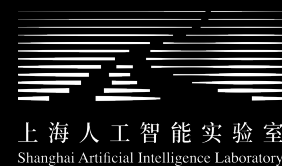




MEGVII 旷视



 **Spotlight Paper**

# DreamLLM ICLR

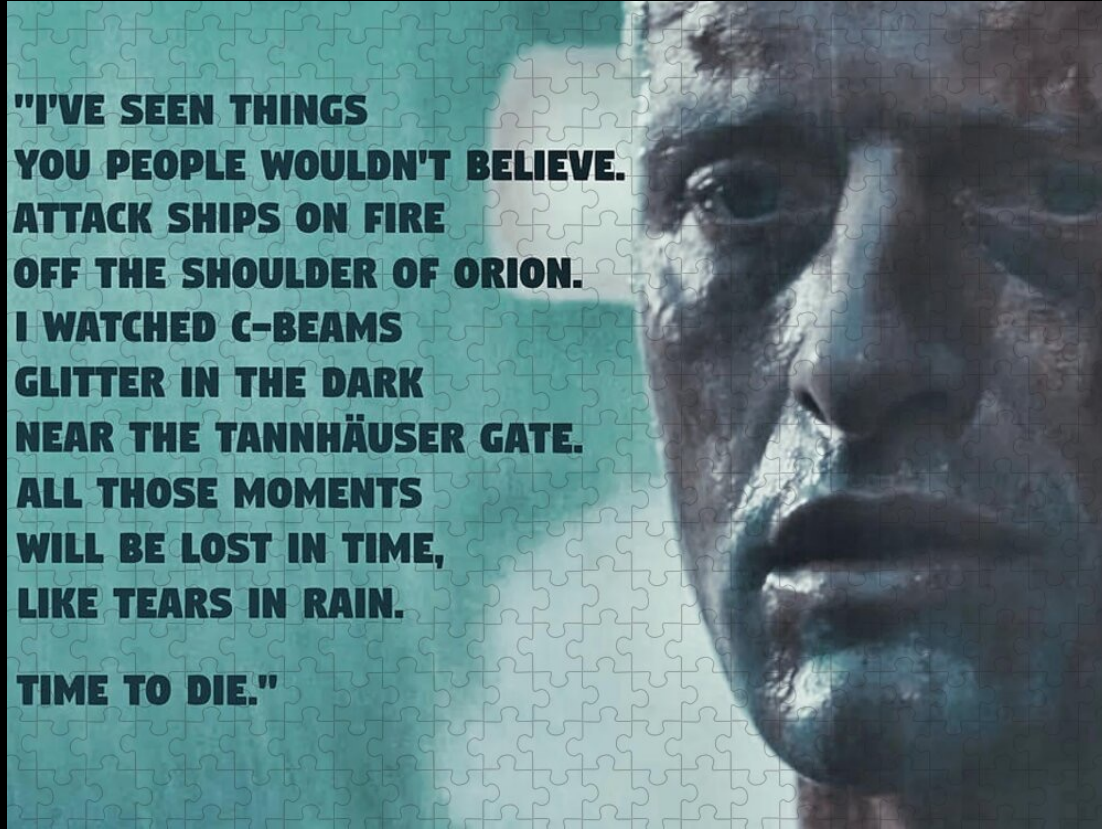
*Synergistic Multimodal Comprehension and Creation*

**Runpei Dong** Chunrui Han Yuang Peng Zekun Qi

Zheng Ge Jinrong Yang Liang Zhao Jianjian Sun Hongyu Zhou Haoran Wei Xiangwen Kong  
Xiangyu Zhang Kaisheng Ma Li Yi

International Conference on Learning Representations (ICLR)  
May 7<sup>th</sup>-11<sup>th</sup> 2024, Vienna Austria

# Do Androids Dream of Electric Sheep?



Blade Runner, 1982



*"What I cannot create, I do not understand"*  
Richard P. Feynman, 1988

**LLMs MIGHT DREAM, WHAT ABOUT MULTIMODAL LLMs?**

## Previous & Concurrent Works

**BUT NO SYNERGY BETWEEN COMPREHENSION AND CREATION**

$I_1, I_2$ : Raw Image

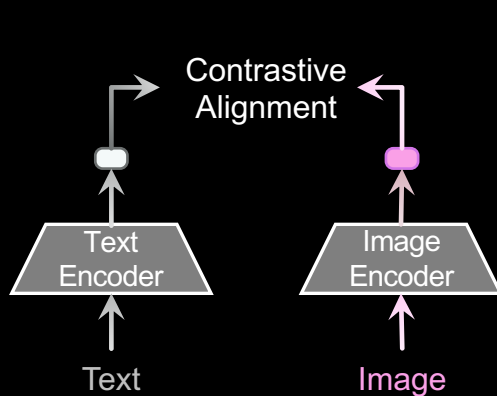
$T_1, T_2$ : Raw Text

Text  
Embedding

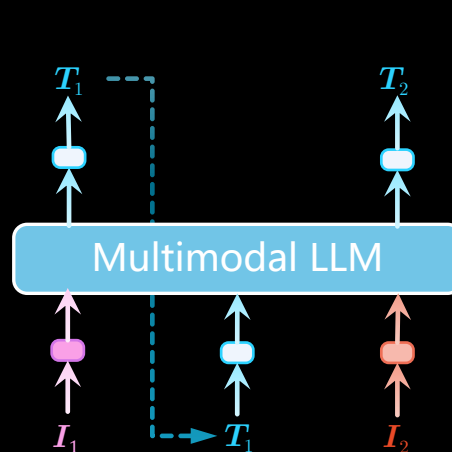
Visual  
Embedding

Inference

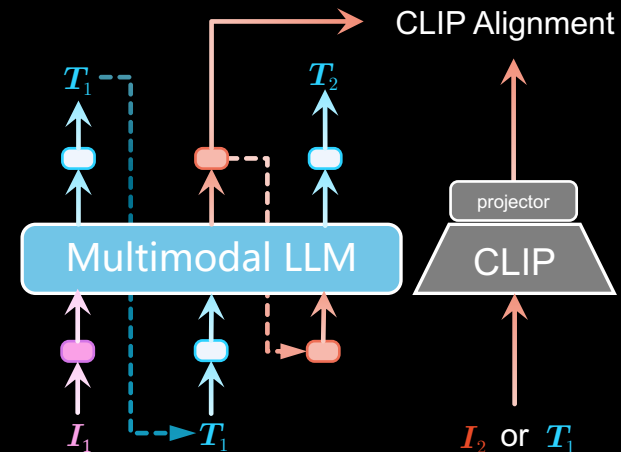
Stream



CLIP-like

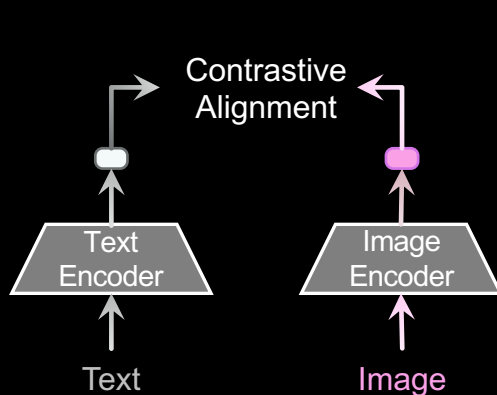
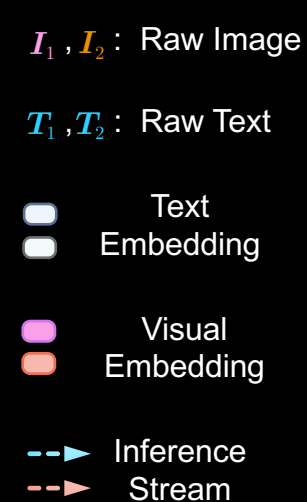


Flamingo/BLIP-like

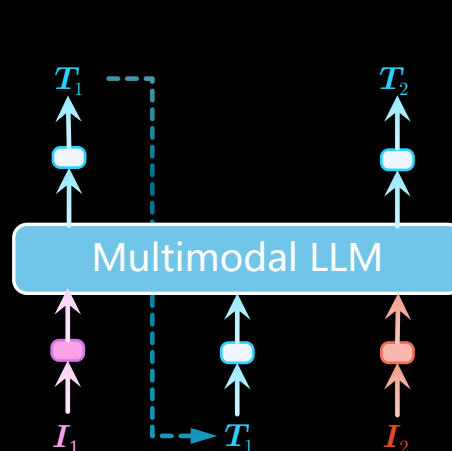


GILL/Emu-like

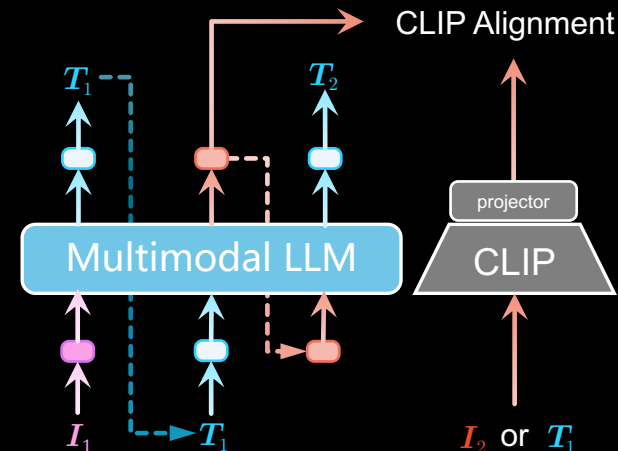
# Fully Autoregressive Multimodal Modeling



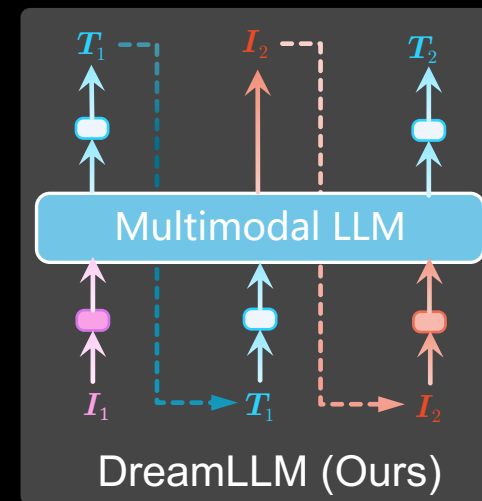
CLIP-like



Flamingo/BLIP-like



GILL/Emu-like



DreamLLM (Ours)

## DreamLLM Key Points

- Fully Autoregressive Multimodal Modeling with raw data multimodal input & output
- First work for instruction-following Interleaved content creation
- Learning Synergy between multimodal comprehension & content creation

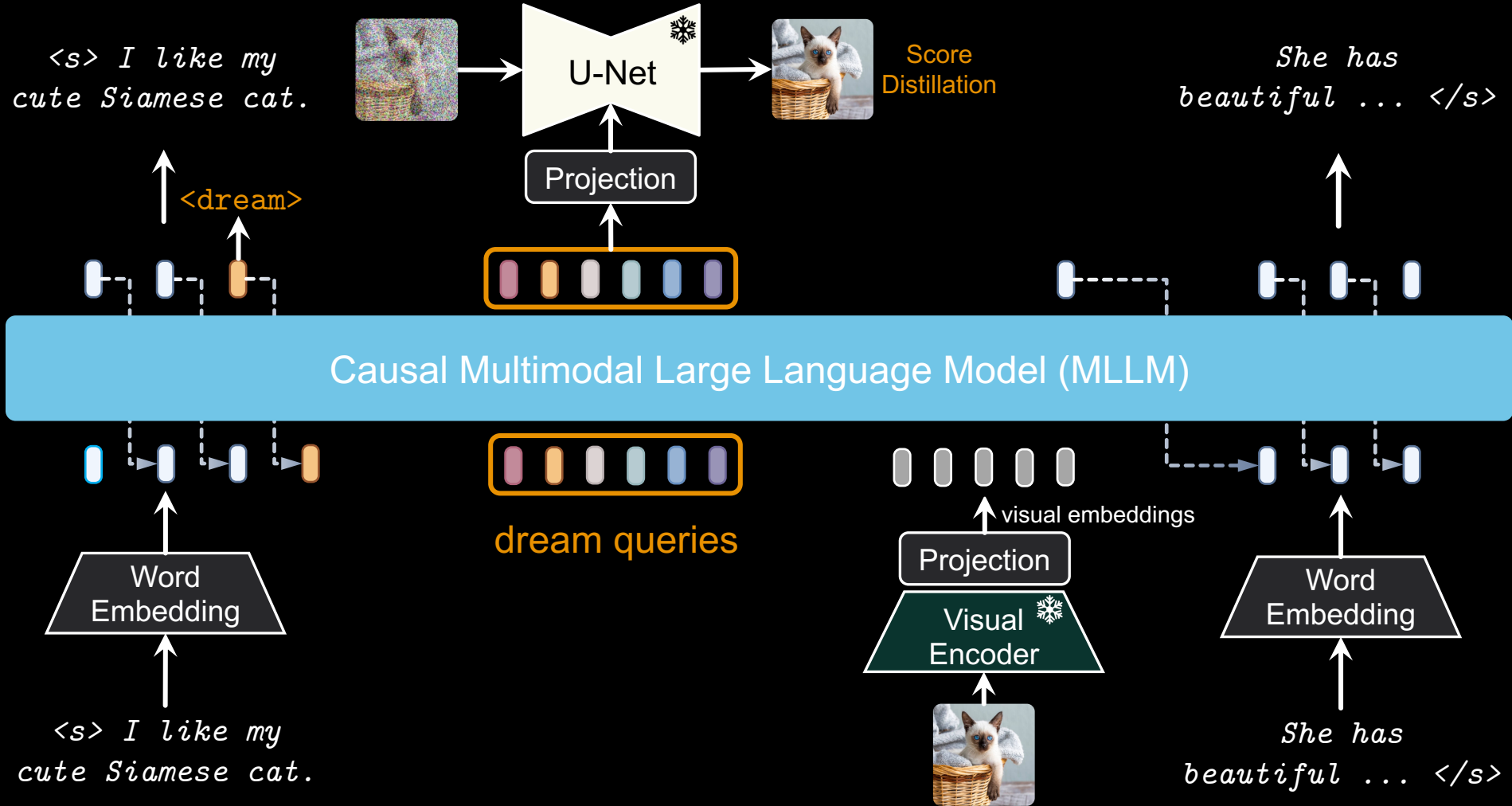
# DreamLLM

## Interleaved Documents

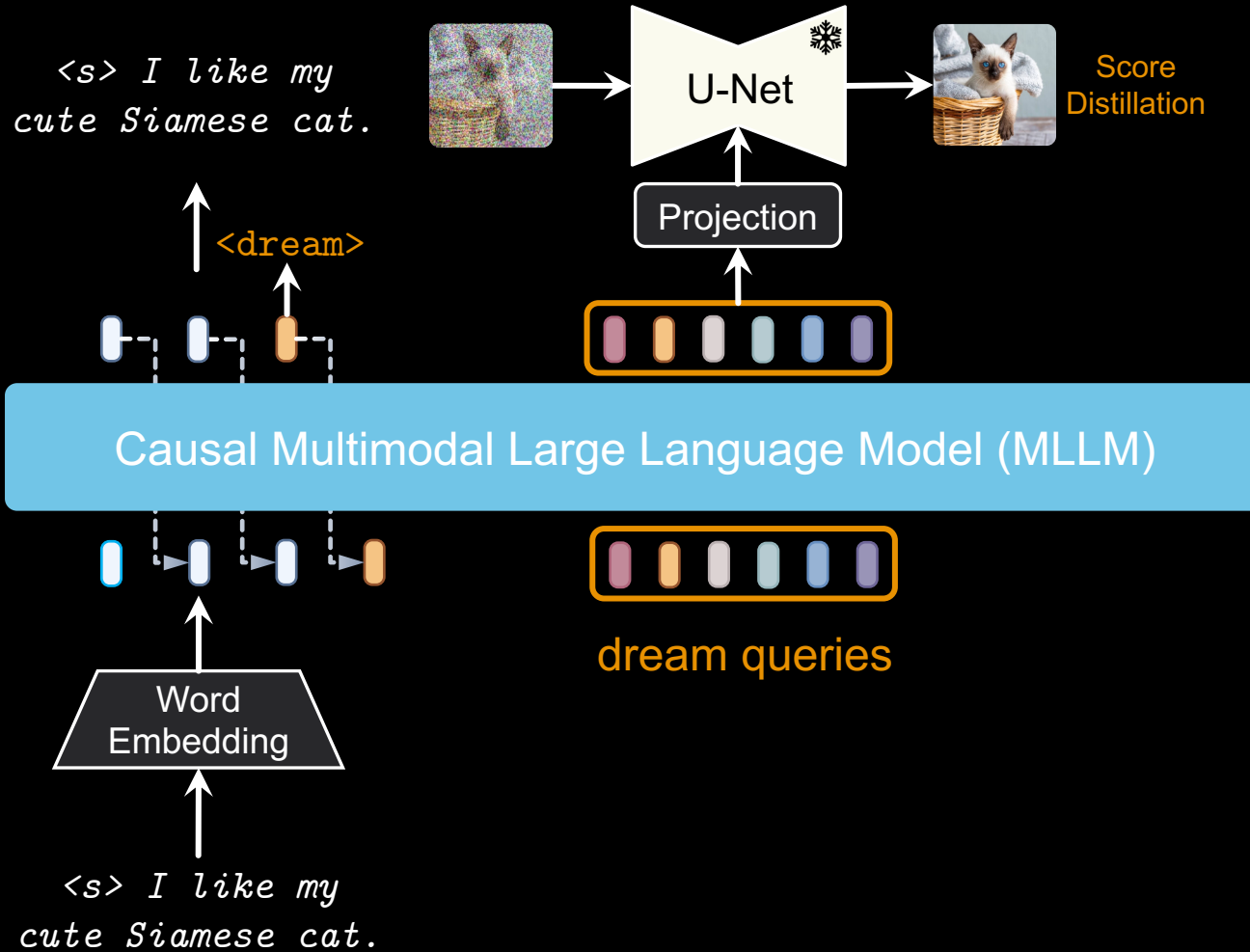
"I like my cute Siamese cat.",



"She has beautiful blue eyes, and she likes to lie on her cozy nest.", ...



# DreamLLM



## Learning Objective

Diffusion denoising instead of feature regression for direct distribution sampling in pixel space!

## SCORE DISTILLATION FOR GENURE DISTRIBUTION LEARNING

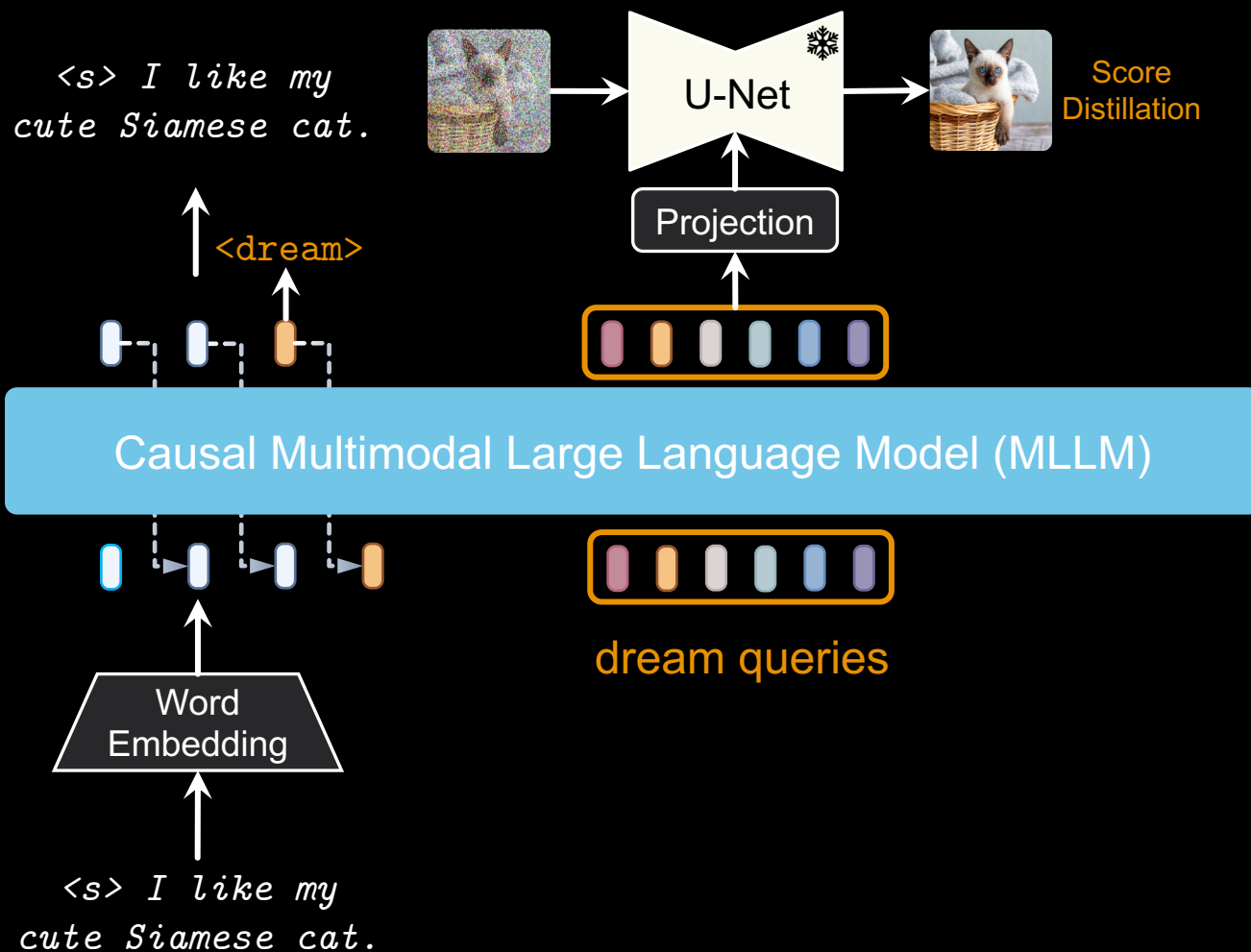
## Conditional Embedding

Query LLMs using dream queries instead of forcing LLMs to output visual representations!

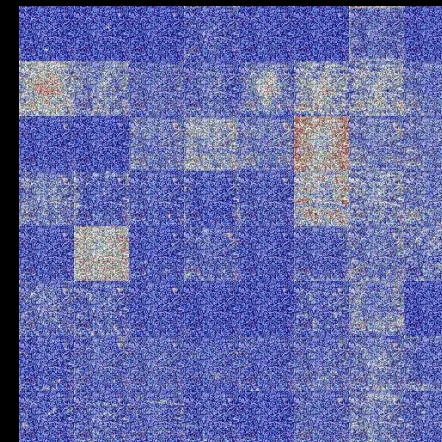
## CONTRASTIVE MODELS ONLY LEARN MODALITY-SHARED BUT NOT MODALITY-SPECIFIC INFORMATION



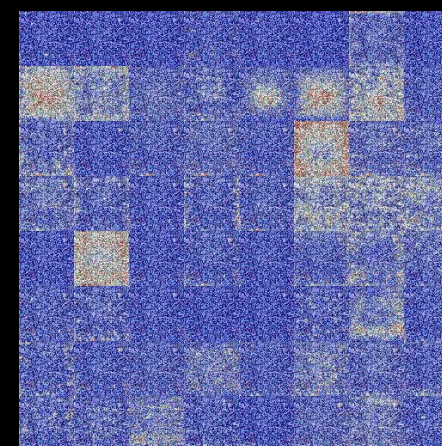
# DreamLLM



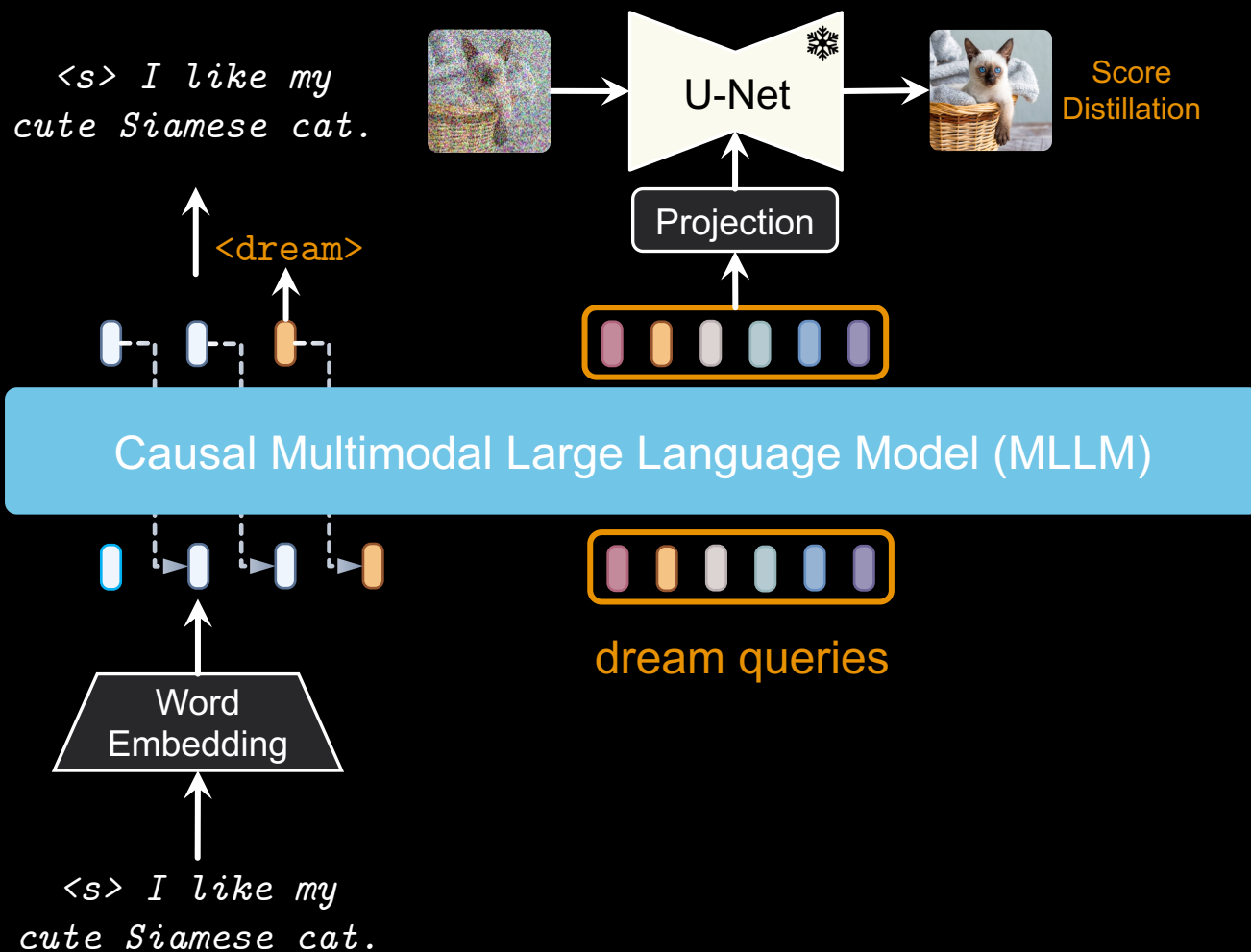
*A cat and a whisky.*



*A polar bear in the forest.*



# DreamLLM



*A cat and a whisky.*



*A polar bear in the forest.*



# Text-to-Image Generation



small bird with a pale yellow underside light brown crown and back gray tail and wing tips tip of tail feather bright yellow black eyes and black strip over eyes



panda mad scientist mixing sparkling chemicals, art station



an espresso machine that makes coffee X, art station, where  $X \in \{\text{"in a warm scene", "from human soul"}\}$



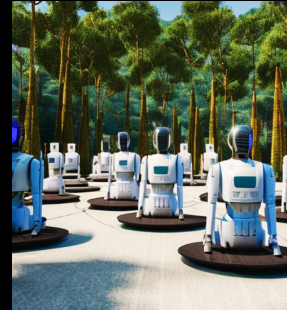
an armchair in the shape of an avocado



a futuristic city X, where  $X \in \{\text{"in a synthwave style", "in vaporwave style", "made of water", "Beijing opera style"}\}$



an astronaut riding a horse X, where  $X \in \{\text{"in a photorealistic style", "in the style of Pop Art", "as a charcoal sketch", "as a golden relief"}\}$



robots meditating in a vipassana retreat



# Multimodal Dialogue



Can you give a short introduction to this movie?



DreamLLM

The Godfather, also known as Il Padrino, is a classic American crime drama film directed by Francis Ford Coppola. It is the first installment in The Godfather trilogy, which also includes The Godfather Part II (1974) and The Godfather Part III (1990). The movie is widely regarded as a masterpiece and a milestone in the mafia genre, exploring the story of the Corleone family, a fictional Italian-American crime family in New York. It follows the transformation of Michael Corleone, the youngest son of the family's aging don, Vito Corleone, as he rises to power and eventually takes on the mantle of his father. The film features a strong cast, which includes Marlon Brando as Vito Corleone, Al Pacino as Michael Corleone, James Caan as Sonny Corleone, Robert Duvall as Tom Hagen, and Diane Keaton as Kay Corleone. The Godfather is known for its iconic scenes, memorable lines, and its impact on the Hollywood film industry.

Good. Show me an illustration of a panda wearing a suit and drinking the martini like the Godfather.



DreamLLM

An illustration of a panda wearing a suit and drinking a martini like the Godfather.



# In-Context Dialogue

Input Prompt

Completion



This is a Welsh Corgi.



This is a Koala.



This is



a woolly Mammoth.



What is the title of this painting? Answer: The Son of Man.



When was this masterpiece painted? Answer: 1818.



Who painted this? Answer:



Claude Monet.



Film name: Harry Potter and the Prisoner of Azkaban



Film name: Blade Runner 2049



Film name:



Joker

# Instruction-Following Interleaved Content Creation



*Help me write the introduction and illustrations about Vienna city, show several high-quality photographs about the beautiful scenes and buildings such as Vienna International Opera House.*



DreamLLM

*The Austrian capital, Vienna, has long been known as the City of Dreams, and there is no better place to experience than this romantic city during the holiday season. The city's historic center is surrounded by a collection of beautiful buildings on a square mile. Stroll along the famous Ring Boulevard and admire the architecture. Some of the buildings date back to the 13th century. The Vienna State Opera House is one of the most beautiful buildings. Founded in 1869, the opera house is the home of the Vienna State Opera and the annual Vienna Opera Ball.*



*Cosmopolitan flair, an eclectic mix of architectural styles and a rich cultural heritage come together in Vienna. The city is home to a number of museums, palaces and churches. The Sacher Torte and the Wiener Schnitzel are local specialties. A visit to the romantic Christmas Market in Vienna is also a unique experience.*





# Multimodal In-context Image Generation

Multimodal Input



*Delicious food.*  
*X.*

DreamLLM



*Teddy bear*

In-context Image Edition

Multimodal Input



*A black dog.*  
*A dog X.*

DreamLLM



*swims in water*



*on the beach*



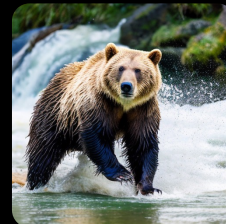
*in the snow*



*lies on sofa*

In-context Subject-Driven Generation

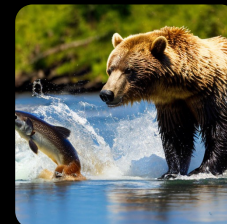
Multimodal Input



*A bear.*



*A salmon. X.*



*A grizzly bear catching a salmon in a crystal clear river surrounded by a forest*



*A ship on the ocean. X.*



*A pod of dolphins leaping out of the water in an ocean, with a ship on the background.*

SD



In-context Compositional Generation



# Image-to-Image Generation

Input Image



*dog*



*swimming in the pool*



*Wearing sunglasses*



*wearing a top hat*



*in the jungle*

Input Image



*teapot*



*on the beach*



*in blue*



*as a lamp*



*in the jungle*

Input Image



*car*



*floating in the water*



*painted in green*



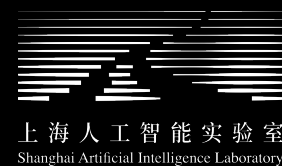
*runs on the mountain*



*in front of wheat field*



MEGVII 旷视



 **Spotlight Paper**

# DreamLLM ICLR

*Synergistic Multimodal Comprehension and Creation*



GitHub

**Runpei Dong** Chunrui Han Yuang Peng Zekun Qi

Zheng Ge Jinrong Yang Liang Zhao Jianjian Sun Hongyu Zhou Haoran Wei Xiangwen Kong  
Xiangyu Zhang Kaisheng Ma Li Yi

## Thanks!

*Please stop by **Halle B** for more details  
or contact me through email.*

[runpei.dong@gmail.com](mailto:runpei.dong@gmail.com)