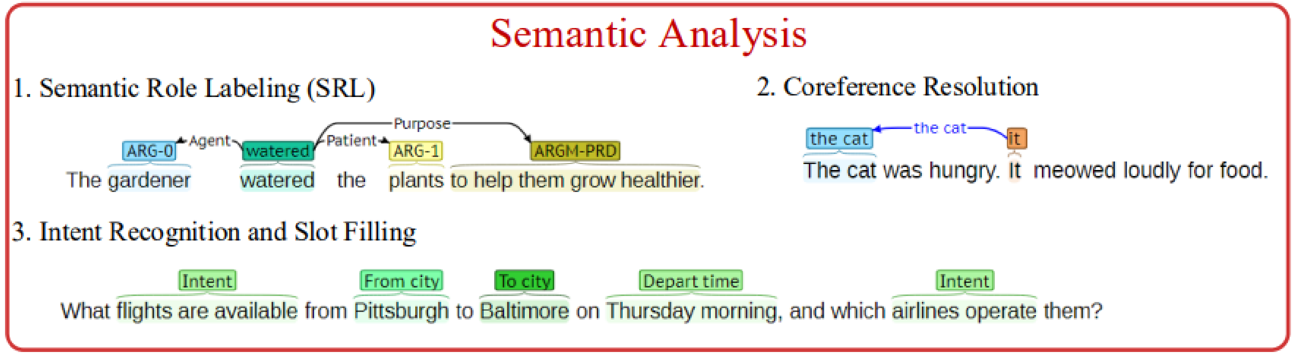
1. Part-of-Speech (POS) Tagging

The river flows gently through the lush valley.

2. Dependency Parsing

The cat chased the mouse .

3. Constituency Parsing



Semantic Analysis

1. Semantic Role Labeling (SRL)

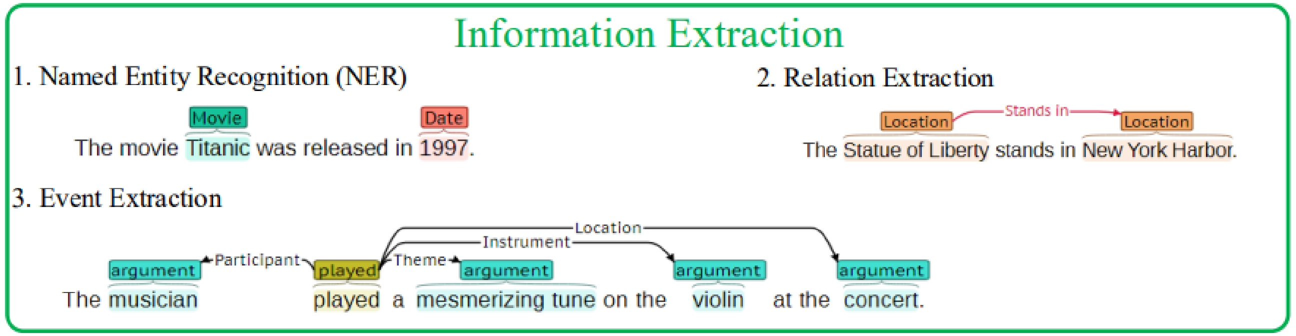
The gardener watered the plants to help them grow healthier.

2. Coreference Resolution

The cat was hungry. It meowed loudly for food.

3. Intent Recognition and Slot Filling

What flights are available from Pittsburgh to Baltimore on Thursday morning, and which airlines operate them?



Information Extraction

1. Named Entity Recognition (NER)

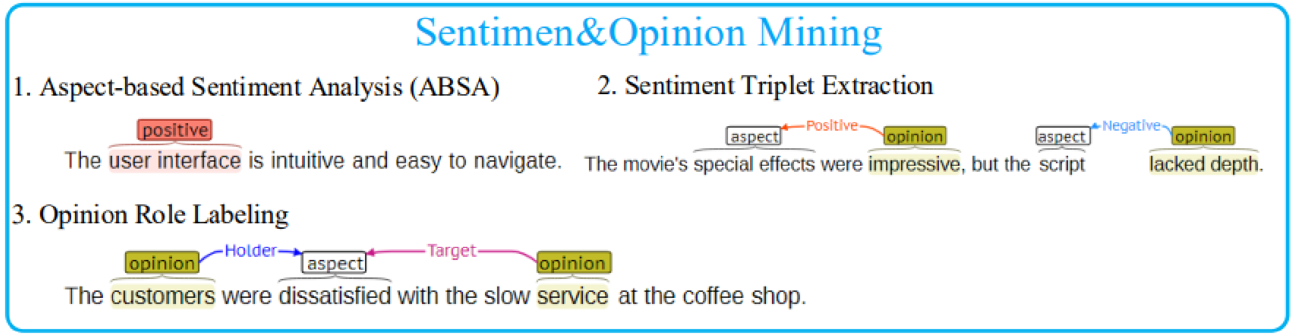
The movie Titanic was released in 1997.

2. Relation Extraction

The Statue of Liberty stands in New York Harbor.

3. Event Extraction

The musician played a mesmerizing tune on the violin at the concert.



Sentimen&Opinion Mining

1. Aspect-based Sentiment Analysis (ABSA)

The user interface is intuitive and easy to navigate.

2. Sentiment Triplet Extraction

The movie's special effects were impressive, but the script lacked depth.

3. Opinion Role Labeling

The customers were dissatisfied with the slow service at the coffee shop.

Thanks

Q&A