Multimodal Generative LLMs: Unification, Interpretability, Evaluation

Mohit Bansal

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Talk Outline

A journey of multimodal generative LLMs for enhancing their unification, interpretable planning/programming, evaluation:

- **Unified/Universal Multimodal Learning** (for Generalizability, Shared Knowledge, Efficiency)
  - VLT5: Unifying Vision-and-Language Tasks via Text Generation [ICML 2021]
  - TVLT: Textless Vision-Language Transformer [NeurIPS 2022]
  - UDOP: Unifying Vision, Text, and Layout for Universal Document Processing [CVPR 2023]
  - CoDi: Any-to-Any Generation via Composable Diffusion [NeurIPS 2023] & *CoDi-2: In-Context, Interleaved, and Interactive Any-to-Any Generation [2023]*
- **Interpretable Multimodal Generation via LLM Planning/Programming** (for Understanding, Control, Faithfulness)
  - VPGen: Step-by-Step Text-to-Image Generation with Interpretable Visual Programming [NeurIPS 2023]
  - VideoDirectorGPT: Consistent Multi-Scene Video Generation via LLM-Guided Planning [2023]
  - DiagrammerGPT: Generating Open-Domain, Open-Platform Diagrams via LLM Planning [2023]
- **Evaluation of Multimodal Generation Models** (of Fine-grained Skills, Faithfulness, Social Biases)
  - DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models [ICCV 2023]
  - VPEval: Step-by-Step Text-to-Image Evaluation with Interpretable Visual Programming [NeurIPS 2023]
  - *Davidsonian Scene Graph: Improving Reliability in Fine-grained Evaluation for Text-to-Image Generation [2023]*
- **Next Big Challenges**: trade-offs, structure, non-verbal, interaction, reasoning, causality, long-distance fine-grained evaluation, efficiencies

Talk Outline

A journey of multimodal generative LLMs for enhancing their unification, interpretable planning/programming, evaluation:

- **Unified/Universal Multimodal Learning** (for Generalizability, Shared Knowledge, Efficiency)
  - VLT5: Unifying Vision-and-Language Tasks via Text Generation [ICML 2021]
  - TVLT: Textless Vision-Language Transformer [NeurIPS 2022]
  - UDOP: Unifying Vision, Text, and Layout for Universal Document Processing [CVPR 2023]
  - CoDi: Any-to-Any Generation via Composable Diffusion [NeurIPS 2023] & *CoDi-2: In-Context, Interleaved, and Interactive Any-to-Any Generation [2023]*
- **Interpretable Multimodal Generation via LLM Planning/Programming** (for Understanding, Control, Faithfulness)
  - VPGen: Step-by-Step Text-to-Image Generation with Interpretable Visual Programming [NeurIPS 2023]
  - VideoDirectorGPT: Consistent Multi-Scene Video Generation via LLM-Guided Planning [2023]
  - DiagrammerGPT: Generating Open-Domain, Open-Platform Diagrams via LLM Planning [2023]
- **Evaluation of Multimodal Generation Models** (of Fine-grained Skills, Faithfulness, Social Biases)
  - DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models [ICCV 2023]
  - VPEval: Step-by-Step Text-to-Image Evaluation with Interpretable Visual Programming [NeurIPS 2023]
  - *Davidsonian Scene Graph: Improving Reliability in Fine-grained Evaluation for Text-to-Image Generation [2023]*
- **Next Big Challenges**: trade-offs, structure, non-verbal, interaction, reasoning, causality, long-distance fine-grained evaluation, efficiencies

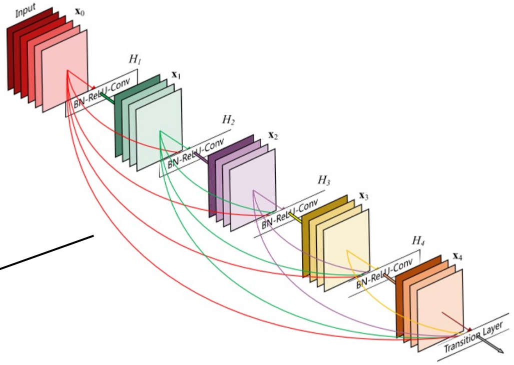
Vision: Pre-training → Fine-tuning

Motivation: the amount of data is limited in downstream tasks
and pre-training enables much more data.

Visual
Pre-training:

ImageNet
[Deng, CVPR 2009]
1.3M Images, 1000 Labels

→ Image Classification →



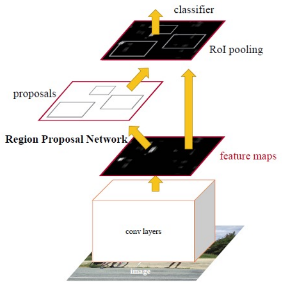
DenseNet
[Huang, CVPR 2017]

Visual
Fine-tuning:

COCO
Common Objects in Context

MS COCO
[Lin, ECCV 2009]
120K Images, 80 Labels

→ Object Detection →



Faster RCNN
[Ren, NeurIPS 2015]

Language: Pre-training → Fine-tuning

Motivation: the amount of data is limited in downstream tasks and pre-training enables much more data.

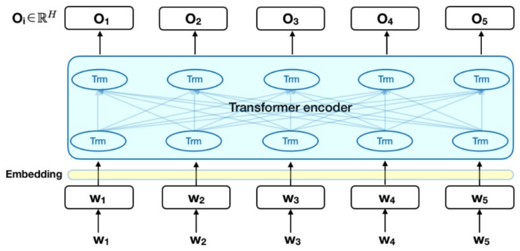
Language Pre-training:

Text in Wikipedia

~2500M Tokens (i.e., Words)

Language Model

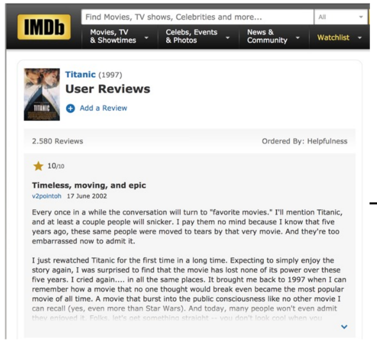
[Peters et al., NAACL 2018], [Devlin et al., NAACL 2019]



Transformer

[Vaswani, NeurIPS 2017]

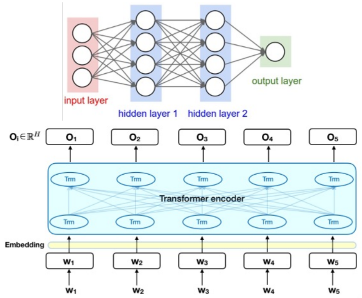
Language Fine-tuning



Movie Review [Maas et al., ACL 2011]

~2.5M Tokens (i.e., Words)

Sentiment Analysis



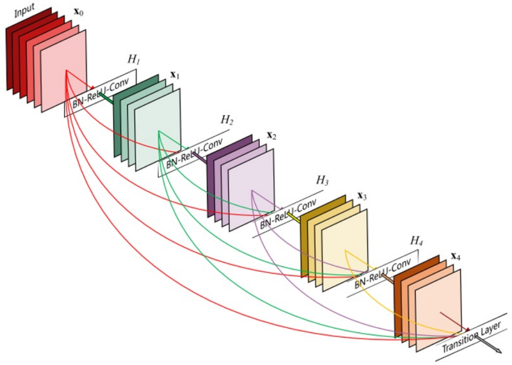
Transformer + Linear Layers

Pre-training of Single Modality Tasks

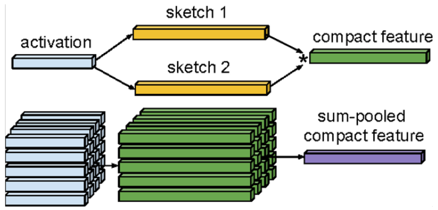
Limitation: Single-modality pre-trained models are not aware of the interactions between vision and language

Visual Pre-training:

Image Classification

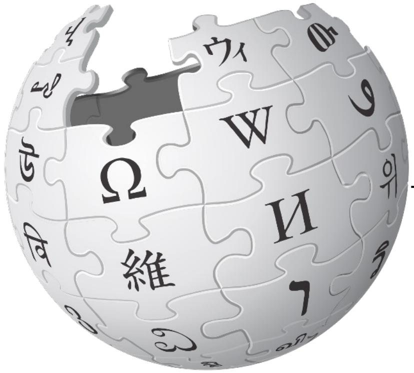


Multimodal Fusion Layers

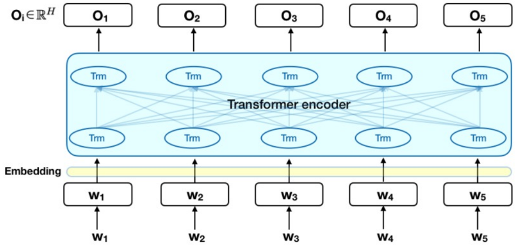


Visual Question Answering, Navigation, Grounding, ...

Language Pre-training:

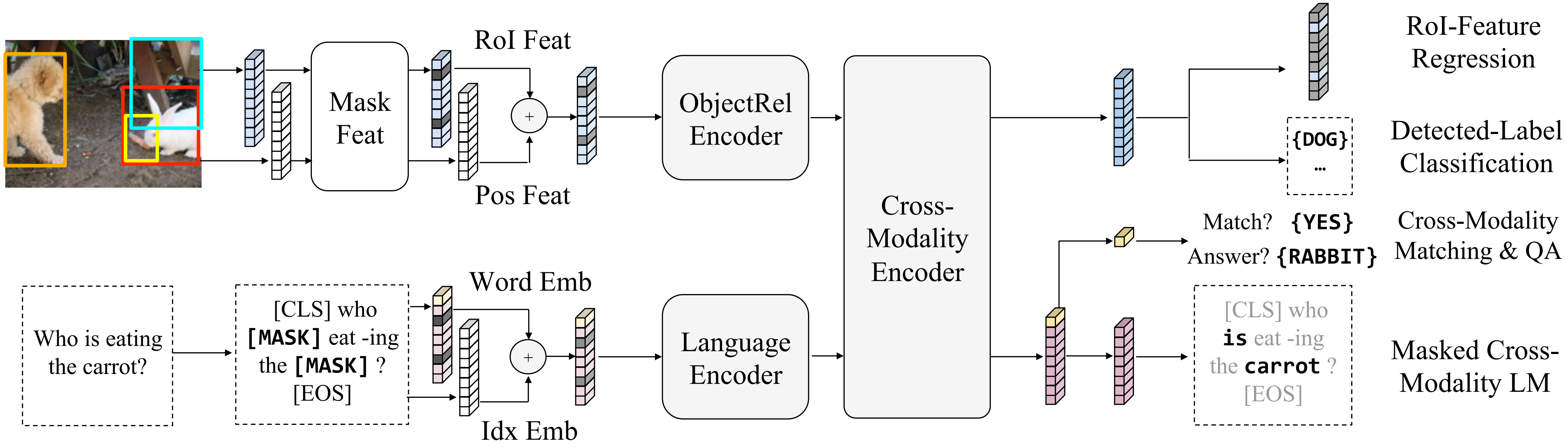


Language Model



Large-Scale Cross-Modal Pre-training: LXMERT

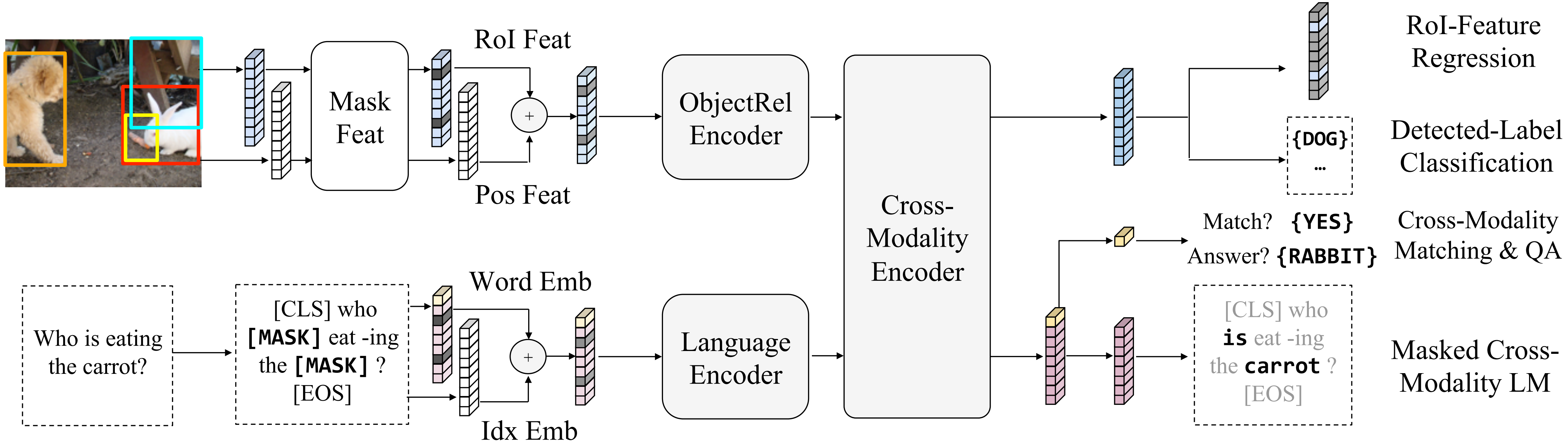
- LXMERT combines knowledge from text, vision and cross-modal matching: vision-language transformers with 3 encoders (object relations, language, cross-modal) & 5 pretraining tasks: masked-LM, masked-Object-Prediction (feature regression+label classification), cross-modality matching, image-QA.



[Tan and Bansal, EMNLP 2019]

Large-Scale Cross-Modal Pre-training: LXMERT

- LXMERT combines knowledge from text, vision and cross-modal matching: vision-language transformers with 3 encoders (object relations, language, cross-modal) & 5 pretraining tasks: masked-LM, masked-Object-Prediction (feature regression+label classification), cross-modality matching, image-QA.

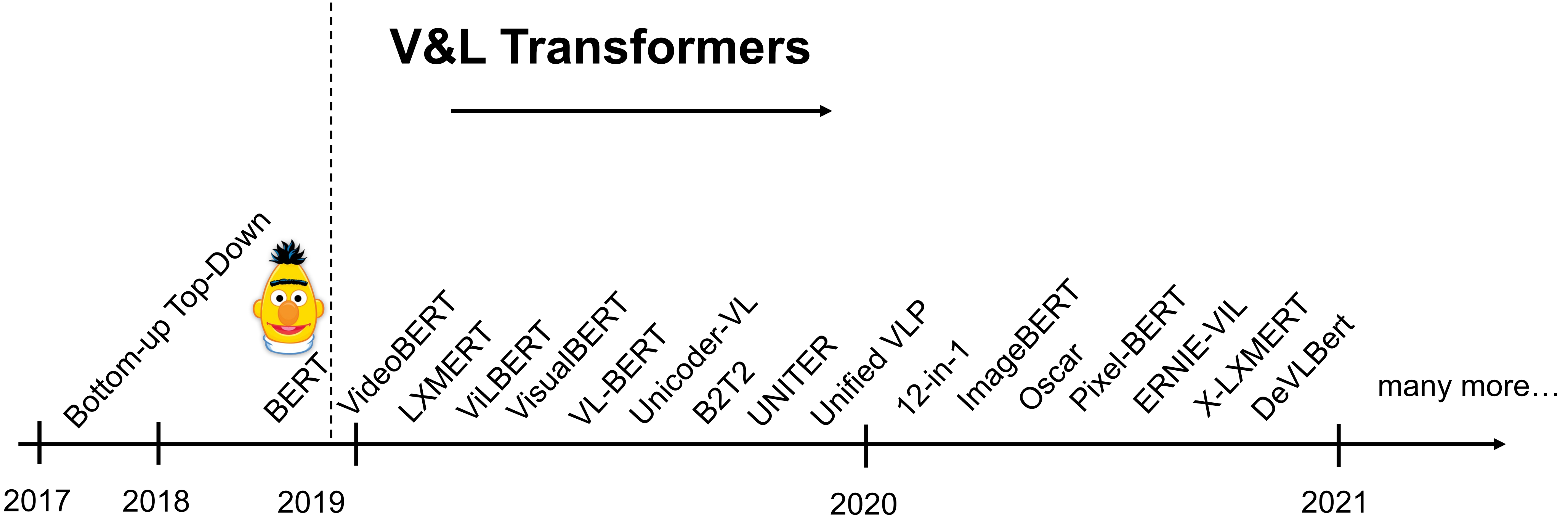


- Achieved big gains + sota on several VL tasks such as VQA, GQA, NLVR2, VizWiz, etc.

[Tan and Bansal, EMNLP 2019]

Tons of Specialized Vision-and-Language Pretraining Models

- Many different architectures (single vs. multi-stream), attention methods, objective functions, encoder/decoders, output heads, specialized modules (OCR/ASR/Tokenizers), etc., etc.!

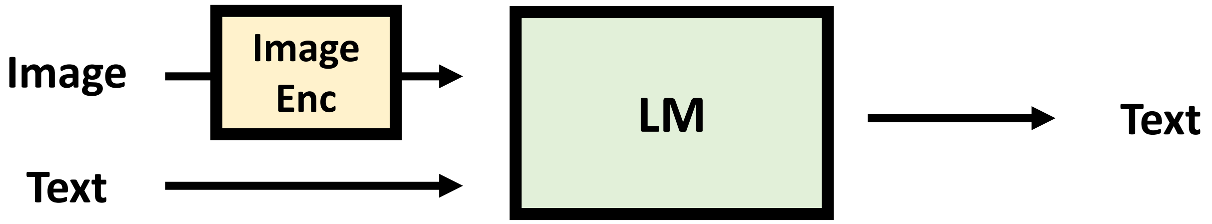


[Sun et al., 2019; Tan and Bansal, 2019; Lu et al., 2019; Li et al., 2019; Su et al., 2019; Chen et al., 2020; Zhou et al., 2020; Li et al., 2020; inter alia]

Part 1: Unified/Universal Multimodal Learning

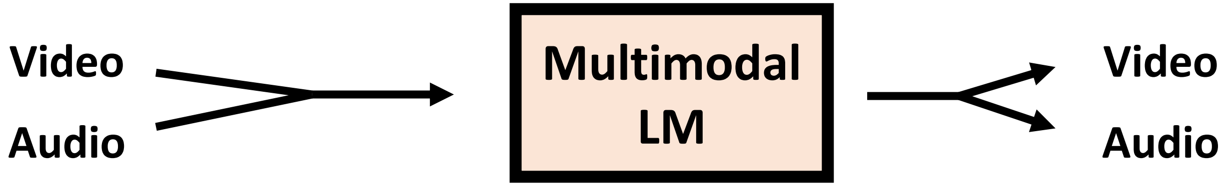
VL-T5 (ICML 2021)

all multimodal tasks via text generation



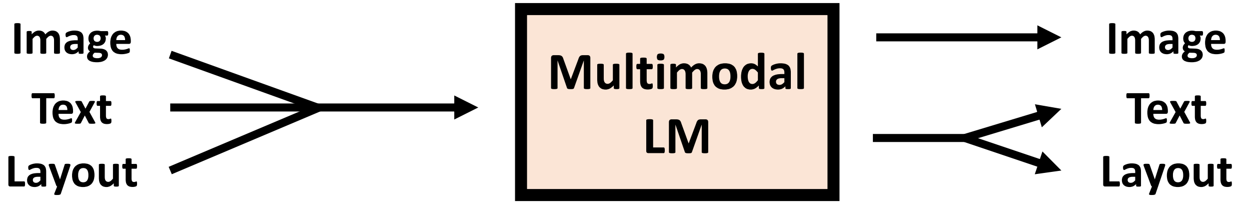
TVLT (NeurIPS 2022)

video modeling without text (audio as images)



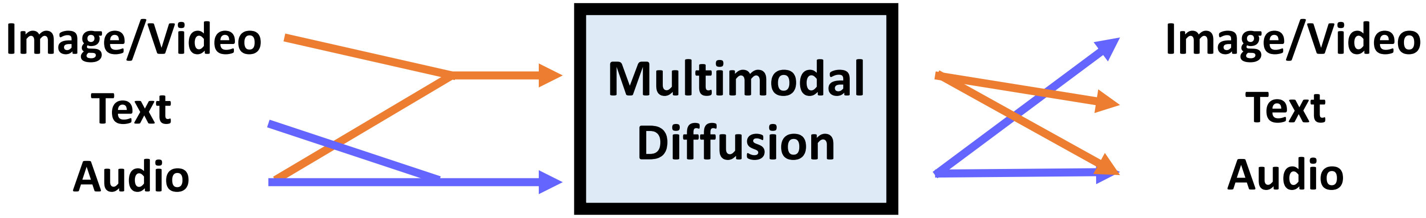
UDOP (CVPR 2023)

document image/text/layout with single architecture



CoDi (NeurIPS 2023)

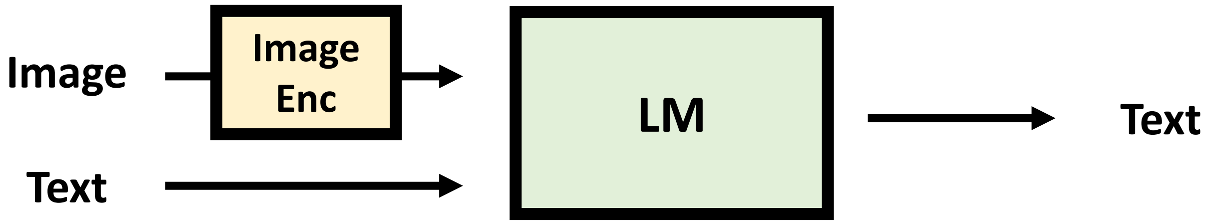
generating any-to-any input-output modality combination



Part 1: Unified/Universal Multimodal Learning

VL-T5 (ICML 2021)

all multimodal tasks via text generation



TVLT (NeurIPS 2022)

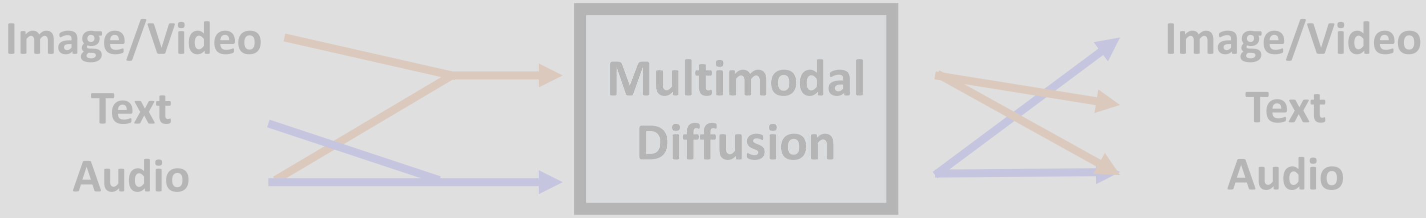
video modeling without text (audio as images)

UDOP (CVPR 2023)

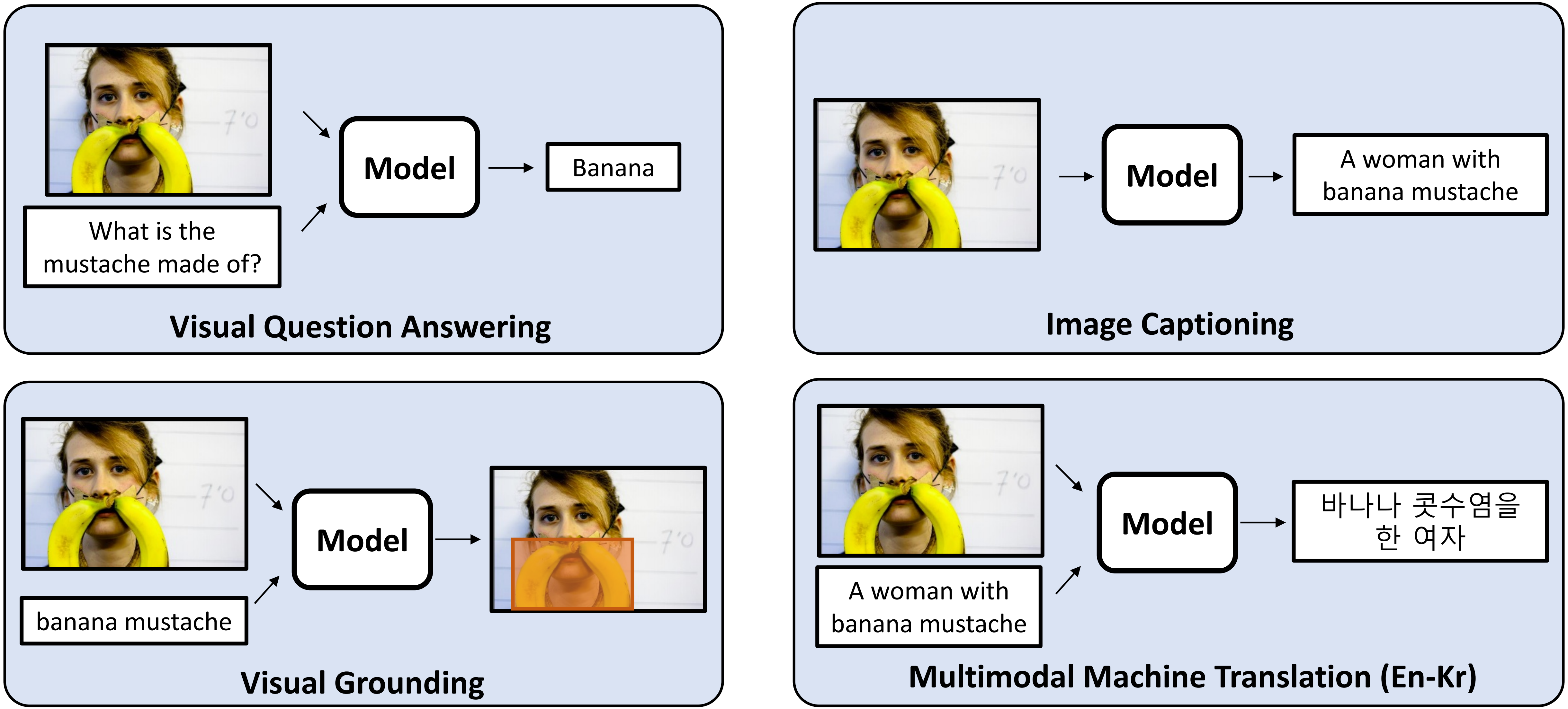
document image/text/layout with single architecture

CoDi (NeurIPS 2023)

generating any-to-any input-output modality combination

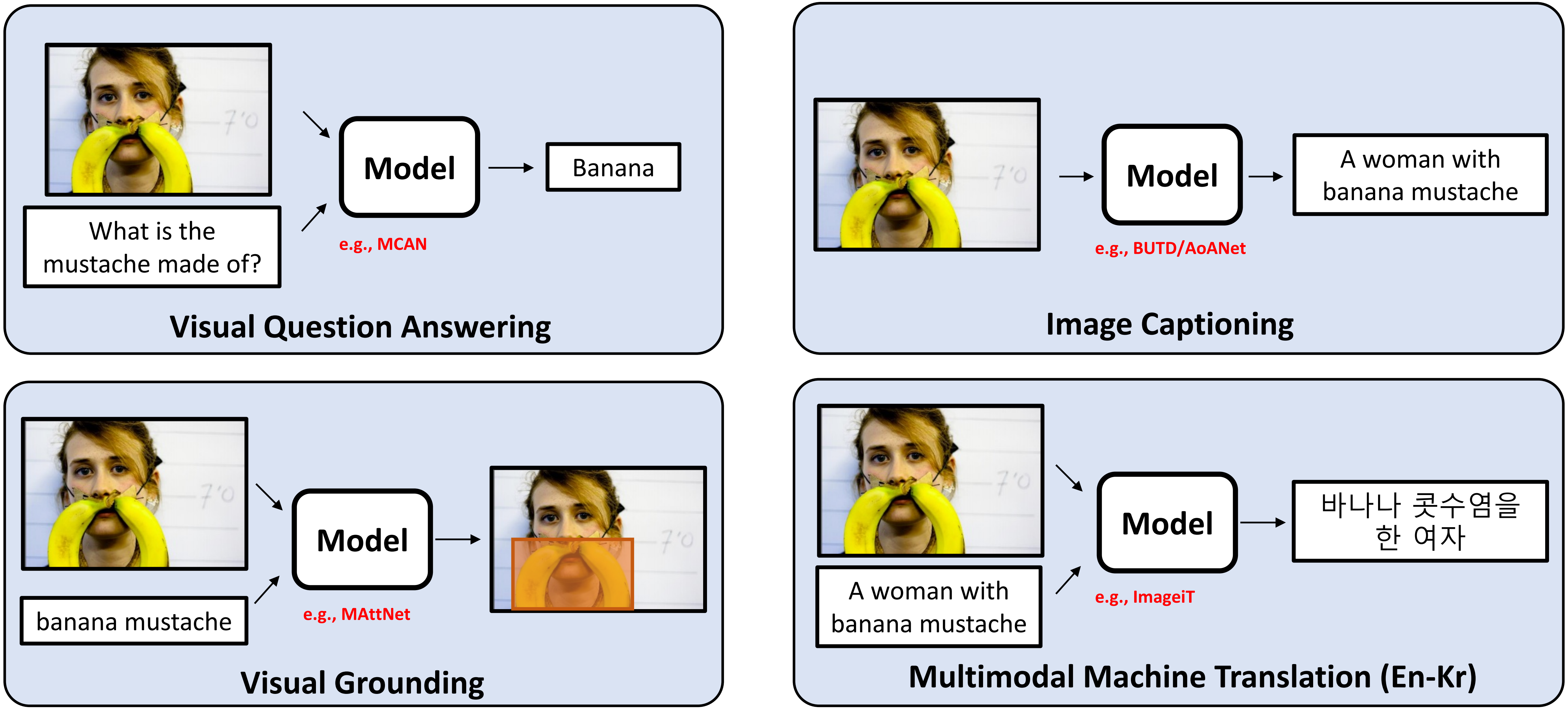


Diverse Vision-and-Language Tasks (and Specialized Models)



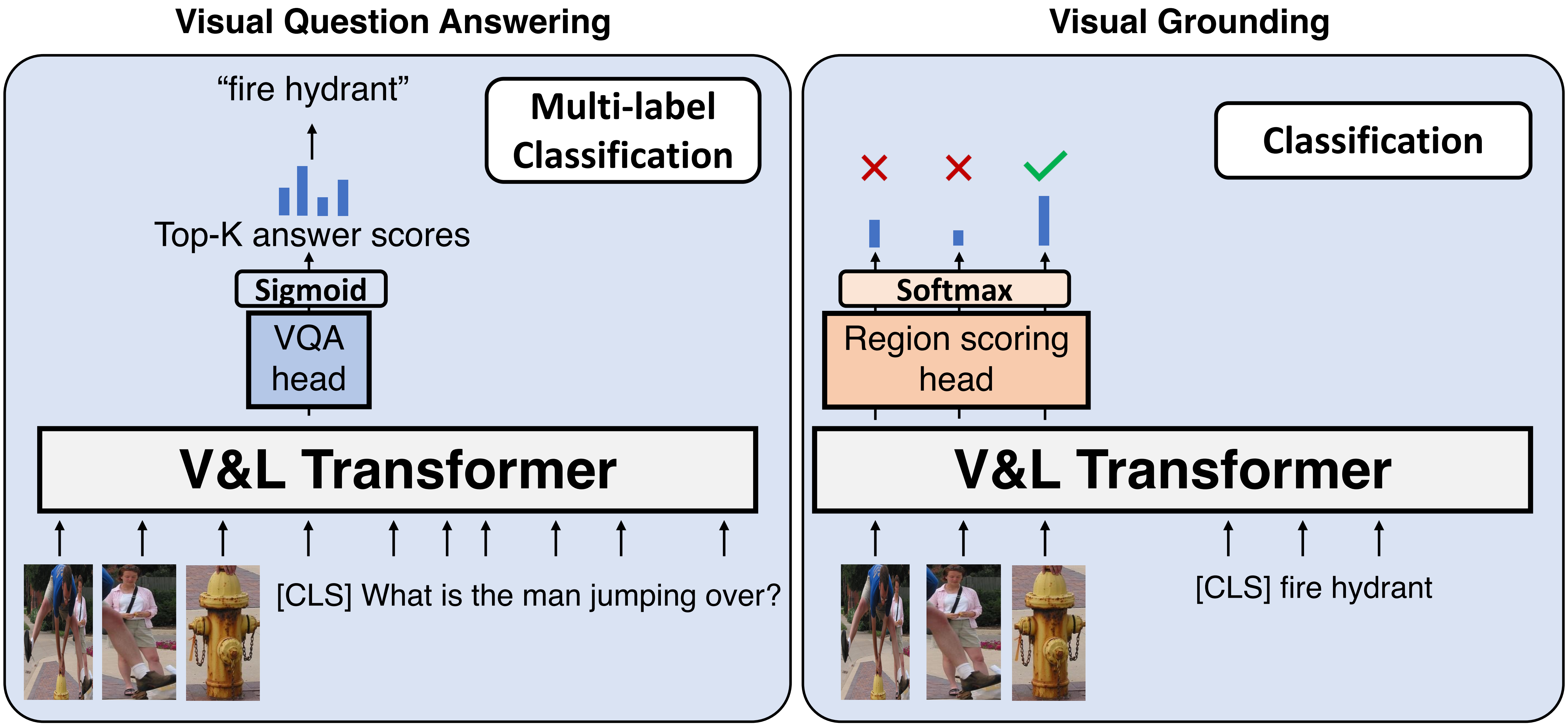
Anderson et al., 2018, Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering
Yu et al., 2019, Deep Modular Co-Attention Networks for Visual Question Answering
Huang et al., 2019, Attention on Attention for Image Captioning
Yu et al., 2018, MAttNet: Modular Attention Network for Referring Expression Comprehension
Long et al., 2021, Generative Imagination Elevates Machine Translation

Diverse Vision-and-Language Tasks (and Specialized Models)



Anderson et al., 2018, Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering
Yu et al., 2019, Deep Modular Co-Attention Networks for Visual Question Answering
Huang et al., 2019, Attention on Attention for Image Captioning
Yu et al., 2018, MAttNet: Modular Attention Network for Referring Expression Comprehension
Long et al., 2021, Generative Imagination Elevates Machine Translation

Task-specific Architectures / Objectives / Modules

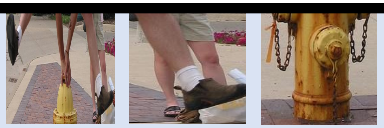


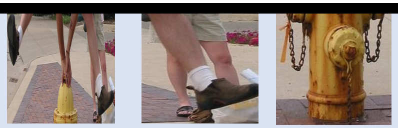
Task-specific Architectures / Objectives

Visual Question Answering

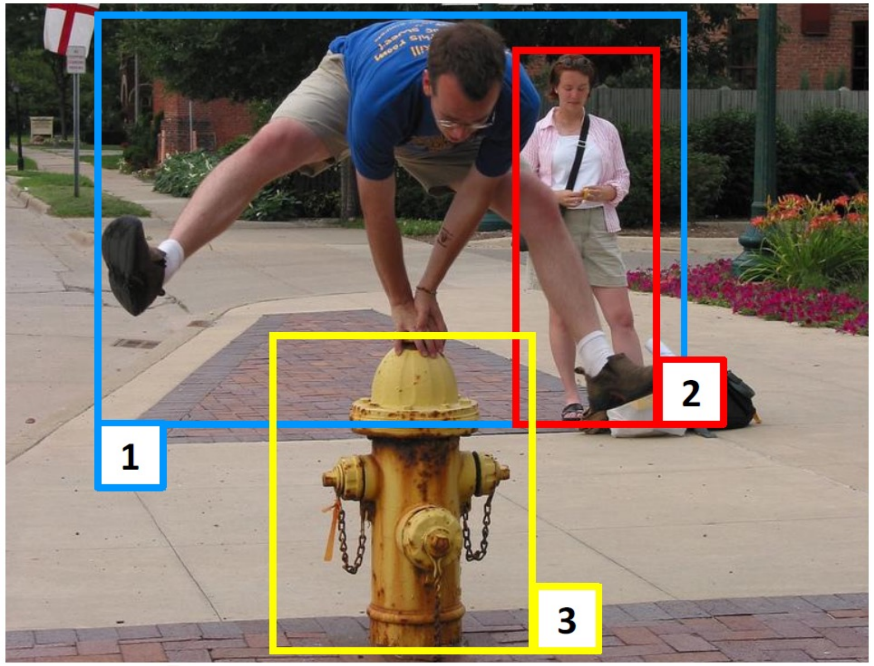
Visual Grounding

Can we tackle all V&L tasks with a single objective?





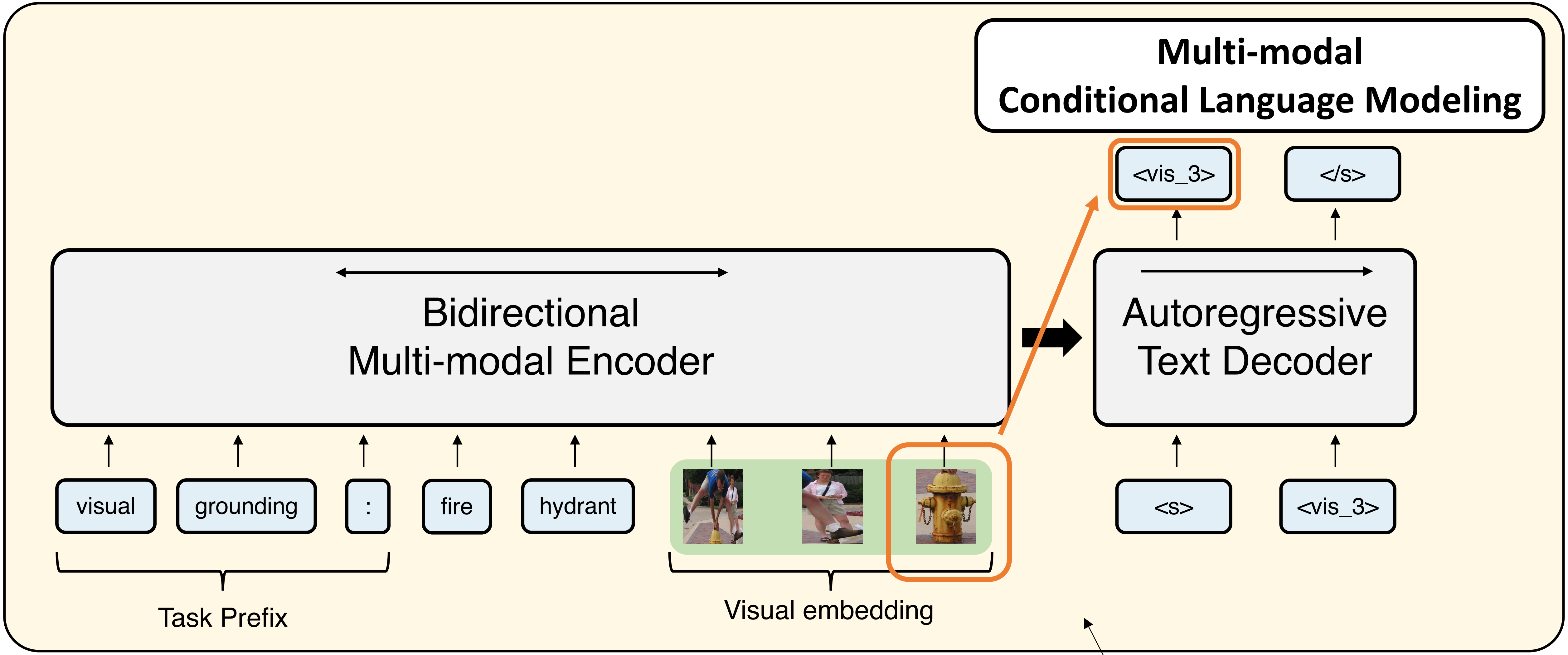
VL-T5: Many Multimodal Tasks as Text Generation



	Text Input		Text Output
Multimodal LM	"span prediction: A <text_1> is <text_2> over fire hydrant"	→	"<text_1> man <text_2> jumping"
Visual QA	"vqa: what is the man jumping over?"	→	"fire hydrant"
Visual Grounding	"visual grounding: yellow fire hydrant"	→	"<vis_3>"
Image-Text Matching	"image text match: A cat is lying on a bed"	→	"false"

[Cho et al., ICML 2021]

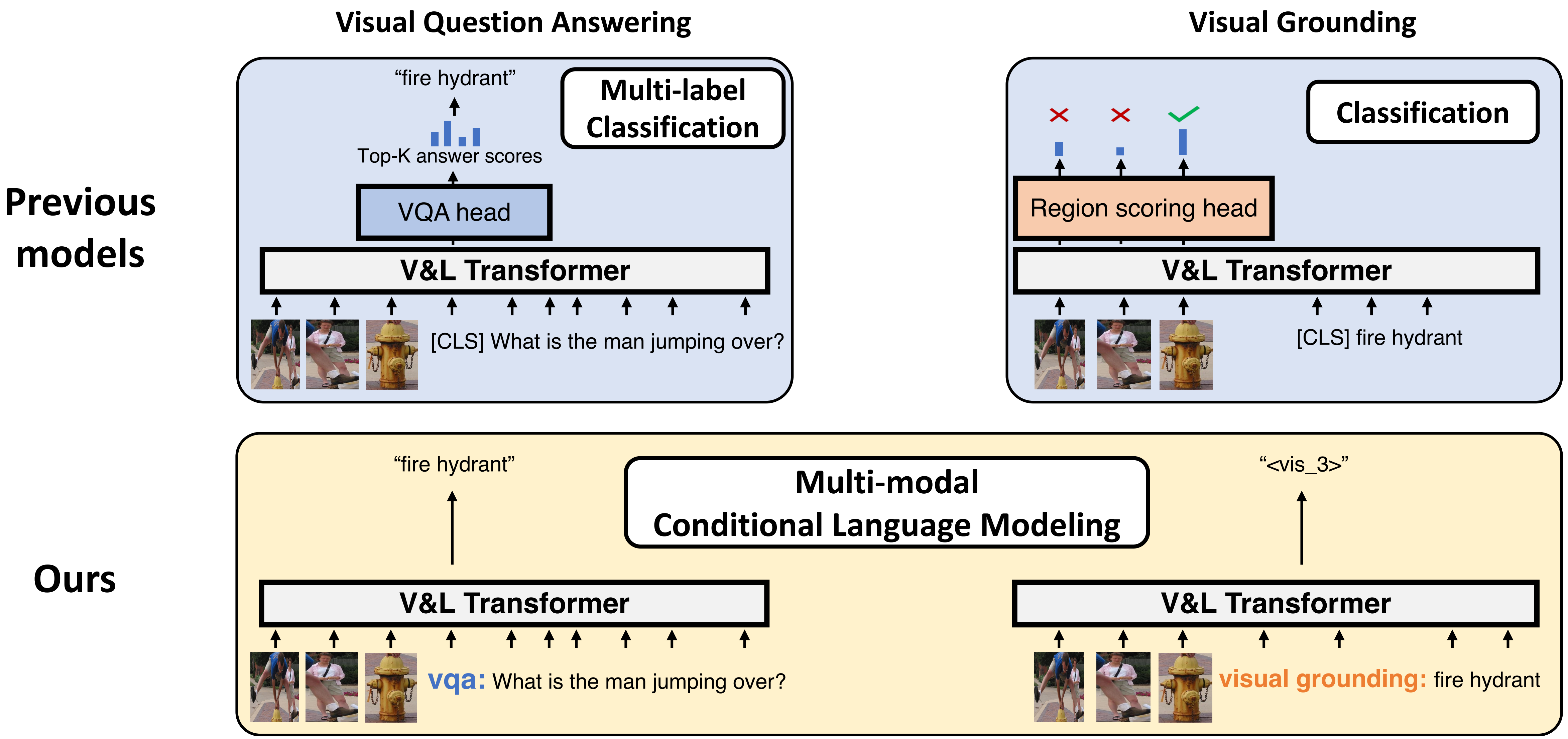
VL-T5: Many Multimodal Tasks as Text Generation



Weights are initialized from off-the-shelf Seq2Seq LMs (e.g., T5)

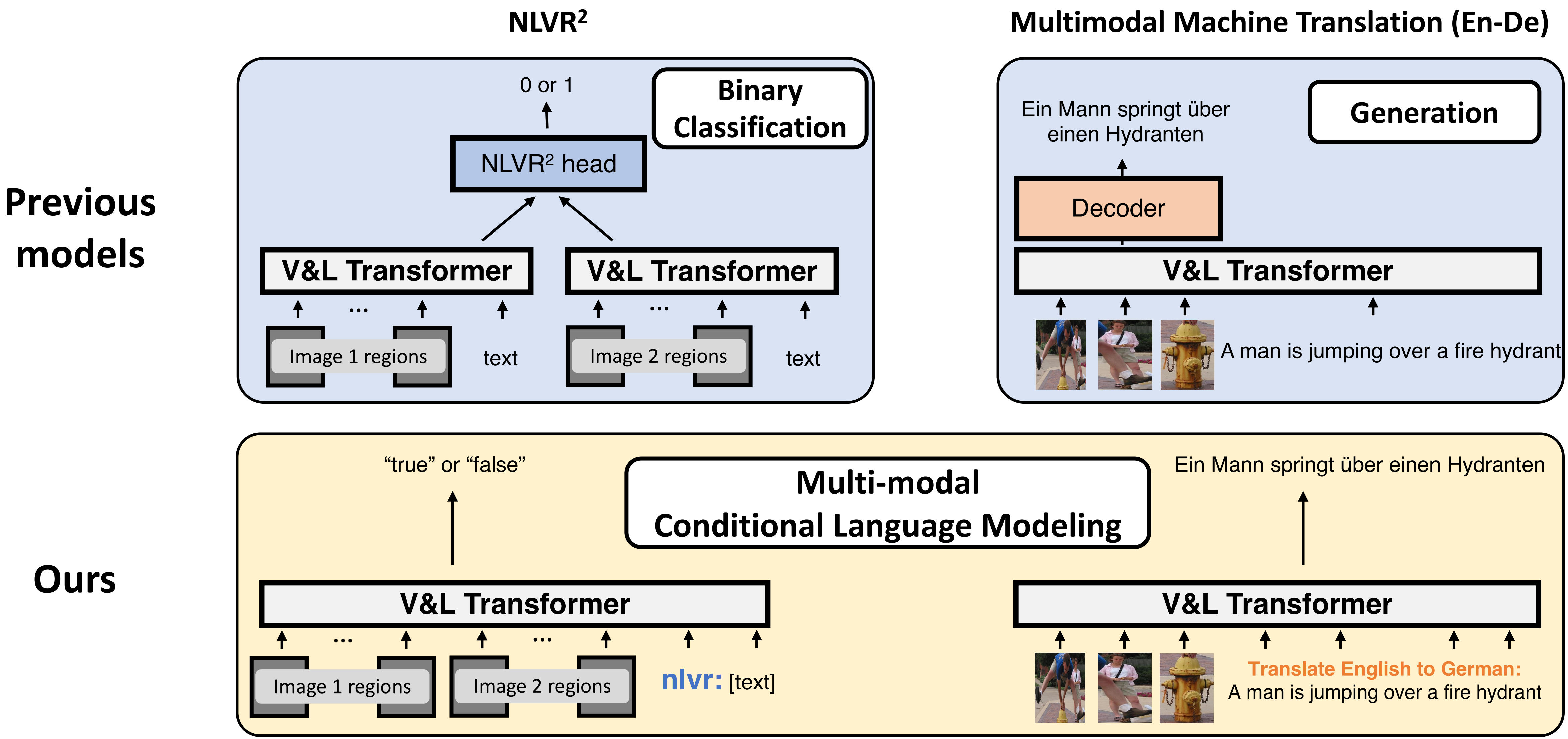
[Cho et al., ICML 2021]

VL-T5: Many Multimodal Tasks as Text Generation



[Cho et al., ICML 2021]

VL-T5: Many Multimodal Tasks as Text Generation



[Cho et al., ICML 2021]

Unified Architecture Comparable to Specialized Models

Method	# Pretrain Images	Discriminative tasks: VQA test-std Acc	GQA test-std Acc	NLVR^2 test-P Acc	RefCOCOg test^d Acc	VCR Q$\rightarrow$ AR test Acc	Generative tasks: COCO Cap Karpathy test CIDEr	Multi30K En-De test 2018 BLEU
LXMERT	180K	72.5	60.3	74.5	-	-	-	-
ViLBERT	3M	70.9	-	-	-	54.8	-	-
UNITER$_{Base}$	4M	72.9	-	77.9	74.5	58.2	-	
Unified VLP	3M	70.7	-	-	-	-	117.7	-
Oscar$_{Base}$	4M	73.4	61.6	78.4	-	-	123.7	-
XGPT	3M	-	-	-	-	-	120.1	-
MeMAD	-	-	-	-	-	-	-	38.5
VL-T5	180K	70.3	60.8	73.6	71.3	58.9	116.5	38.6
VL-BART	180K	71.3	60.5	70.3	22.4*	48.9	116.6	28.1

[Cho et al., ICML 2021]

Multi-task Learning with Single Shared Set of Parameters

Method	Finetuning tasks	# Params	Discriminative tasks					Generative tasks	
			VQA Karpathy test Acc	GQA test-dev Acc	$NLVR^2$ test-P Acc	RefCOCOg $test^d$ Acc	VCR val Acc	COCO Caption Karpathy test CIDEr	Multi30K En-De test2018 BLEU
VL-T5	single task	7P	67.9	60.0	73.6	71.3	57.5	116.1	38.6
VL-T5	all tasks	P	67.2	58.9	71.6	69.4	55.3	110.8	37.6

Similar performance with $1/7^{th}$ = 14% parameters!

Multi-task Learning with Single Shared Set of Parameters

Method	Finetuning tasks	# Params	Discriminative tasks					Generative tasks	
			VQA Karpathy test Acc	GQA test-dev Acc	$NLVR^2$ test-P Acc	RefCOCOg $test^d$ Acc	VCR val Acc	COCO Caption Karpathy test CIDEr	Multi30K En-De test2018 BLEU
VL-T5	single task	7P	67.9	60.0	73.6	71.3	57.5	116.1	38.6
VL-T5	all tasks	P	67.2	58.9	71.6	69.4	55.3	110.8	37.6

Similar performance with $1/7^{th}$ = 14% parameters!

- Also performs better on rare/unseen categories!

Multi-task Learning with Single Shared Set of Parameters

Method	Finetuning tasks	# Params	Discriminative tasks					Generative tasks	
			VQA Karpathy test Acc	GQA test-dev Acc	NLVR^2 test-P Acc	RefCOCOg test^d Acc	VCR val Acc	COCO Caption Karpathy test CIDEr	Multi30K En-De test2018 BLEU
VL-T5	single task	7P	67.9	60.0	73.6	71.3	57.5	116.1	38.6
VL-T5	all tasks	P	67.2	58.9	71.6	69.4	55.3	110.8	37.6

Similar performance with 1/7th = 14% parameters!

- Also performs better on rare/unseen categories!
- Many follow-up useful works on unification:

  e.g., SimVLM, Flamingo, OFA, UnifiedIO, BLIP-2, CoCa, PaLI, etc.

Wang et al., 2021, SimVLM: Simple Visual Language Model Pretraining with Weak Supervision
Alayrac et al., 2022, Flamingo: a Visual Language Model for Few-Shot Learning
Wang et al., 2022, OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework
Lu et al., 2022, Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks
Li et al., 2023, BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models
Yu et al., 2022, CoCa: Contrastive Captioners are Image-Text Foundation Models
Chen et al., 2023, PaLI: A Jointly-Scaled Multilingual Language-Image Model

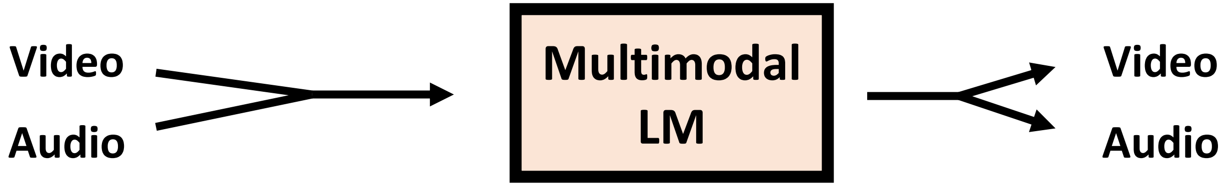
Part 1: Unified/Universal Multimodal Learning

VL-T5 (ICML 2021)

all multimodal tasks via text generation

TVLT (NeurIPS 2022)

video modeling without text (audio as images)

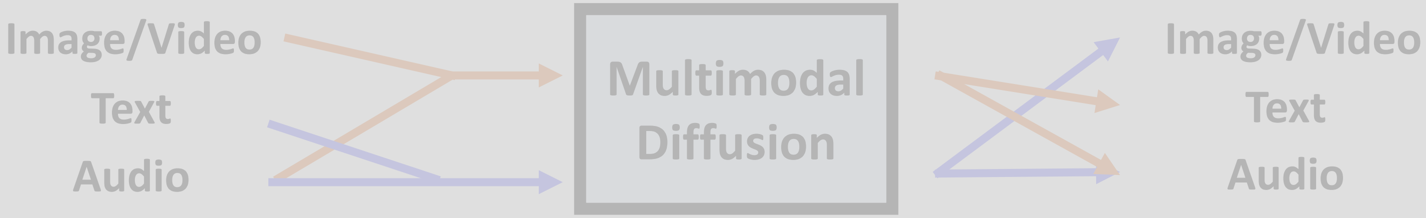


UDOP (CVPR 2023)

document image/text/layout with single architecture

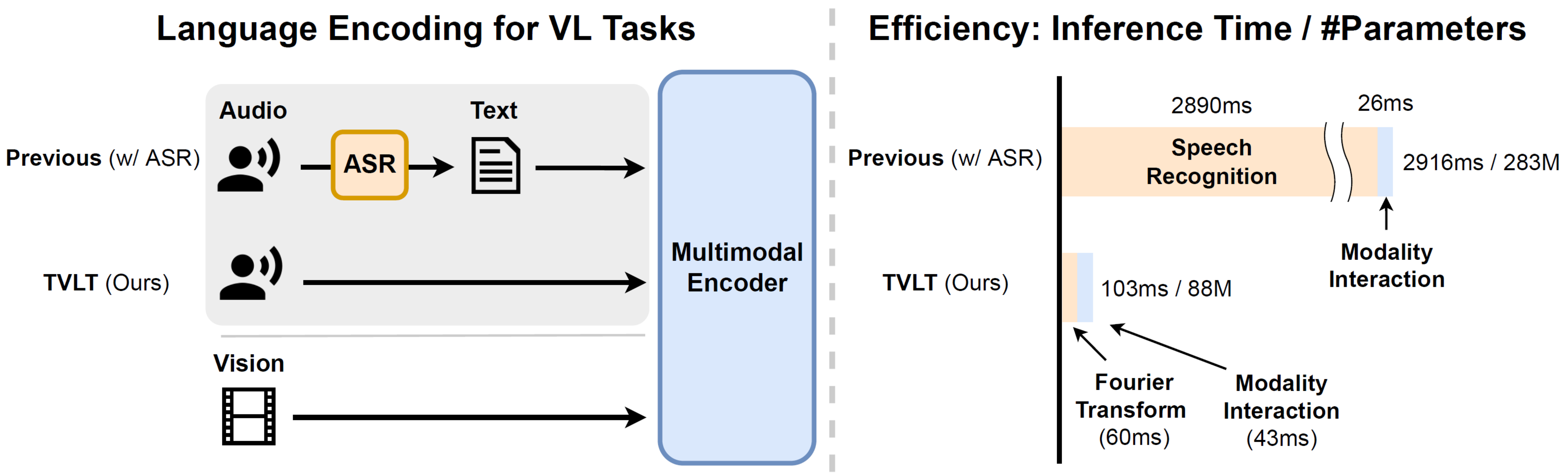
CoDi (NeurIPS 2023)

generating any-to-any input-output modality combination



TVLT: Textless Vision-Language Transformer

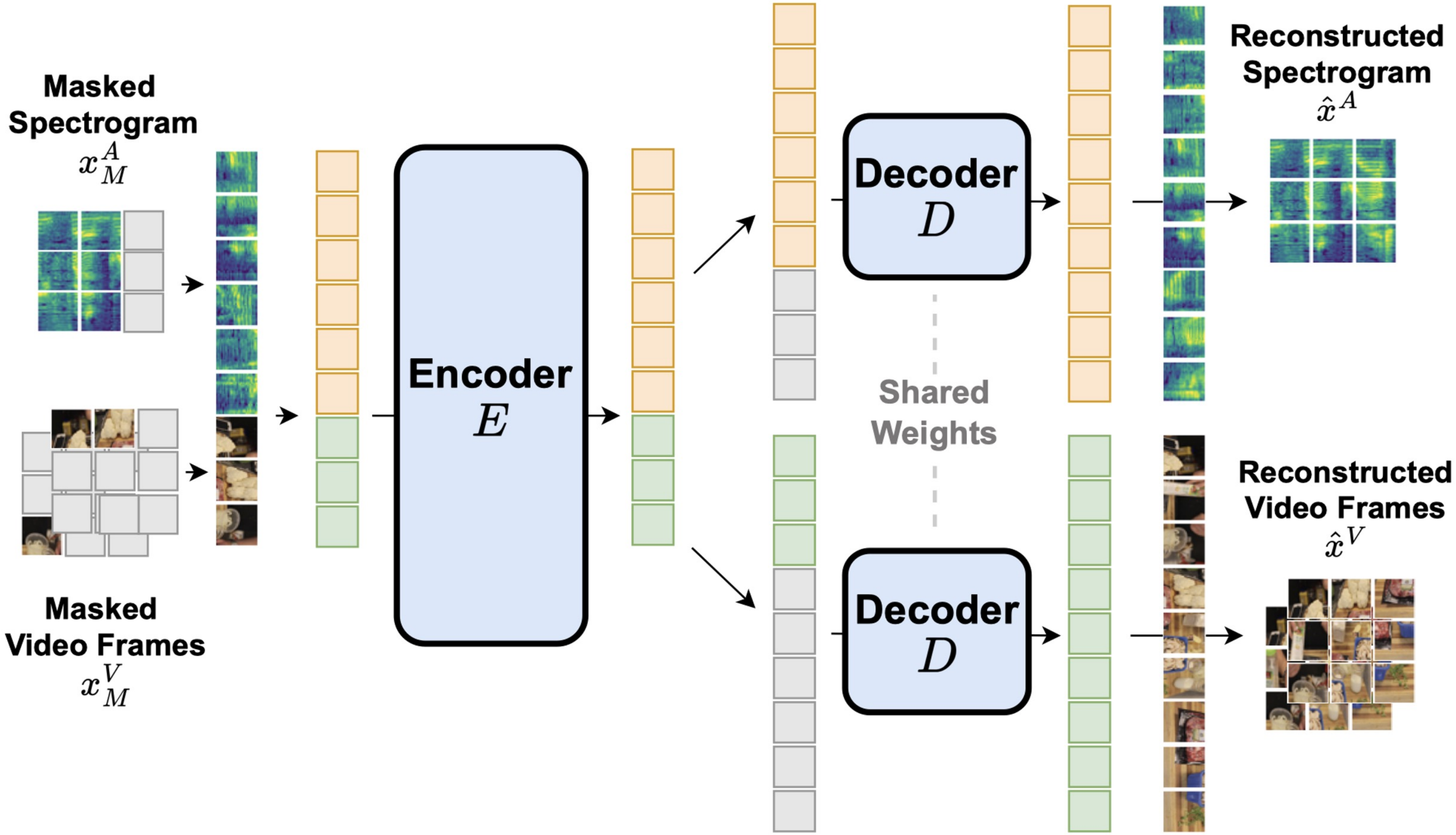
- Unified textless, audio-based homogeneous vision-language transformer
- No ASR/tokenizer/text modules, 28x inference speed, 1/3 #params!



[Tang et al., NeurIPS 2022]

TVLT: Textless Vision-Language Transformer

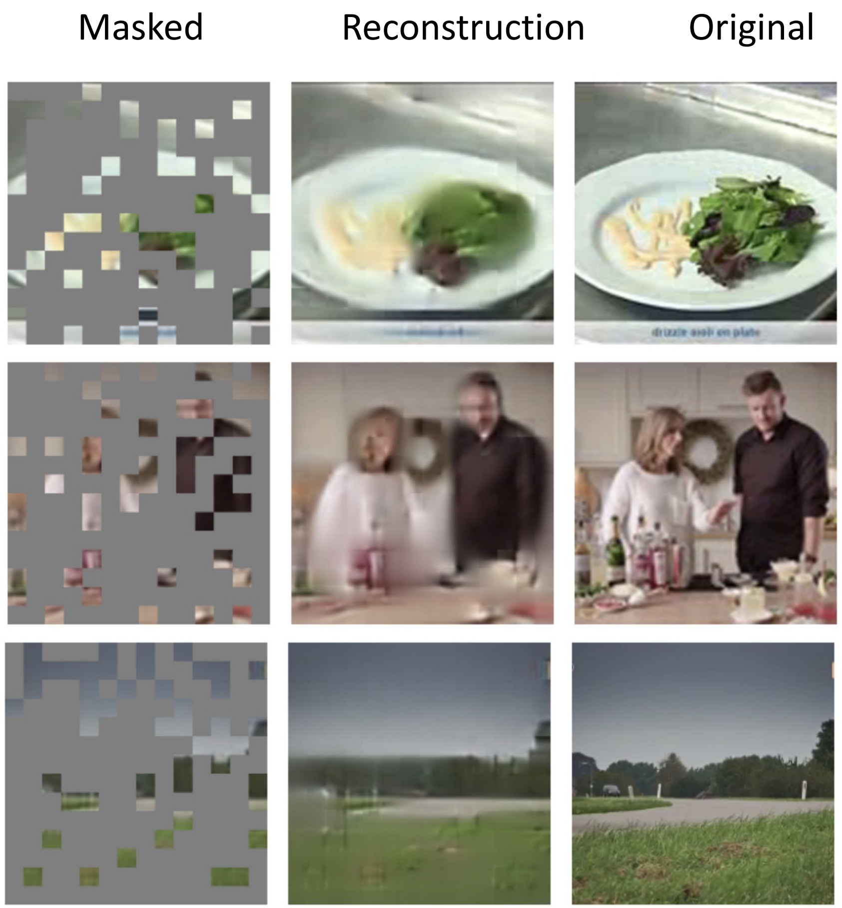
- Unified ViT-style patch embeddings for both video and audio inputs
- MAE-style enc-dec: multimodal joint encoder; decoder weights are shared for video & audio decoding
- Two objectives: (1) masked autoencoding, (2) contrastive learning

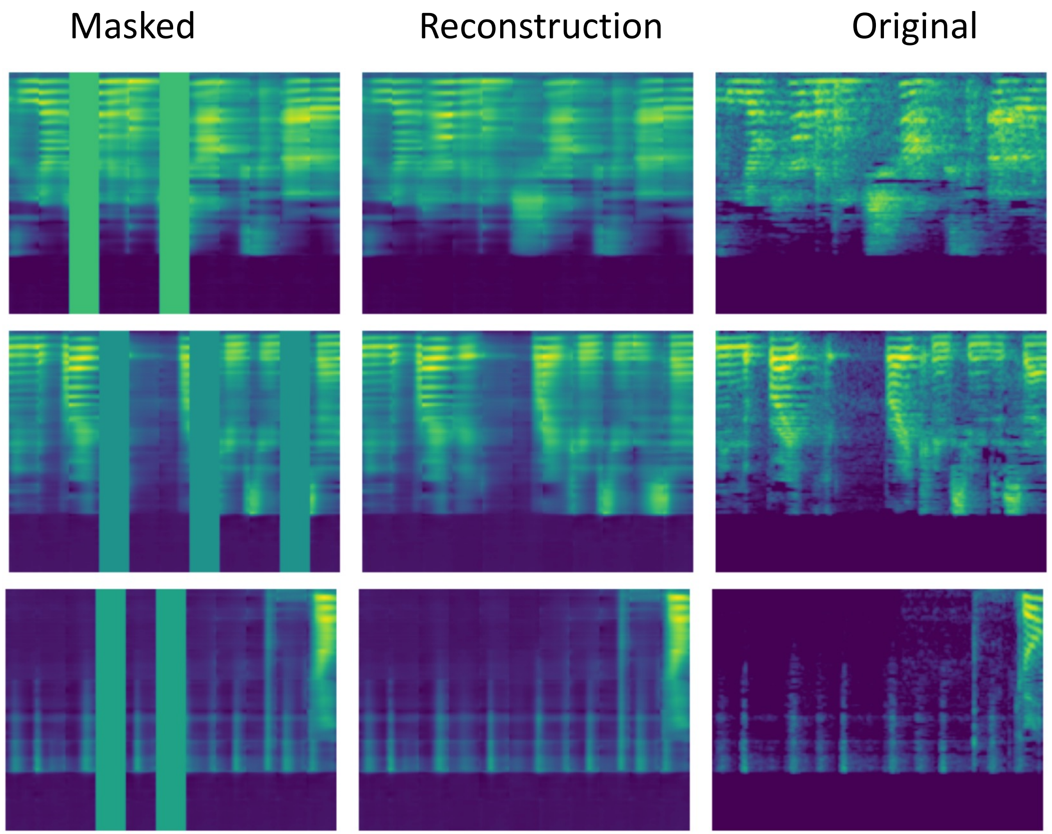


[Tang et al., NeurIPS 2022]

TVLT: Textless Vision-Language Transformer

- Results: Audio-based TVLT (w/o any text modules) performs competitively with text-based model on diverse tasks: image-retrieval, video-retrieval, visual-QA, multimodal sentiment analysis, emotion analysis (while also being much more efficient = 28x faster, 1/3 #parameters)!





[Tang et al., NeurIPS 2022]

Part 1: Unified/Universal Multimodal Learning

VL-T5 (ICML 2021)

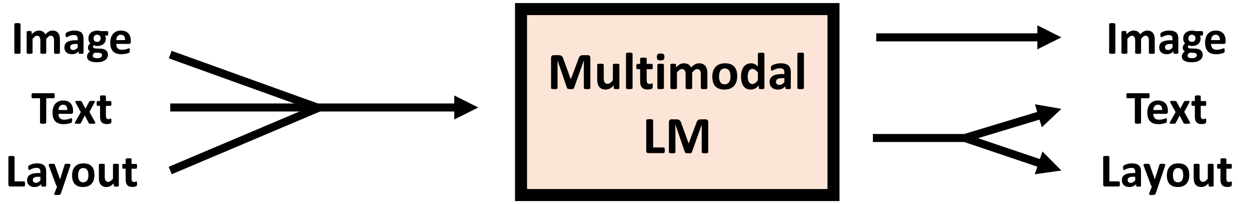
all multimodal tasks via text generation

TVLT (NeurIPS 2022)

video modeling without text (audio as images)

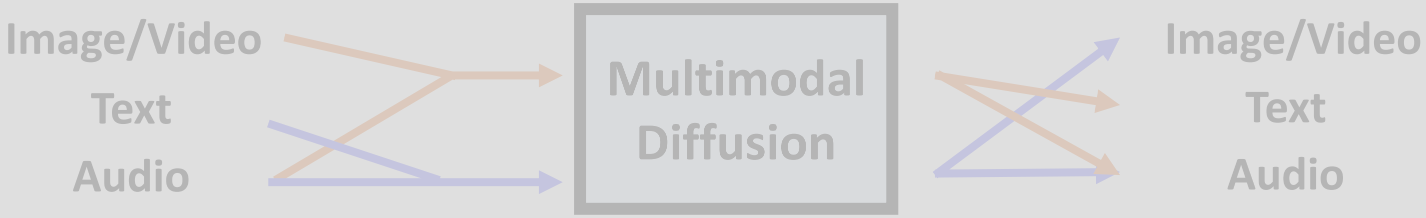
UDOP (CVPR 2023)

document image/text/layout with single architecture



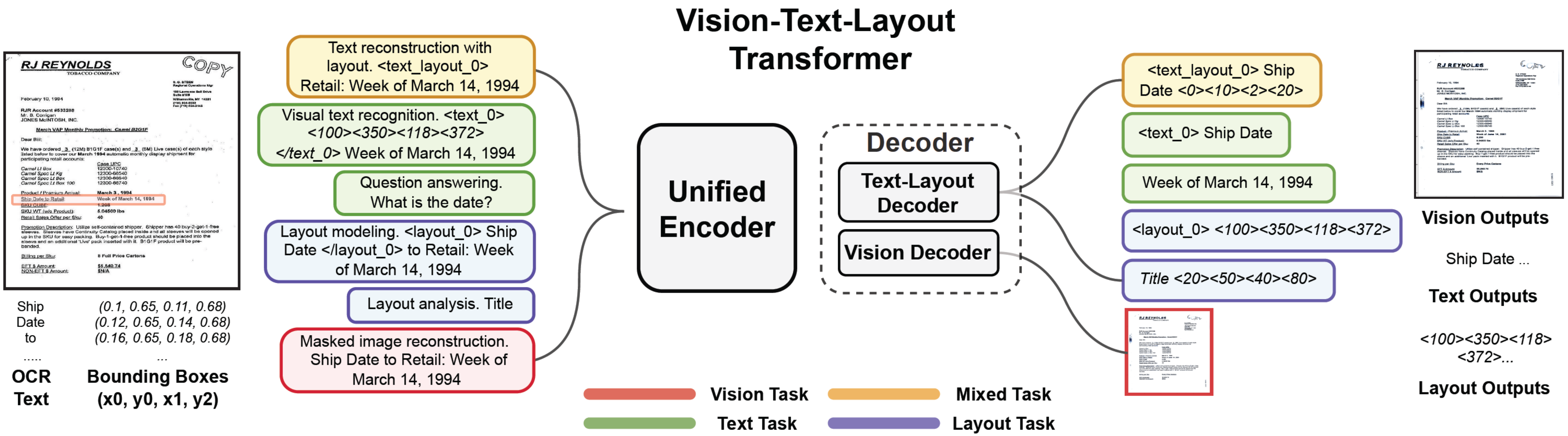
CoDi (NeurIPS 2023)

generating any-to-any input-output modality combination



UDOP: Unifying Vision, Text, Layout for Universal Document Processing

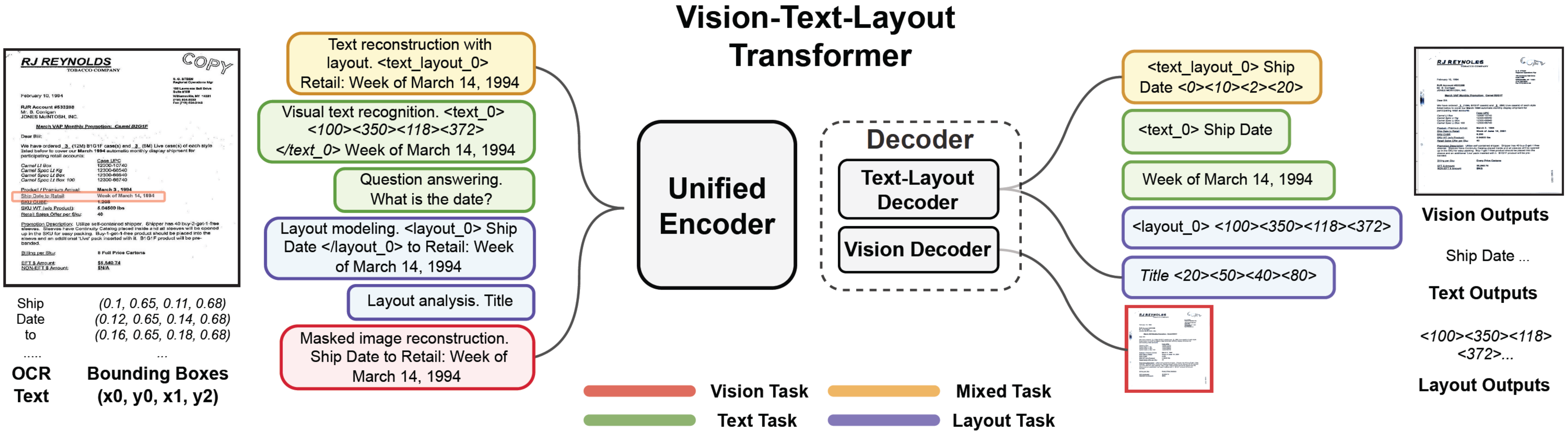
- Unifies text, image, layout modalities (*w/o specialized modules incl. OCR*) with varied task formats, doing document understanding + generation/editing from text+layout modalities via masked image reconstruction.



[Tang et al., CVPR 2023]

UDOP: Unifying Vision, Text, Layout for Universal Document Processing

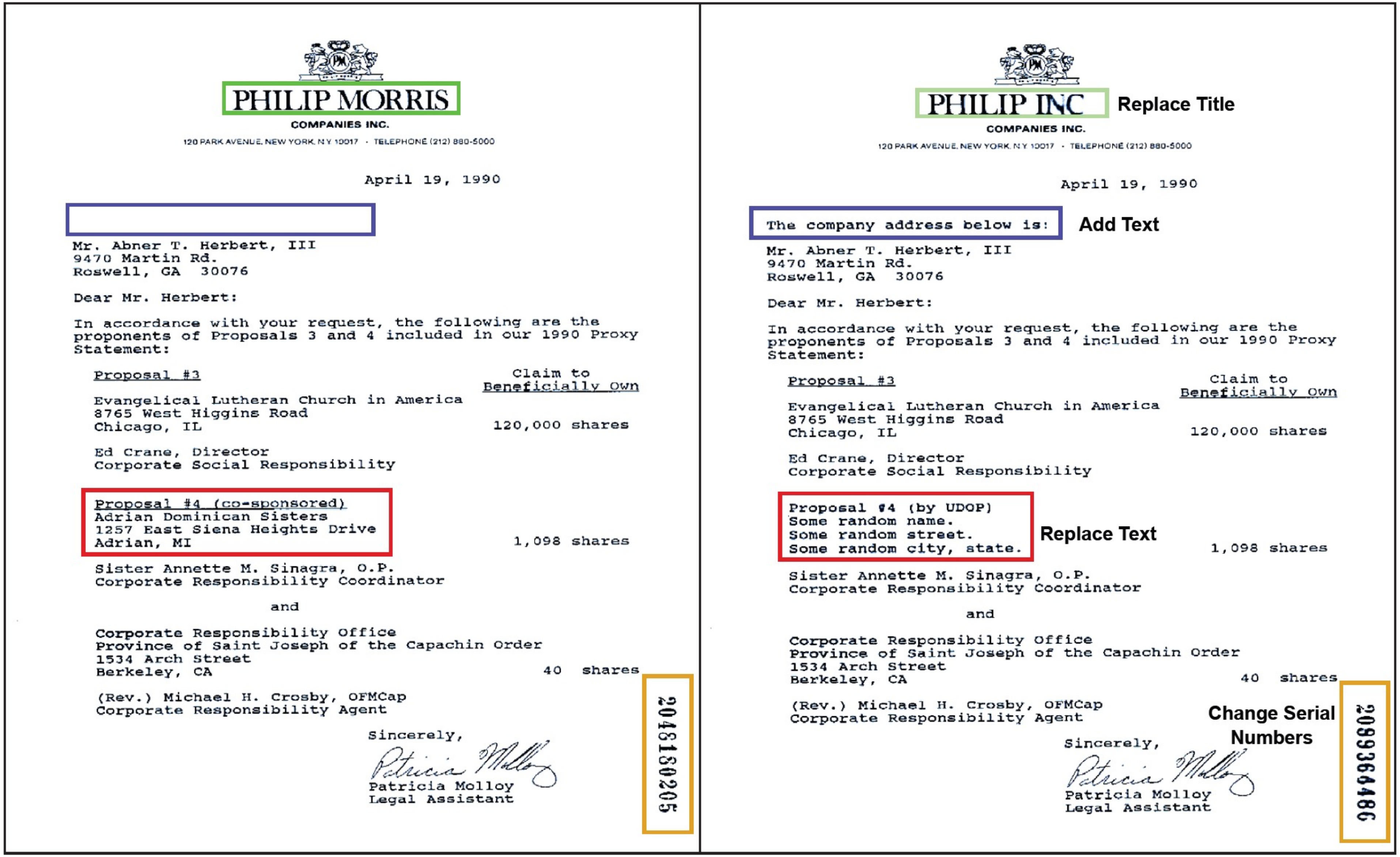
- Unifies text, image, layout modalities (*w/o specialized modules incl. OCR*) with varied task formats, doing document understanding + generation/editing from text+layout modalities via masked image reconstruction.



- State-of-the-art & rank-1 on 8 DocAI tasks / DUE-benchmark, e.g., document-VQA, table-NLI, table-QA, doc-IE, etc. across *diverse data domains like finance reports, academic papers, and websites*.

[Tang et al., CVPR 2023]

UDOP: Unifying Vision, Text, Layout for Universal Document Processing



[Tang et al., CVPR 2023]

Part 1: Unified/Universal Multimodal Learning

VL-T5 (ICML 2021)

all multimodal tasks via text generation

TVLT (NeurIPS 2022)

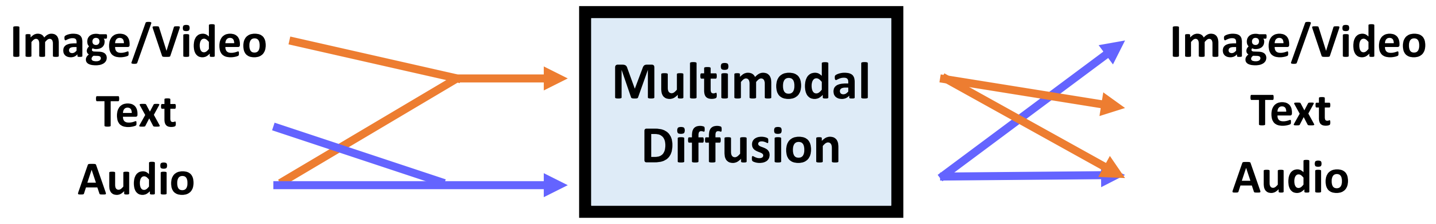
video modeling without text (audio as images)

UDOP (CVPR 2023)

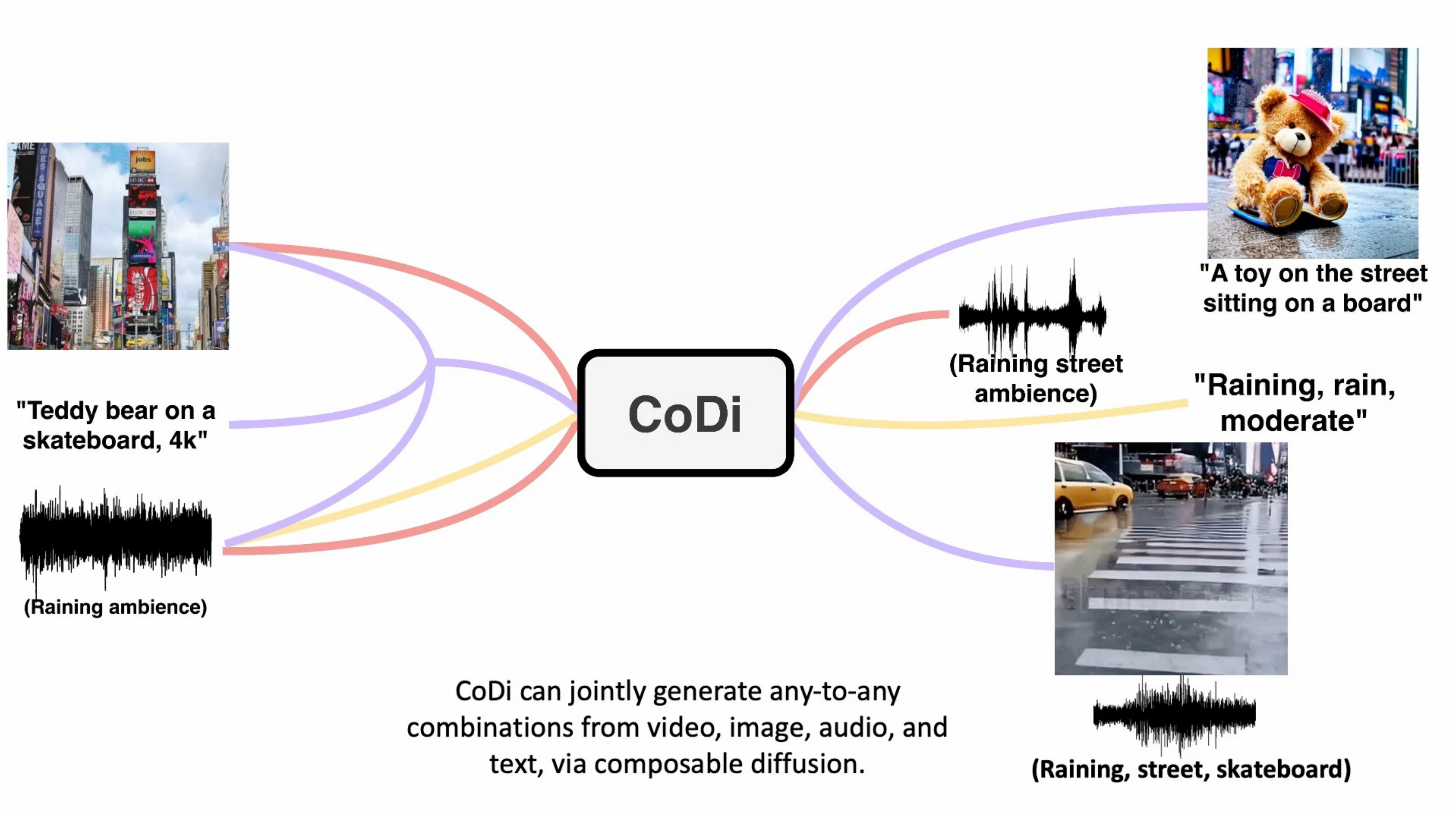
document image/text/layout with single architecture

CoDi (NeurIPS 2023)

generating any-to-any input-output modality combination



CoDi: Any-to-Any Multimodal Generation



https://codi-gen.github.io/

[Tang et al., NeurIPS 2023]

CoDi: Any-to-Any Multimodal Generation

- New generative-AI foundation model that allows **any combination of input modalities** & **generates any combination of output modalities** (text, audio, image, video) – can help create diverse **'many-modal' stories** using different types of inputs on the storyboard!

- **BUT** training such a model presents **significant costs**, as the # combinations for input and output modalities scales **exponentially** & training datasets **missing** for many combinations of modalities.

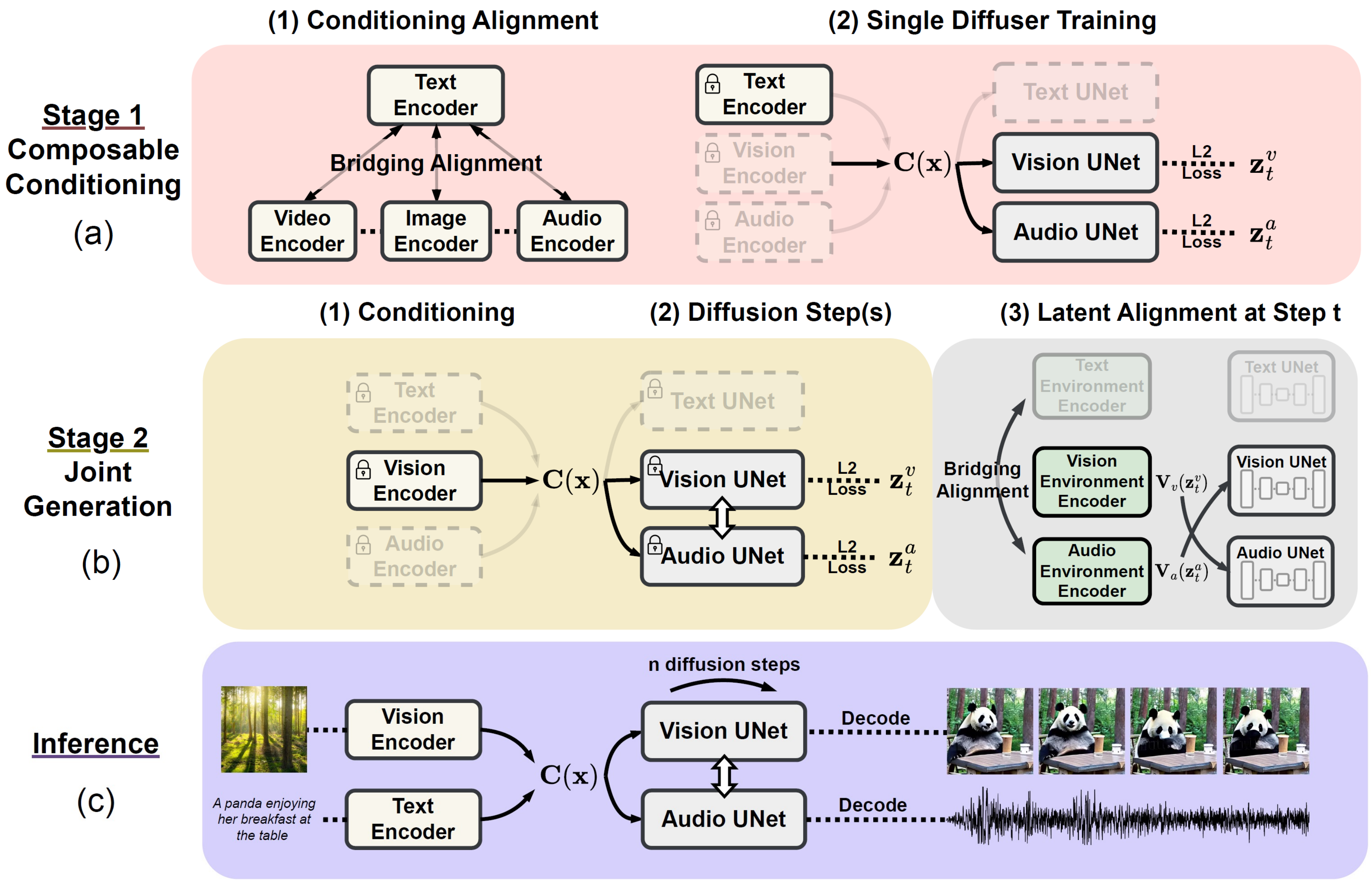
- We propose "**Bridging Alignment**" strategy to **efficiently model the exponential number** of input-output combinations with a **linear number** of training objectives.

- Allows CoDi to freely condition on any input combination+generate any group of modalities, even if not present in the training data.

https://codi-gen.github.io/

[Tang et al., NeurIPS 2023]

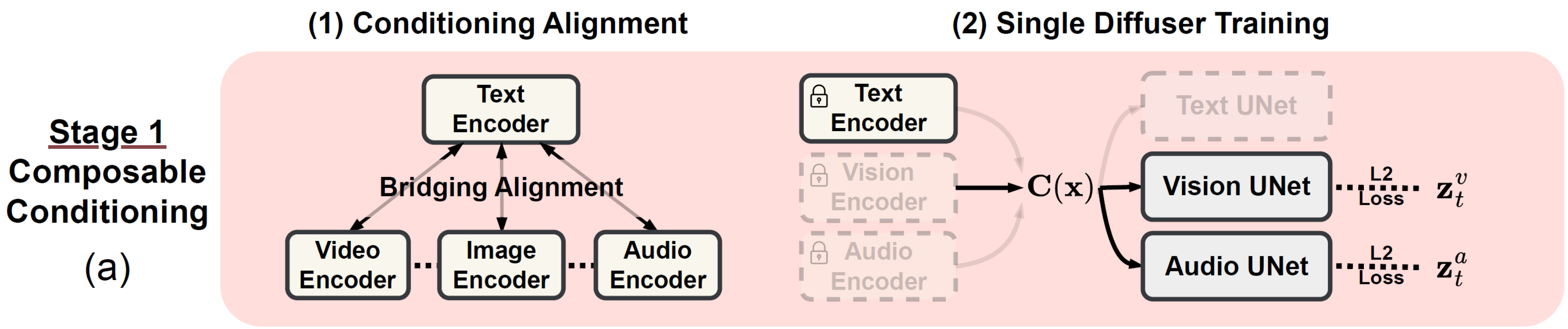
CoDi: Any-to-Any Multimodal Generation



https://codi-gen.github.io/

[Tang et al., NeurIPS 2023]

CoDi: Any-to-Any Multimodal Generation

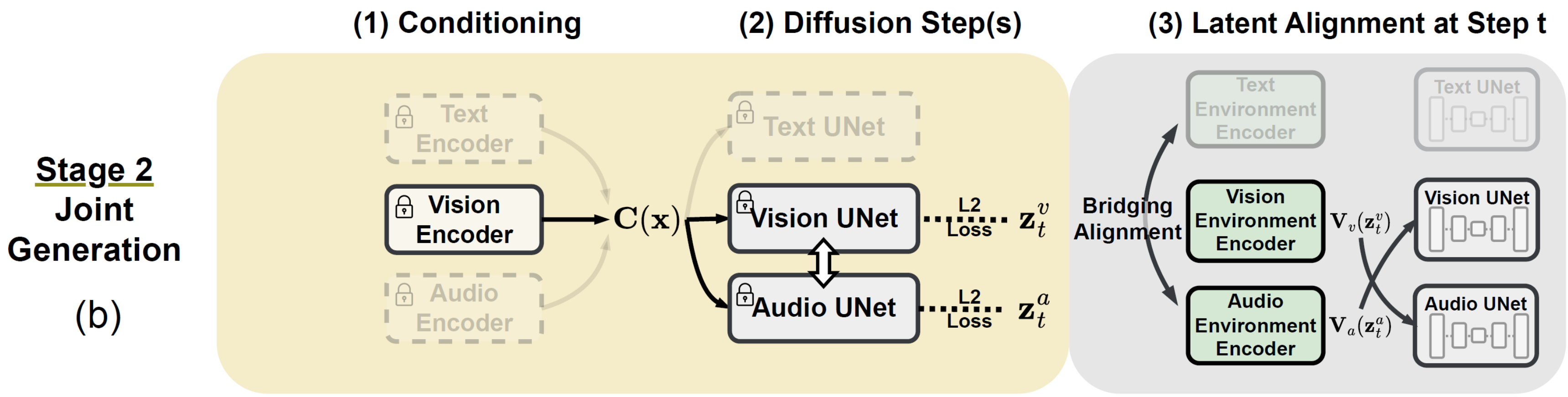


- <u>**Stage 1**</u>: We train a **latent diffusion model (LDM)** for each modality. They can be trained **independently**, ensuring high-quality generation for each modality. For conditional generation, e.g., *audio+language→image*, the input modalities are projected into a **shared feature space**, and the **output LDM attends to this combination of input features**.

- This multimodal conditioning mechanism prepares the diffusion model to **condition on any combination of modalities without directly training** for such settings.

https://codi-gen.github.io/ [Tang et al., NeurIPS 2023]

CoDi: Any-to-Any Multimodal Generation

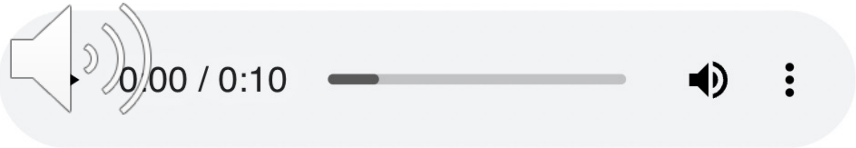


- **<u>Stage 2</u>**: We add a **cross-attention module** to each LDM and an **environment encoder** to project the **LDM latent variables into a shared/mixed space**.

- This enables CoDi to seamlessly **mix/generate any group of output modalities, w/o training** on all generation combinations (with linear # training objectives).

CoDi: Any-to-Any Multimodal Generation

Audio + Image → Text + Image



CoDi: Any-to-Any Multimodal Generation

Audio + Image → Text + Image

0:00 / 0:10

"Playing piano in a forest."

CoDi: Any-to-Any Multimodal Generation

Text + Image → Video

"Eating on a coffee table."

https://codi-gen.github.io/

[Tang et al., NeurIPS 2023]

CoDi: Any-to-Any Multimodal Generation

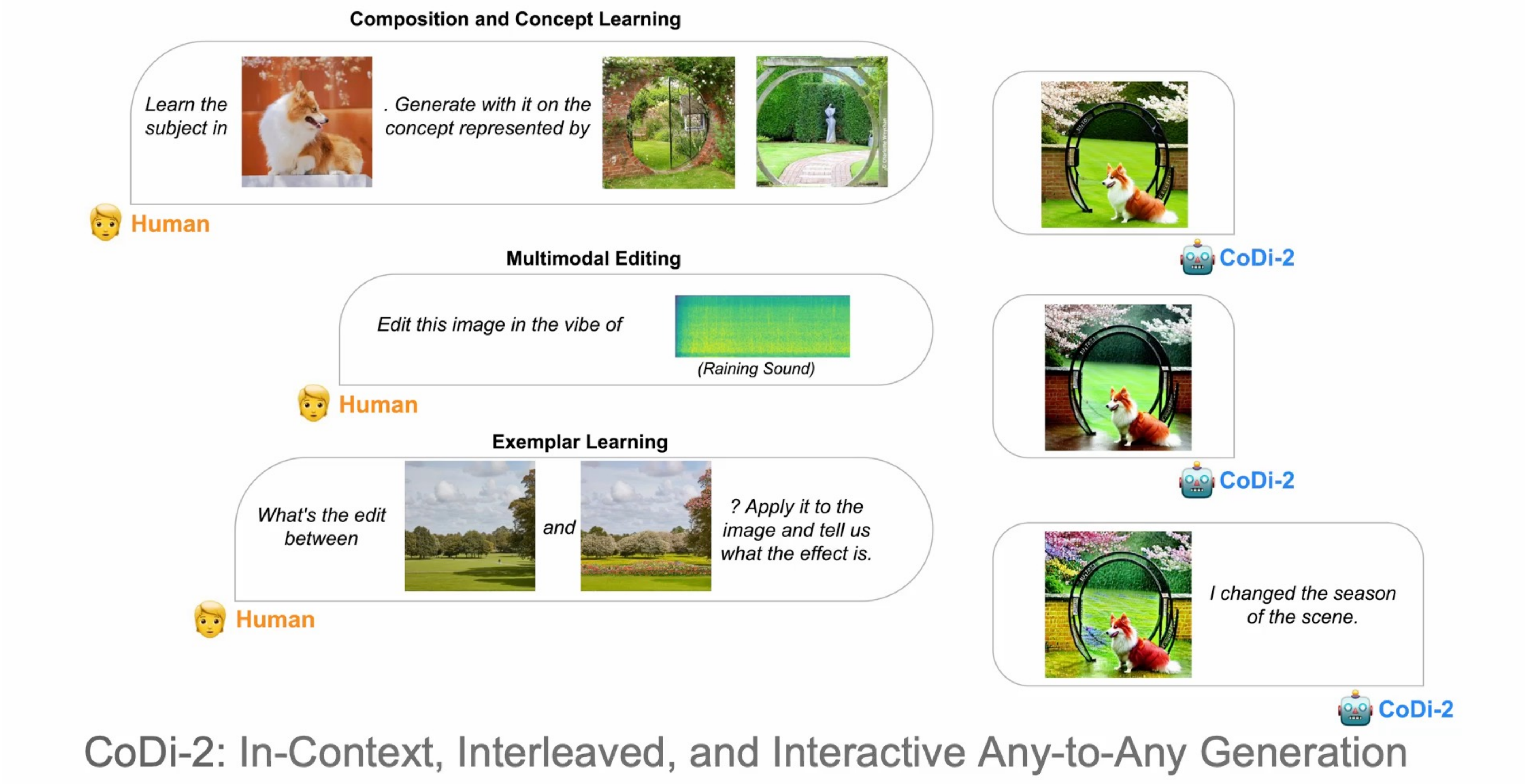
Text + Image → Video

"Eating on a coffee table."

https://codi-gen.github.io/

[Tang et al., NeurIPS 2023]

CoDi-2: In-Context, Interleaved, Interactive Any-to-Any Generation



https://codi-2.github.io/

[Tang et al., Preprint 2023]

Talk Outline

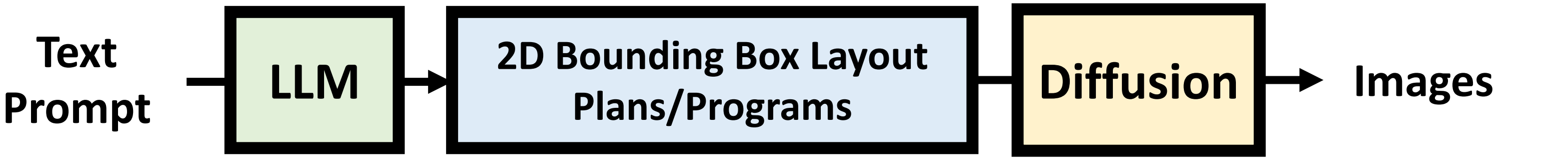
A journey of multimodal generative LLMs for enhancing their unification, interpretable planning/programming, evaluation:

- **Unified/Universal Multimodal Learning** (for Generalizability, Shared Knowledge, Efficiency)
  - VLT5: Unifying Vision-and-Language Tasks via Text Generation [ICML 2021]
  - TVLT: Textless Vision-Language Transformer [NeurIPS 2022]
  - UDOP: Unifying Vision, Text, and Layout for Universal Document Processing [CVPR 2023]
  - CoDi: Any-to-Any Generation via Composable Diffusion [NeurIPS 2023] & *CoDi-2: In-Context, Interleaved, and Interactive Any-to-Any Generation [2023]*
- **Interpretable Multimodal Generation via LLM Planning/Programming** (for Understanding, Control, Faithfulness)
  - VPGen: Step-by-Step Text-to-Image Generation with Interpretable Visual Programming [NeurIPS 2023]
  - VideoDirectorGPT: Consistent Multi-Scene Video Generation via LLM-Guided Planning [2023]
  - DiagrammerGPT: Generating Open-Domain, Open-Platform Diagrams via LLM Planning [2023]
- **Evaluation of Multimodal Generation Models** (of Fine-grained Skills, Faithfulness, Social Biases)
  - DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models [ICCV 2023]
  - VPEval: Step-by-Step Text-to-Image Evaluation with Interpretable Visual Programming [NeurIPS 2023]
  - *Davidsonian Scene Graph: Improving Reliability in Fine-grained Evaluation for Text-to-Image Generation [2023]*
- **Next Big Challenges**: trade-offs, structure, non-verbal, interaction, reasoning, causality, long-distance fine-grained evaluation, efficiencies

Part 2: Interpretable Multimodal Generation with LLM Planning

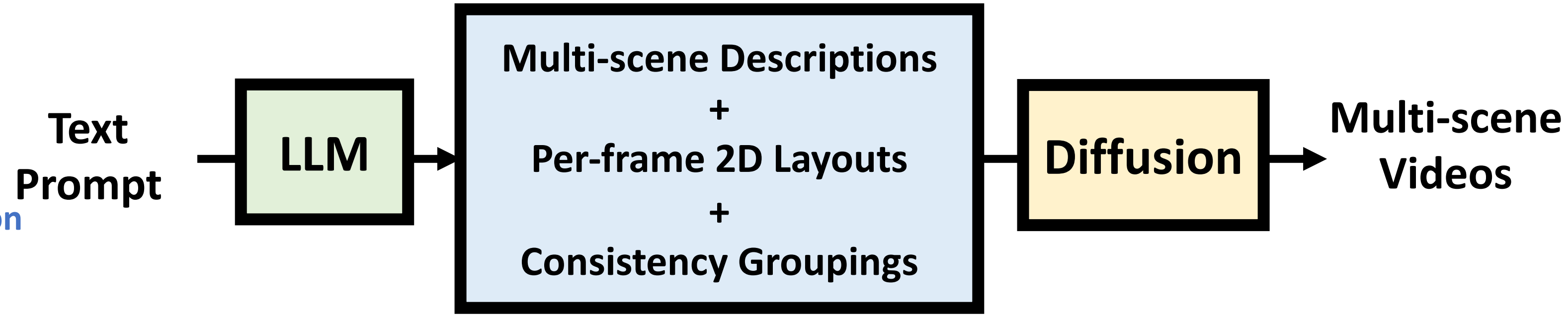
VPGen (NeurIPS 2023)

LLM Planning for Image Generation



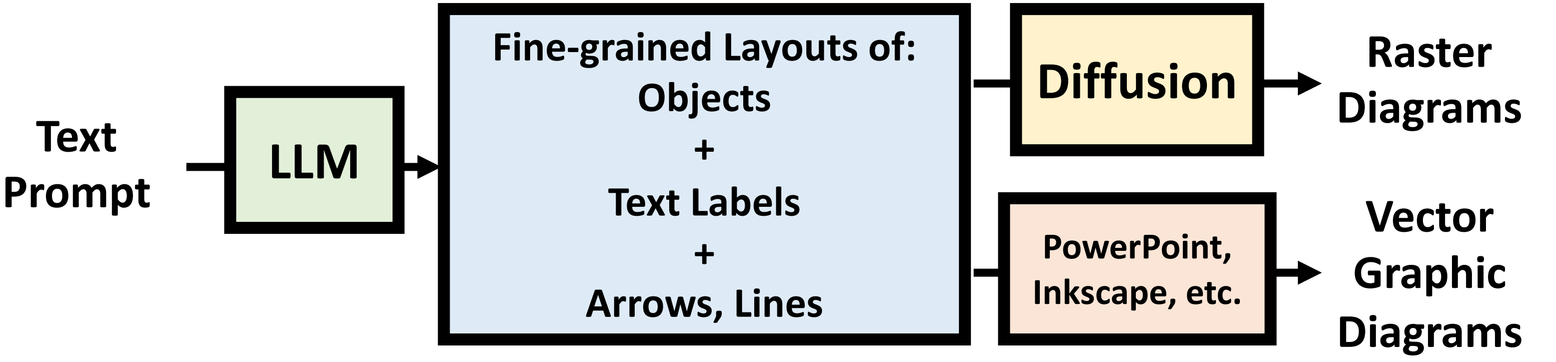
VideoDirectorGPT (2023)

LLM Planning for Multi-Scene, Consistent Video Generation



DiagrammerGPT (2023)

LLM Planning for Open-Domain Diagram Generation



Background: Text-to-Image Generation with Blackbox Models

`two Pikachus on a table`

`A truck is behind a motorcycle`

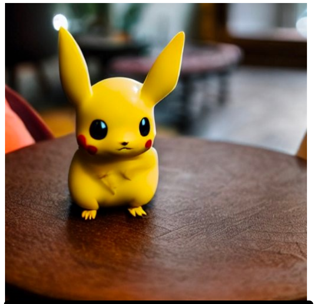
Blackbox Text-to-Image Generation (e.g., DALL-E, Imagen, Stable Diffusion)

[Cho et al., NeurIPS 2023]

Background: Text-to-Image Generation with Blackbox Models

two Pikachus on a table

Good visual quality!



one Pikachu ✗

A truck is **behind** a motorcycle

truck is **below** motorcycle ✗

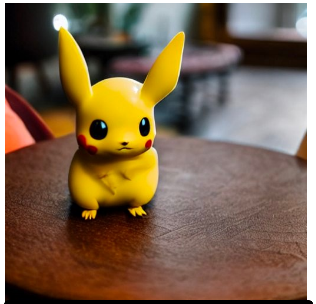
[Cho et al., NeurIPS 2023]

Background: Text-to-Image Generation with Blackbox Models

two Pikachus on a table

A truck is behind a motorcycle

Good visual quality! But important semantic issues...



one Pikachu ❌

truck is below motorcycle ❌

[Cho et al., NeurIPS 2023]

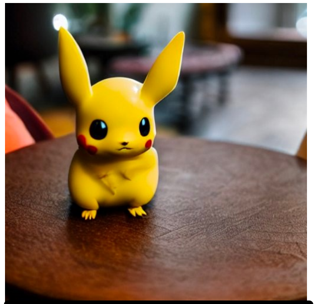
Background: Text-to-Image Generation with Blackbox Models

two Pikachus on a table

A truck is behind a motorcycle

Good visual quality! But important semantic issues…

- **lack of fine-grained layout planning/control**
- **lack of interpretability behind generation process**
- **lack of faithfulness to input (incl. hallucinations)**



one Pikachu ❌

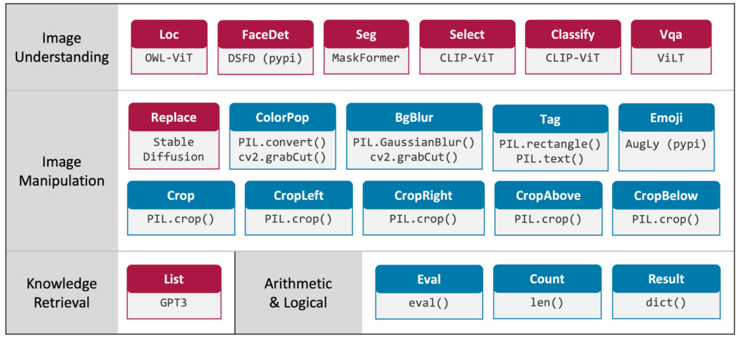
truck is below motorcycle ❌

[Cho et al., NeurIPS 2023]

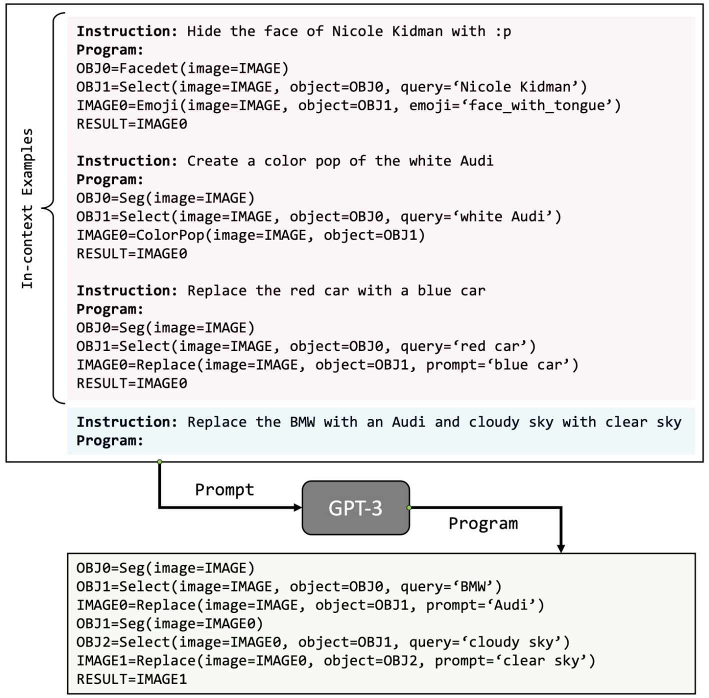
Background: Visual Programming

NMNs (Andreas et al., CVPR 2016), MattNet (Yu et al., CVPR 2018)...SummProg (Saha et al., ICLR 2023), VisProg (Gupta and Kembhavi, CVPR 2023), ViperGPT (Surís et al., 2023)

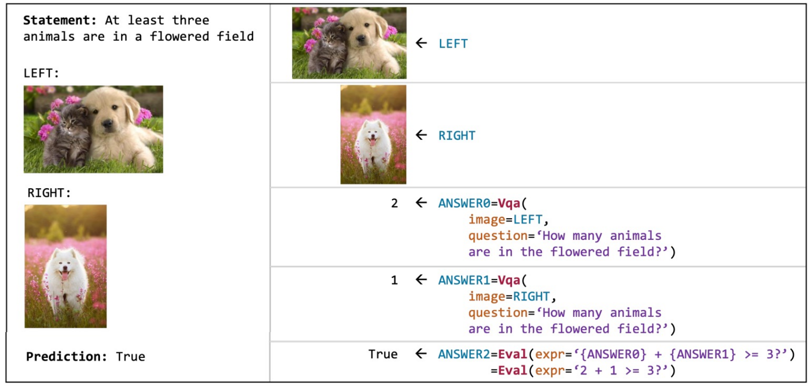
1) Define visual modules



2) Generate programs w/ LLM



3) Execute programs for reasoning tasks



Andreas et al., 2016, Neural Module Networks
Yu et al., 2018, MAttNet: Modular Attention Network for Referring Expression Comprehension
Saha et al., 2023, Summarization programs: Interpretable abstractive summarization with neural modular trees
Gupta and Kembhavi, 2023, Visual Programming: Compositional Visual Reasoning Without Training
Surís et al., 2023, ViperGPT: Visual Inference via Python Execution for Reasoning

VPGen: Visual Programming for Step-by-Step T2I Generation

```
two Pikachus
on a table
```

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

[Cho et al., NeurIPS 2023]

VPGen: Visual Programming for Step-by-Step T2I Generation

two Pikachus
on a table

Object/Count Generation

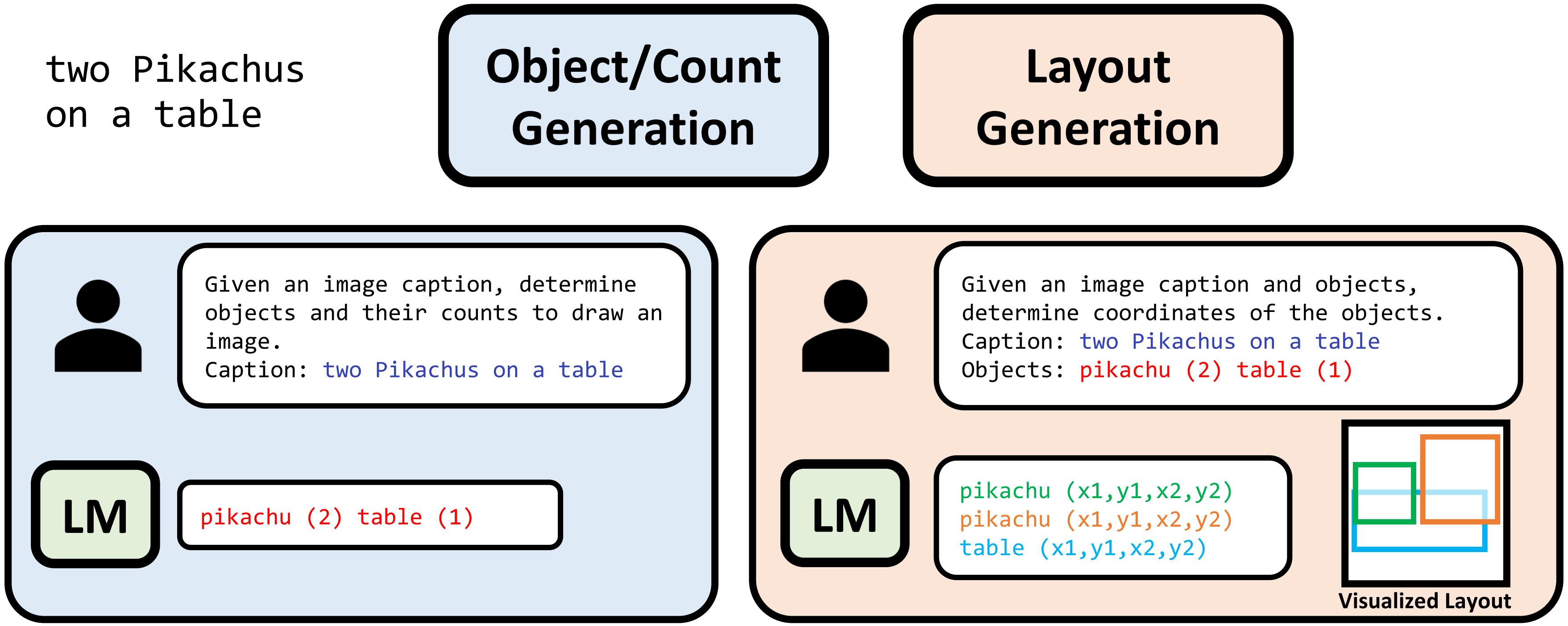
Given an image caption, determine objects and their counts to draw an image.
Caption: two Pikachus on a table

LM

pikachu (2) table (1)

VPGen: Visual Programming for Step-by-Step T2I Generation

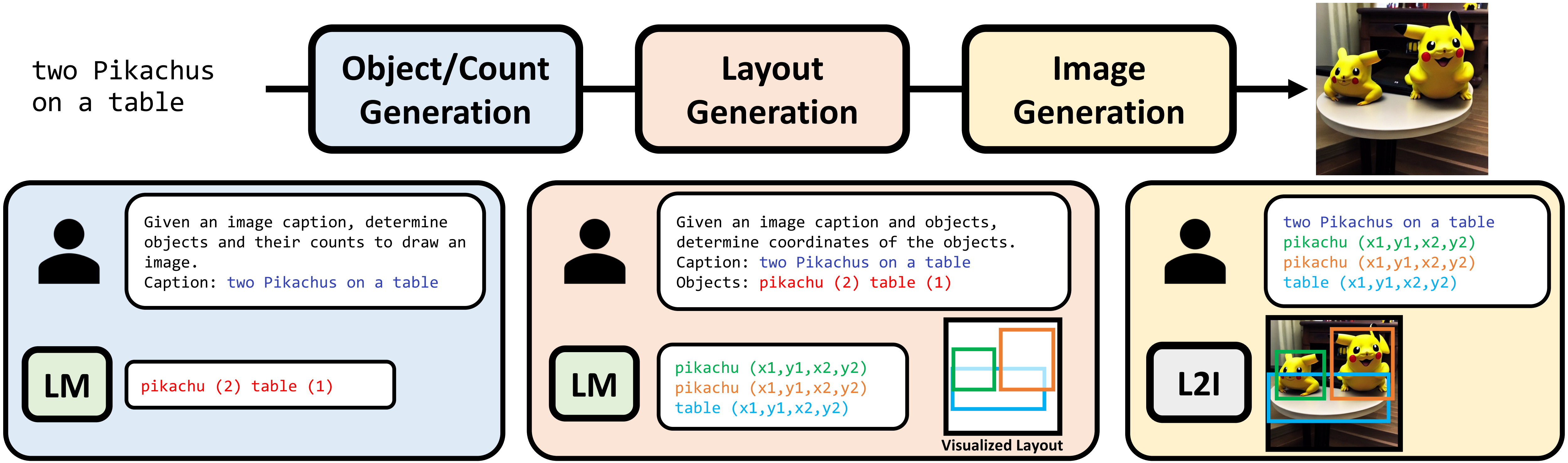
two Pikachus
on a table



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

[Cho et al., NeurIPS 2023]

VPGen: Visual Programming for Step-by-Step T2I Generation



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

[Cho et al., NeurIPS 2023]

Skill-based Results

Our VPGen shows improved spatial control

- Generation via layout programs promotes better **understanding+planning** of structure/scale/spatial relations (also allows **explicit control** over these properties via manual, **interpretable corrections of unfaithful parts**)!

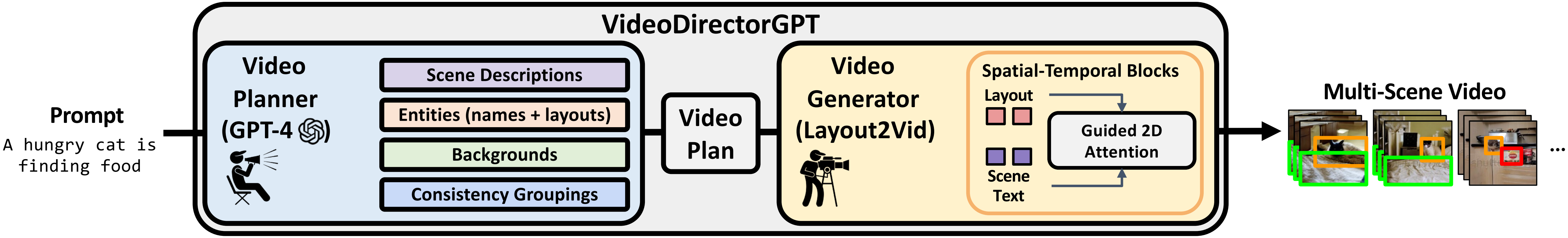
Model	VPEVAL Skill Score (%) ↑					
	Object	Count	Spatial	Scale	Text Rendering	Average
Stable Diffusion v1.4	**97.3**	47.4	22.9	11.9	**8.9**	37.7
Stable Diffusion v2.1	96.5	53.9	31.3	14.3	6.9	40.6
Karlo	95.0	59.5	24.0	16.4	**8.9**	40.8
minDALL-E	79.8	29.3	7.0	6.2	0.0	24.4
DALL-E Mega	94.0	45.6	17.0	8.5	0.0	33.0
VPGEN (F30)	96.8	55.0	39.0	23.3	5.2	43.9
VPGEN (F30+C+P)	96.8	**72.2**	**56.1**	**26.3**	3.7	**51.0**

Large improvements on structural control:
- Counting
- Spatial relation
- Relative size/scale comparison

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

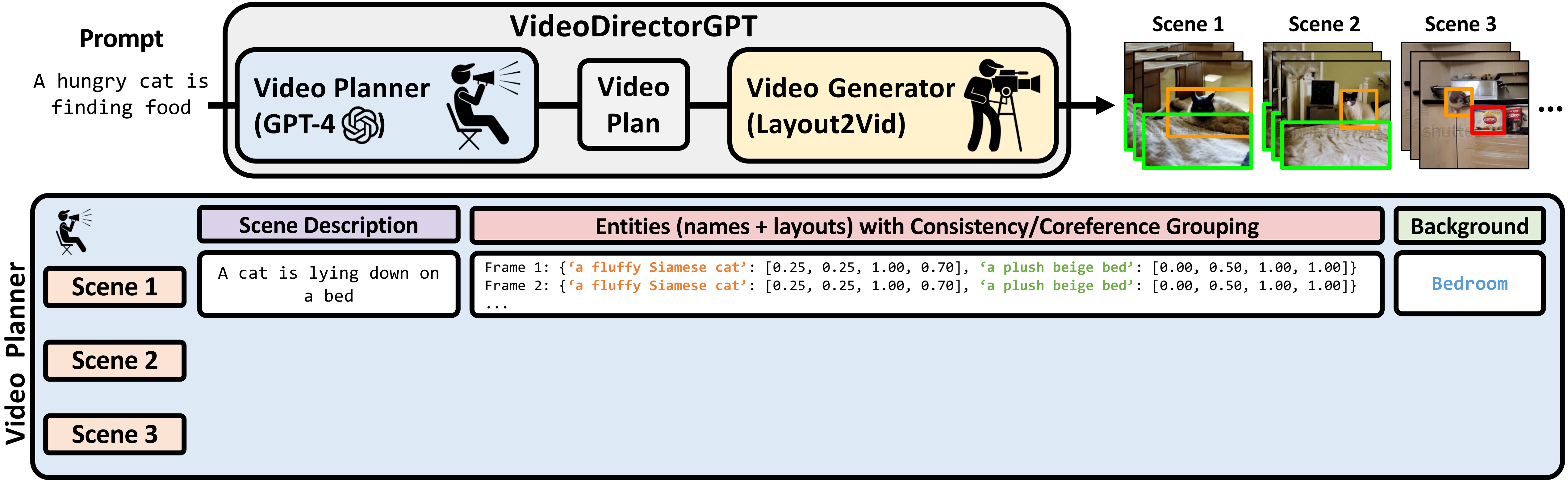
[Cho et al., NeurIPS 2023]

VideoDirectorGPT: Consistent Multi-Scene Video Generation via LLM-Guided Planning



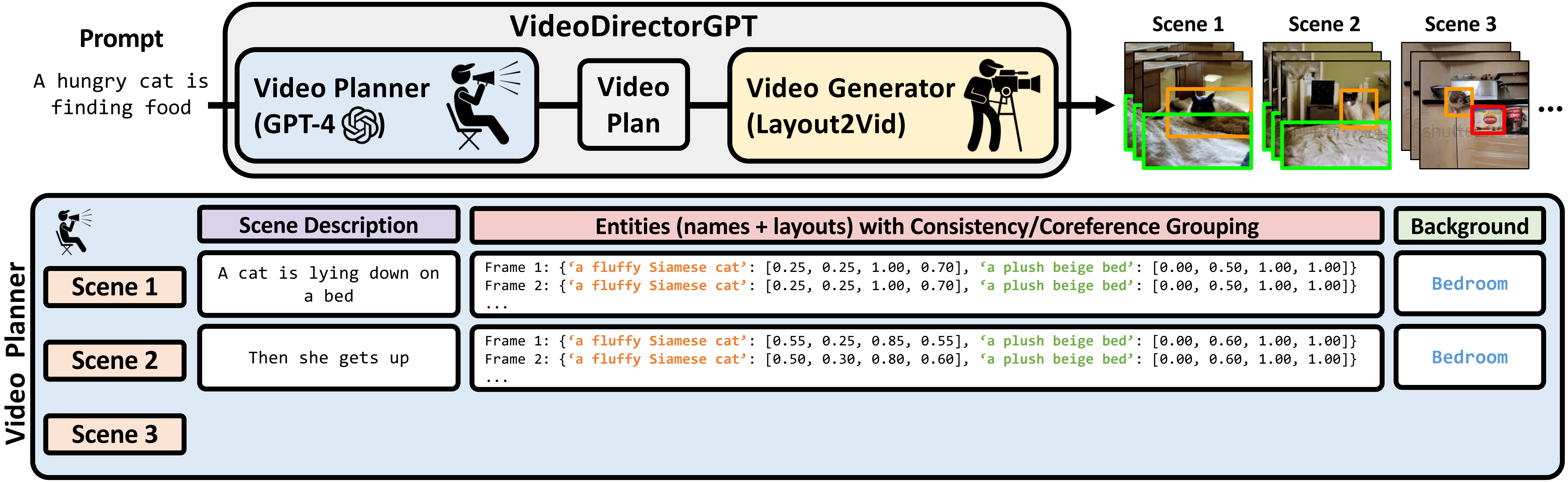
https://videodirectorgpt.github.io/

[Lin et al., 2023]



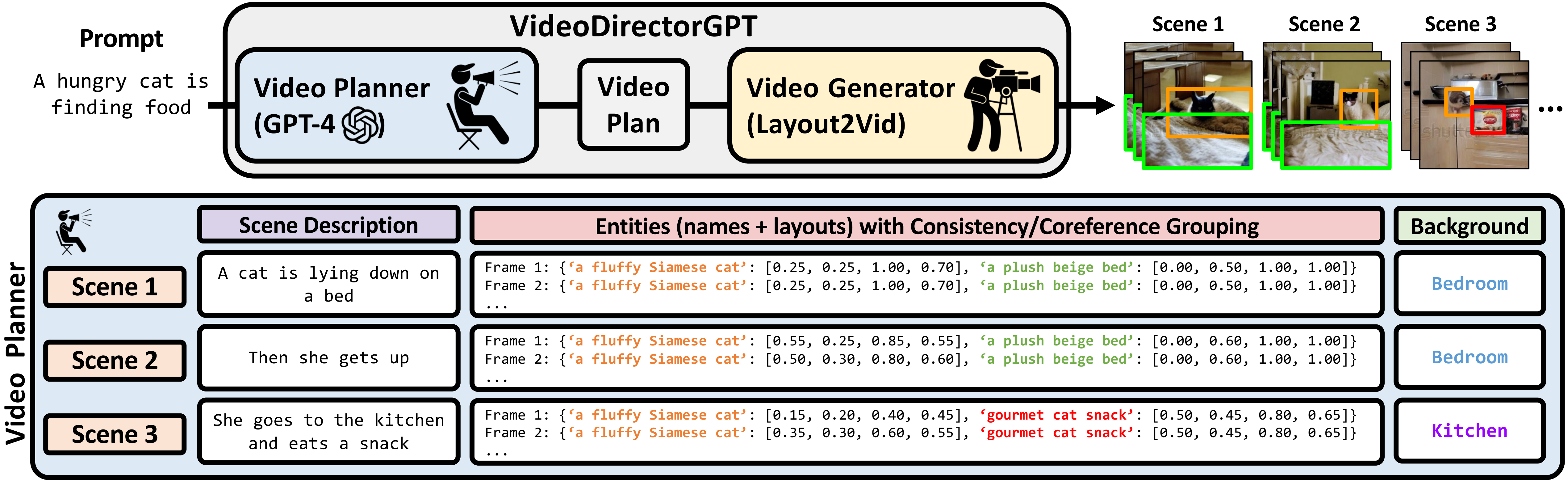
https://videodirectorgpt.github.io/

[Lin et al., 2023]



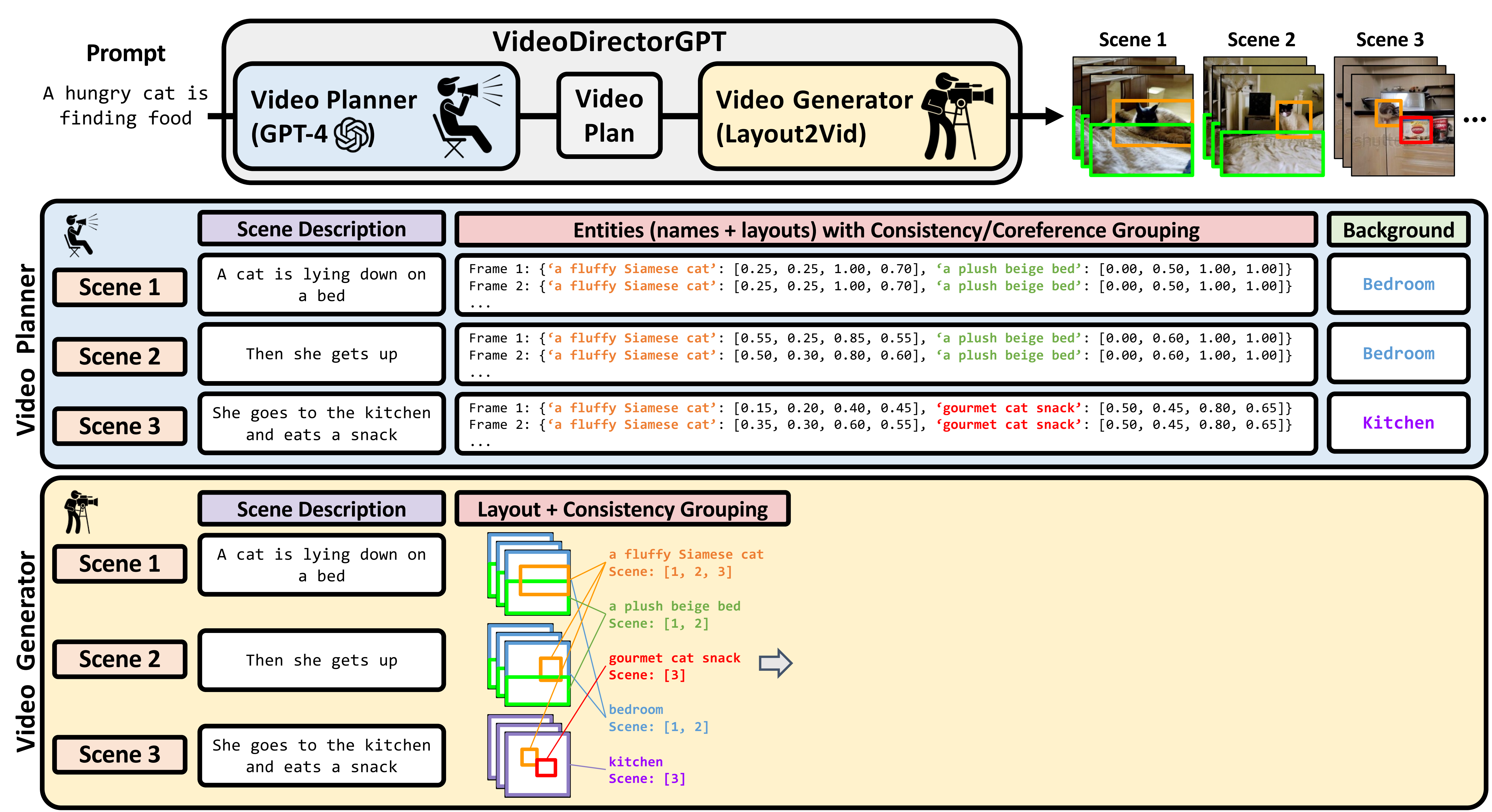
https://videodirectorgpt.github.io/

[Lin et al., 2023]



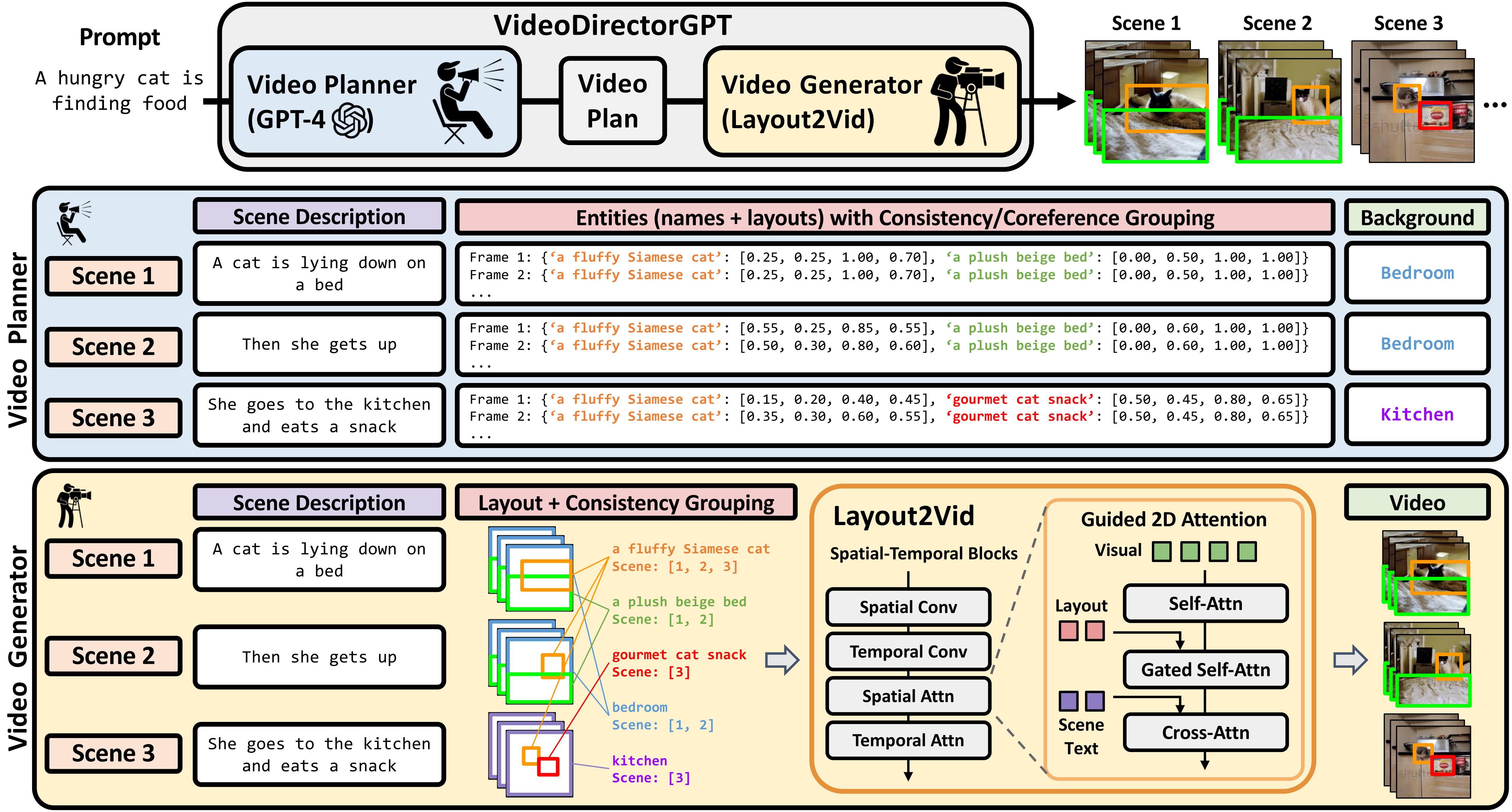
https://videodirectorgpt.github.io/

[Lin et al., 2023]



https://videodirectorgpt.github.io/

[Lin et al., 2023]



https://videodirectorgpt.github.io/

[Lin et al., 2023]

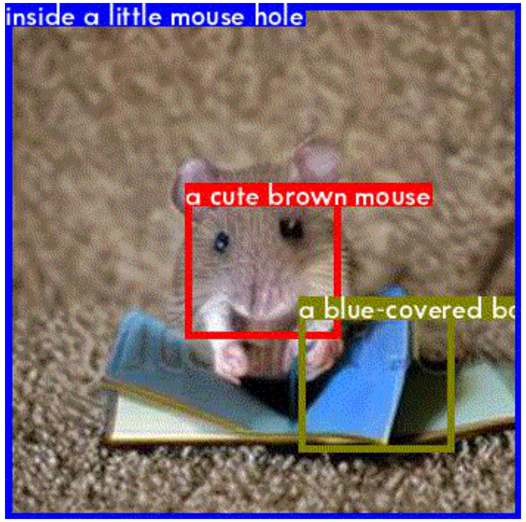
Multi-Sentence to Multi-Scene Video (Coref-SV)

Scene 1: **mouse** is holding a book and makes a happy face.
Scene 2: **he** looks happy and talks.
Scene 3: **he** is pulling petals off the flower.
Scene 4: **he** is ripping a petal from the flower.
Scene 5: **he** is holding a flower by **his** right paw.
Scene 6: one paw pulls the last petal off the flower.
Scene 7: **he** is smiling and talking while holding a flower on **his** right paw.

ModelScopeT2V

❌ fails to keep "mouse" through all scenes

VideoDirectorGPT (Ours)



✓ the "mouse" is consistent through all scenes + layout control

https://videodirectorgpt.github.io/

[Lin et al., 2023]

Single Sentence to Multi-Scene Video (HiREST)

make a strawberry surprise

GPT-4 generated sub-scene descriptions:

- a young man in a red apron washes ripe red strawberries in a silver sink
- a young man in a red apron carefully cuts the strawberries on a wooden chopping board with a sharp knife
- a young man in a red apron places cut strawberries, banana, and Greek yogurt into an electric blender
- a young man in a red apron blends ingredients together until smooth in an electric blender
- a young man in a red apron pours the smoothie into a tall glass
- a young man in a red apron places a scoop of vanilla ice cream on top of the smoothie in a tall glass
- a young man in a red apron places a strawberry on top of the ice cream for garnishing
- a young man in a red apron serves the Strawberry Surprise on a ceramic plate

ModelScopeT2V

❌ no actual process shown on how to "make" the strawberry surprise

VideoDirectorGPT (Ours)

✔ step-by-step + consistent process on how to "make" the strawberry surprise

https://videodirectorgpt.github.io/

[Lin et al., 2023]

Single Sentence to Single-Scene Video (ActionBench-Direction)

pushing **stuffed animal** from **left to right**

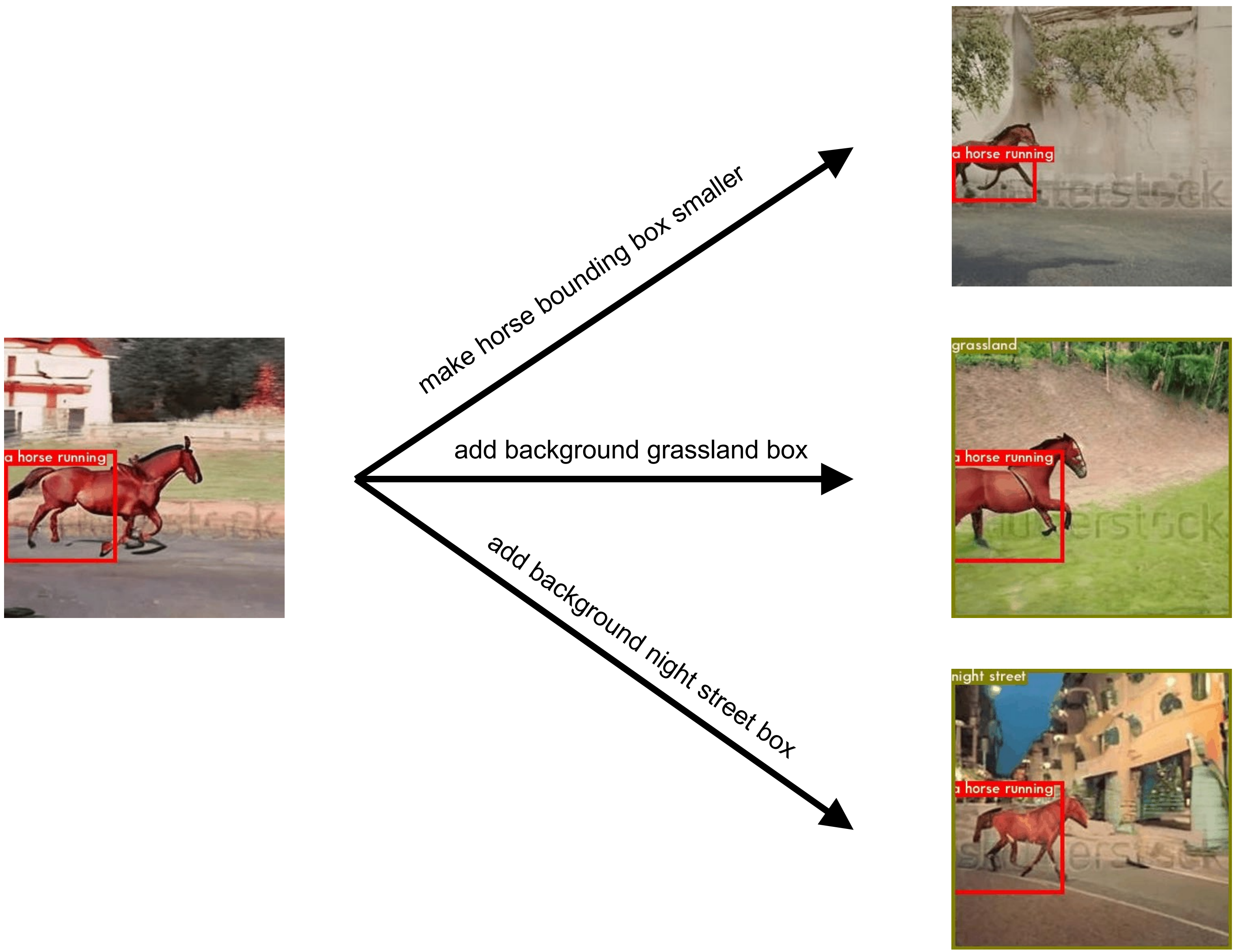
ModelScopeT2V

❌ fails to move the "stuffed animal"

VideoDirectorGPT (Ours)

✓ correctly moves the "stuffed animal" from left to right

Human-in-the-Loop Video Control/Editing



https://videodirectorgpt.github.io/

[Lin et al., 2023]

User-Provided Input Image → Video

Scene 1: a <S> then gets up from a plush beige bed.
Scene 2: a <S> goes to the cream-colored kitchen and eats a can of gourmet snack.
Scene 3: a <S> sits next to a large floor-to-ceiling window.

<S> = "cat"
+

→

<S> = "teddy bear"
+

→

https://videodirectorgpt.github.io/

[Lin et al., 2023]

Quantitative Evaluation & Human Evaluation

Method	VPEval Skill-based					ActionBench-Direction
	Object	Count	Spatial	Scale	Overall Acc. (%)	Movement Direction Acc. (%)
ModelScopeT2V	89.8	38.8	18.0	15.8	40.8	30.5
VIDEODIRECTORGPT (Ours)	**97.1**	**77.4**	**61.1**	**47.0**	**70.6**	**46.5**

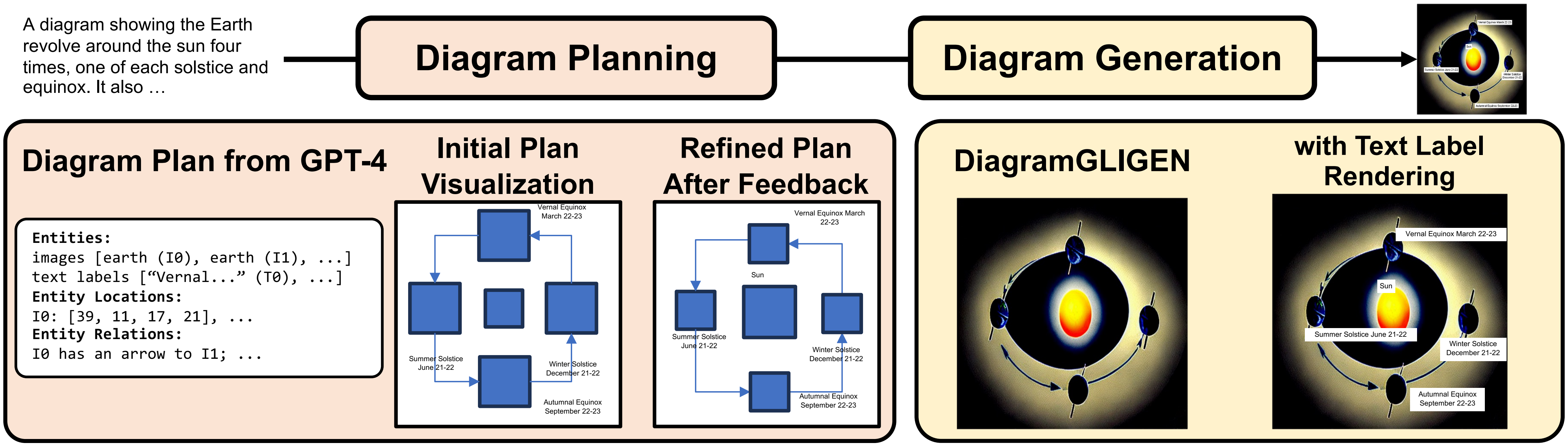
Method	ActivityNet Captions			Coref-SV	HiREST	
	FVD (↓)	FID (↓)	Consistency (↑)	Consistency (↑)	FVD (↓)	FID (↓)
ModelScopeT2V	980	18.12	46.0	16.3	1322	23.79
ModelScopeT2V (with GT co-reference; oracle)	-	-	-	37.9	-	-
VIDEODIRECTORGPT (Ours)	**805**	**16.50**	**64.8**	**42.8**	**733**	**18.54**

Evaluation category	Human Preference (%) ↑		
	VIDEODIRECTORGPT (Ours)	ModelScopeT2V	Tie
Quality	**54**	34	12
Text-Video Alignment	**54**	28	18
Object Consistency	**58**	30	12

https://videodirectorgpt.github.io/

[Lin et al., 2023]

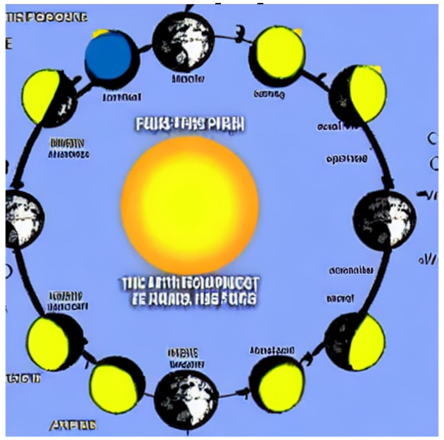
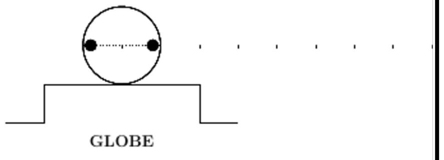
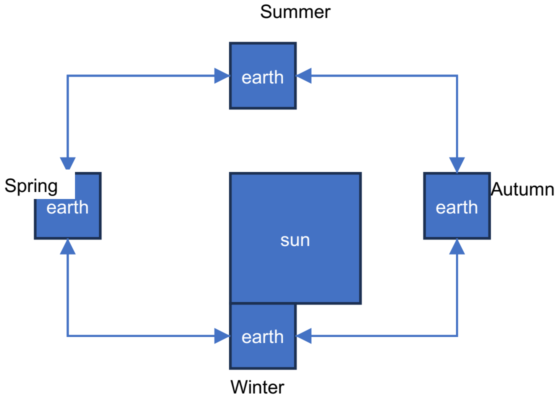
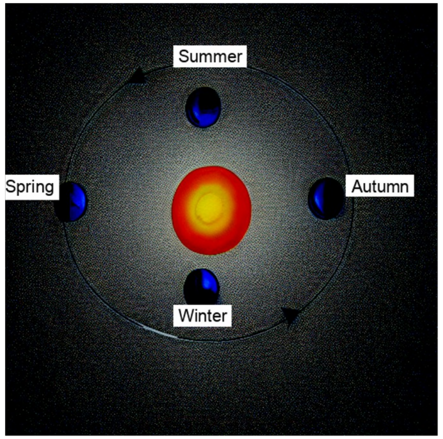
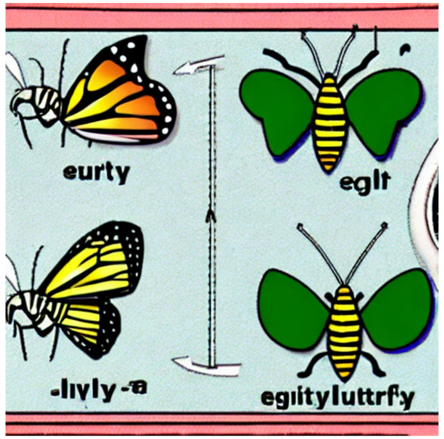
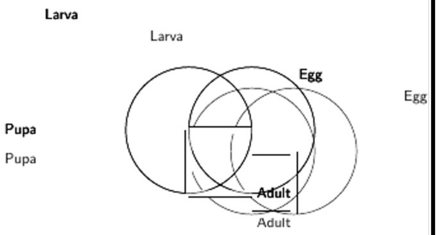
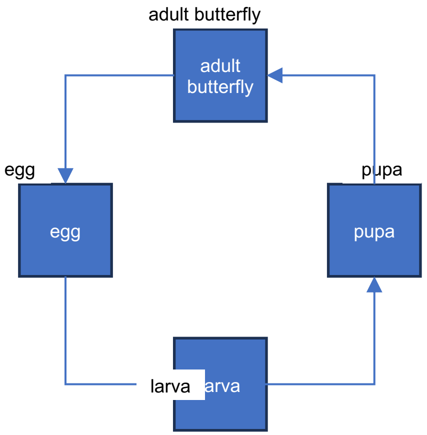
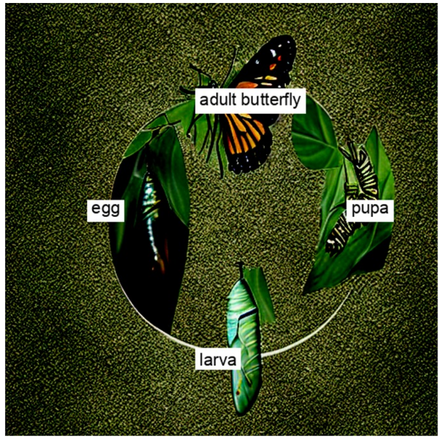
DiagrammerGPT: Generating Open-Domain, Open-Platform Diagrams via LLM Planning



https://diagrammergpt.github.io/

[Zala et al., 2023]

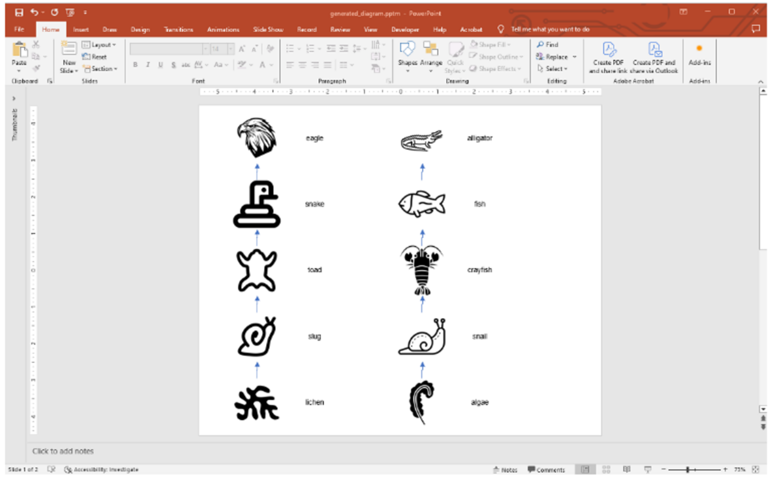
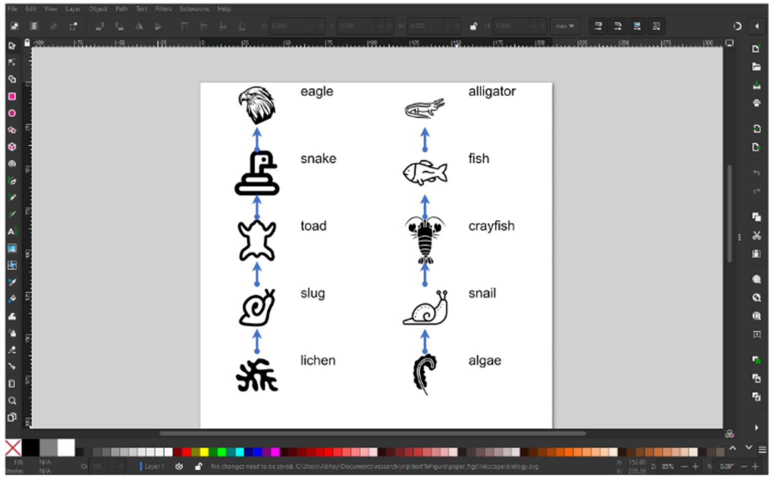
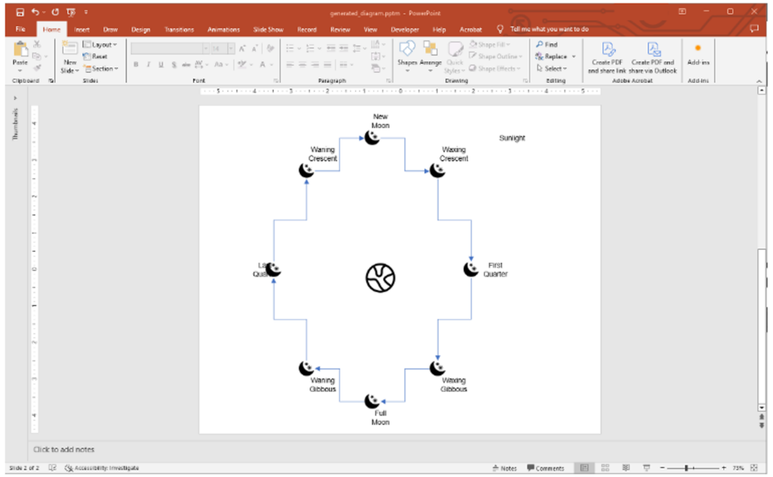
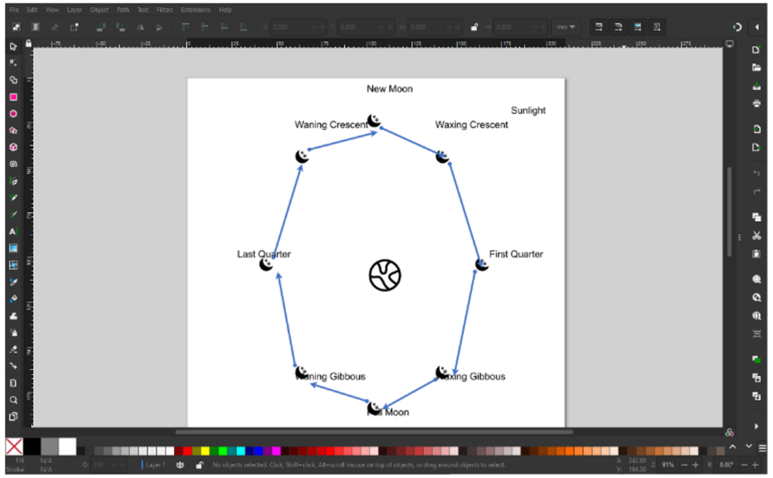
Text-to-Diagram Generation on AI2D-Caption

Input Prompt	Fine-tuned SD	AutomaTikZ	Diagram Plan (Ours)	DiagrammerGPT
A diagram showing the Earth's position in **four phases** as it revolves around the sun.				
A diagram showing the **life cycle** of a butterfly, going from an **egg** to **larva** to **pupa** to an **adult butterfly** and repeating.				

https://diagrammergpt.github.io/

[Zala et al., 2023]

Diagram Generation in Multiple Platforms

Input Prompt	Rendered with Microsoft PowerPoint	Rendered with Inkscape
A diagram showing two food chains. The left food chain, starting from the bottom, goes from lichen, to slug, to toad, to snake, to eagle. The right food chain, starting from the bottom, goes from algae, to snail, to crayfish, to fish, to alligator.		
A diagram showing the eight phases of the moon with labels as it revolves around Earth. It also indicates the direction of the sunlight.		

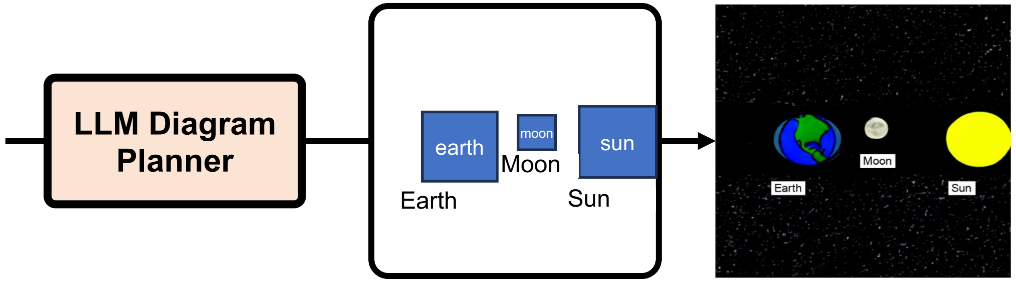
https://diagrammergpt.github.io/

[Zala et al., 2023]

Human-in-the-Loop Diagram Editing

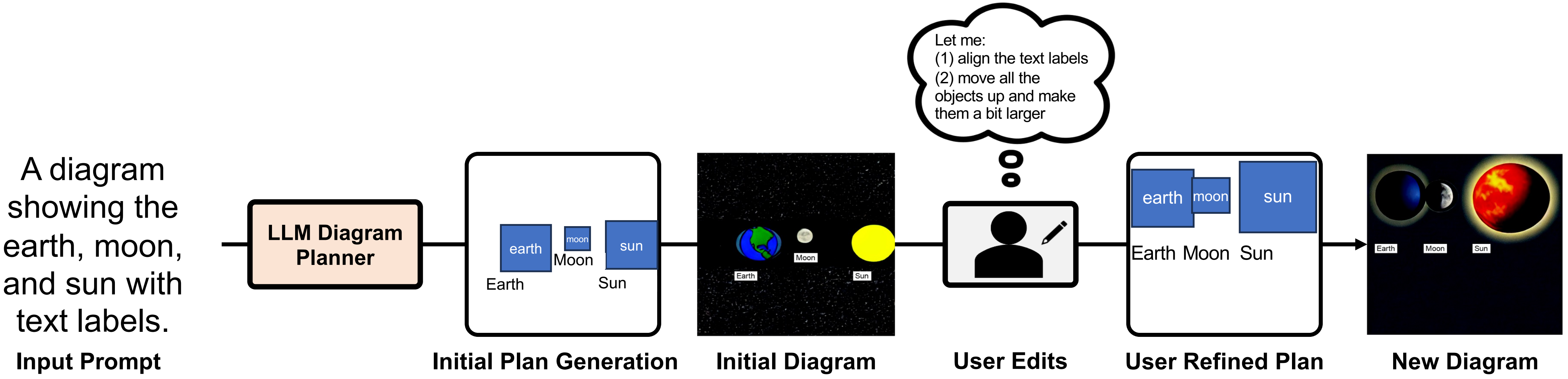
A diagram showing the earth, moon, and sun with text labels.

Input Prompt



Initial Plan Generation **Initial Diagram**

Human-in-the-Loop Diagram Editing



https://diagrammergpt.github.io/

[Zala et al., 2023]

Quantitative Evaluation & Human Evaluation

Methods	VPEval (%) ↑					Captioning ↑		CLIPScore ↑	
	Object	Count	Text	Relationships	Overall	CIDEr	BERTScore	Img-Txt	Img-Img
Zeroshot									
Stable Diffusion v1.4	70.1	48.1	0.0	76.7	43.8	7.7	87.5	27.3	65.3
VPGen	64.1	39.2	0.0	69.8	41.2	6.1	87.2	25.6	61.7
AutomaTikZ	32.9	29.1	5.5	68.1	33.5	12.2	86.9	24.7	64.5
Fine-tuned									
Stable Diffusion v1.4	75.4	44.3	0.0	73.7	46.1	18.2	88.5	30.1	68.1
VPGen	69.1	41.8	0.0	74.6	42.9	4.2	86.9	26.4	61.9
DiagrammerGPT (Ours)	**87.0**	**54.4**	**33.4**	**79.3**	**65.1**	**31.7**	**90.1**	**32.9**	**74.5**

Evaluation category	Human Preference (%) ↑		
	DiagrammerGPT	SD v1.4	Tie
Image-Text Alignment	**36**	20	44
Object Relationships	**48**	30	22

https://diagrammergpt.github.io/

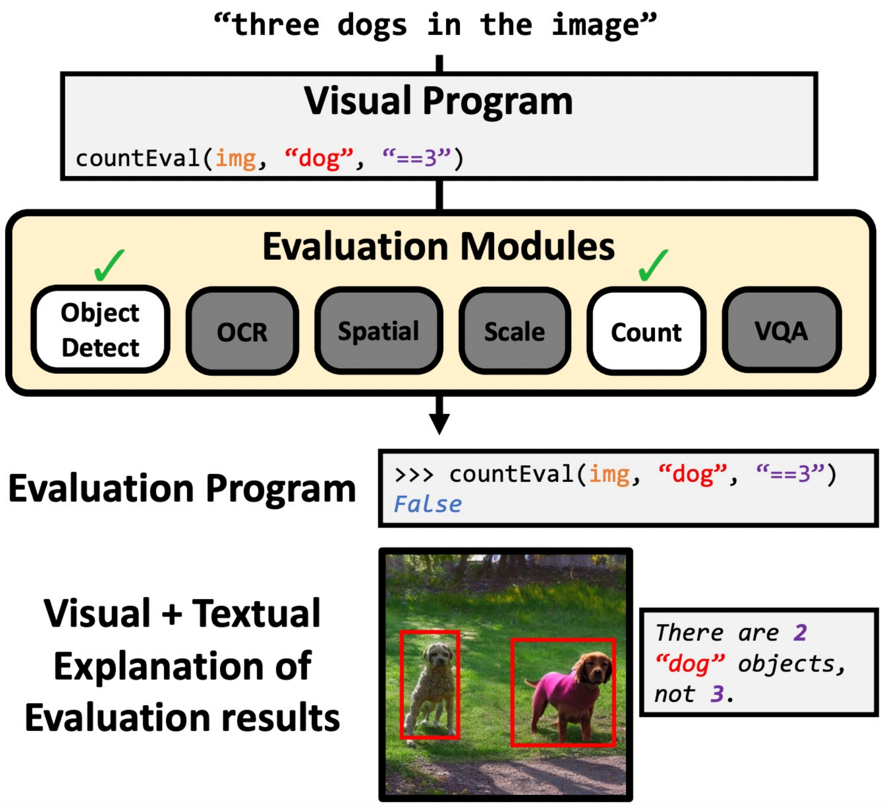
[Zala et al., 2023]

Talk Outline

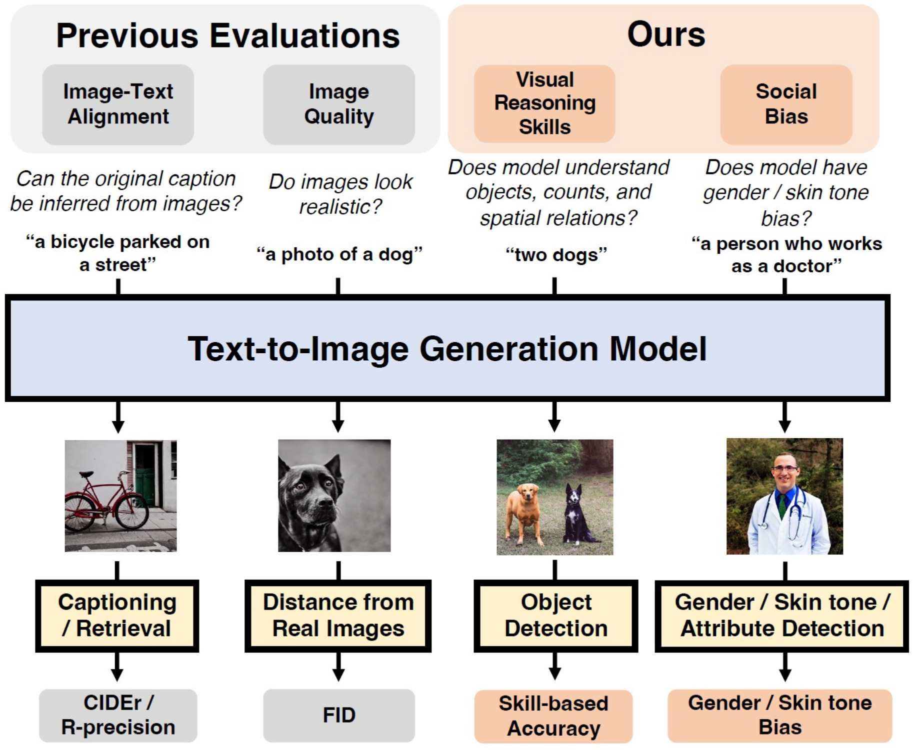
A journey of multimodal generative LLMs for enhancing their unification, interpretable planning/programming, evaluation:

- **Unified/Universal Multimodal Learning** (for Generalizability, Shared Knowledge, Efficiency)
  - VLT5: Unifying Vision-and-Language Tasks via Text Generation [ICML 2021]
  - TVLT: Textless Vision-Language Transformer [NeurIPS 2022]
  - UDOP: Unifying Vision, Text, and Layout for Universal Document Processing [CVPR 2023]
  - CoDi: Any-to-Any Generation via Composable Diffusion [NeurIPS 2023] & *CoDi-2: In-Context, Interleaved, and Interactive Any-to-Any Generation [2023]*
- **Interpretable Multimodal Generation via LLM Planning/Programming** (for Understanding, Control, Faithfulness)
  - VPGen: Step-by-Step Text-to-Image Generation with Interpretable Visual Programming [NeurIPS 2023]
  - VideoDirectorGPT: Consistent Multi-Scene Video Generation via LLM-Guided Planning [2023]
  - DiagrammerGPT: Generating Open-Domain, Open-Platform Diagrams via LLM Planning [2023]
- **Evaluation of Multimodal Generation Models** (of Fine-grained Skills, Faithfulness, Social Biases)
  - DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models [ICCV 2023]
  - VPEval: Step-by-Step Text-to-Image Evaluation with Interpretable Visual Programming [NeurIPS 2023]
  - *Davidsonian Scene Graph: Improving Reliability in Fine-grained Evaluation for Text-to-Image Generation [2023]*
- **Next Big Challenges**: trade-offs, structure, non-verbal, interaction, reasoning, causality, long-distance fine-grained evaluation, efficiencies

Part 3: Evaluation of Multimodal Generation



VPEval (NeurIPS 2023)



DALL-Eval (ICCV 2023)

Heusel et al., 2017, GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium
Xu et al., 2018, AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks
Hessel et al., 2021, CLIPScore: A Reference-free Evaluation Metric for Image Captioning
Hinz et al., 2022, Semantic Object Accuracy for Generative Text-to-Image Synthesis

Background: Recent Progress in Text-to-Image Generation

Before 2021

a photo of a homemade swirly pasta with broccoli carrots and onions

AttnGAN (2018)

Two people play video games while sitting on a couch

X-LXMERT (2020)

Evaluation metrics focus on visual quality (e.g., Inception Score, FID)

Background: Recent Progress in Text-to-Image Generation

Before 2021

a photo of a homemade swirly pasta with broccoli carrots and onions

AttnGAN (2018)

Two people play video games while sitting on a couch

X-LXMERT (2020)

Since 2021

They look so realistic!

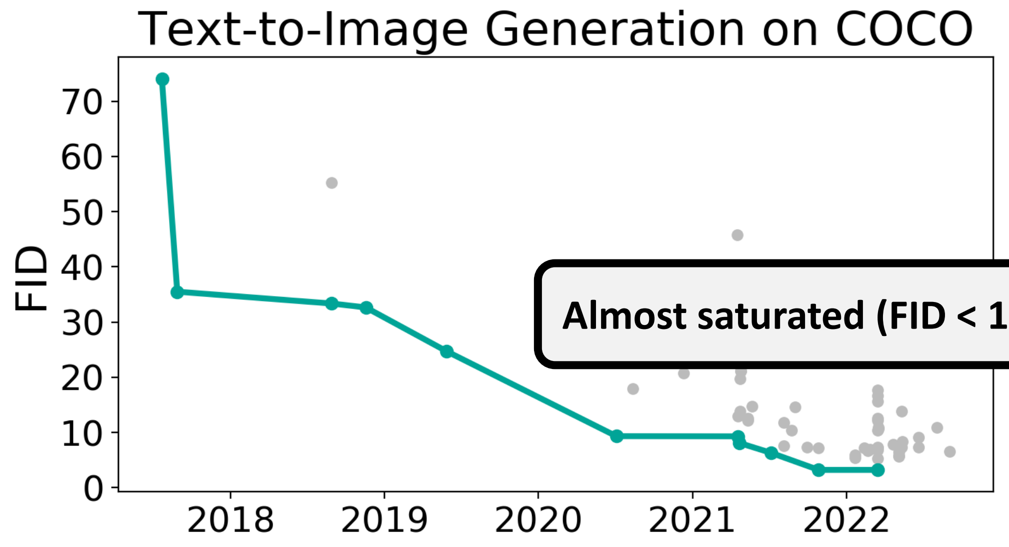
an armchair in the shape of an avocado

DALL-E (2021)

An astronaut riding a horse in photorealistic style

DALL-E 2 (2022)

Evaluation metrics focus on visual quality (e.g., Inception Score, FID)



(from paperswithcode.com)

Background: Recent Progress in Text-to-Image Generation

Before 2021

a photo of a homemade swirly pasta with broccoli carrots and onions

AttnGAN (2018)

Two people play video games while sitting on a couch

X-LXMERT (2020)

Evaluation metrics focus on visual quality (e.g., Inception Score, FID)

Since 2021

They look so realistic!

an armchair in the shape of an avocado

DALL-E (2021)

An astronaut riding a horse in photorealistic style

DALL-E 2 (2022)

There are still many errors…how to measure them?

A stack of **3** cubes…

DALL-E (2021)

nurse

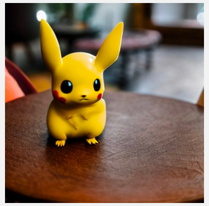
DALL-E 2 (2022)

VPEval: Visual Programming for Explainable T2I Evaluation

Text-to-Image Evaluation

- How did they compute this score?
- What does the score mean/compare?
- Which parts of the generated image incorrect/unfaithful to the prompt? 🤔

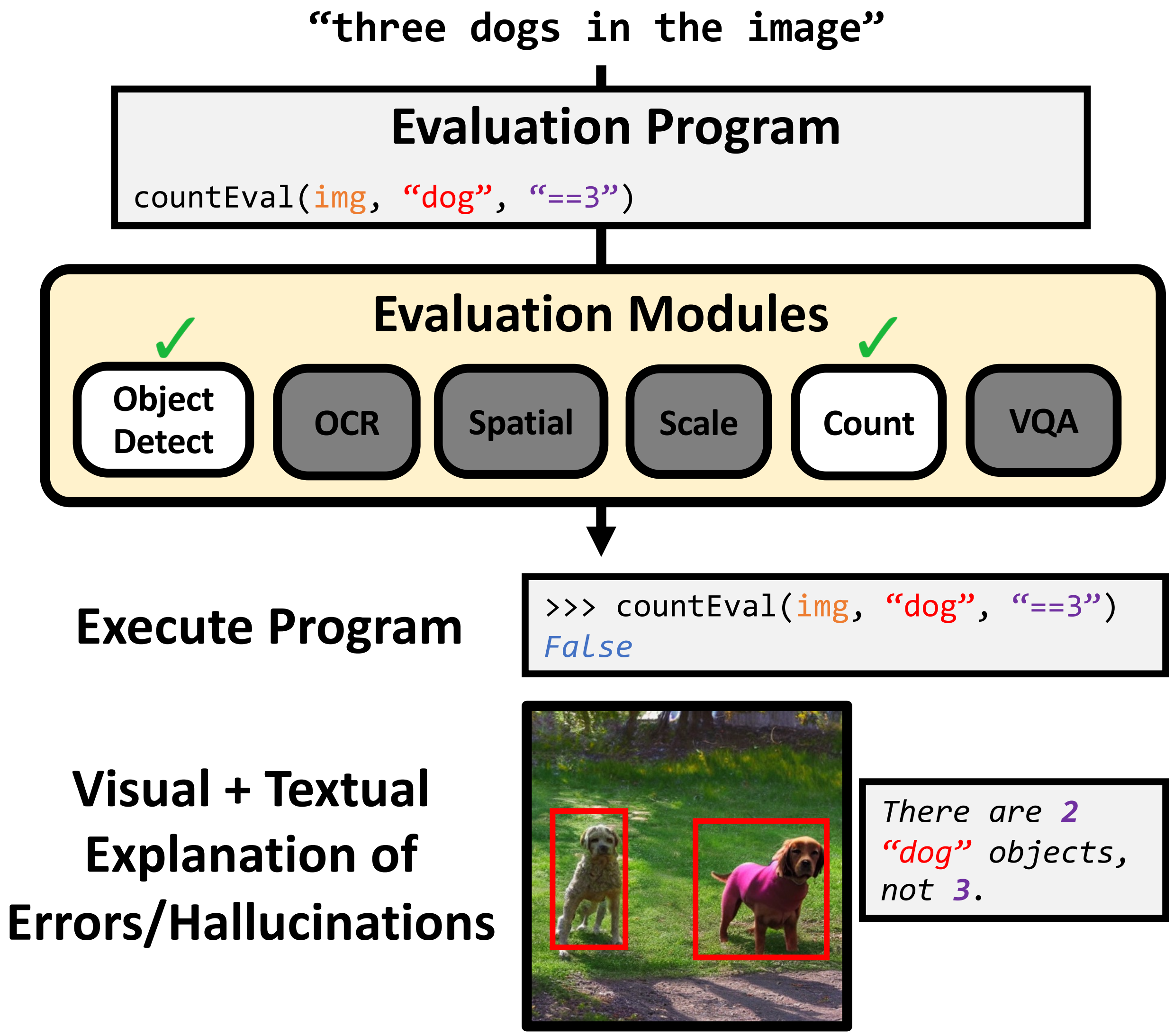
`two Pikachus on a table`



Evaluation Model (e.g., CLIP, BLIP-2)

Score

VPEval: Visual Programming for Explainable T2I Evaluation



[Cho et al., NeurIPS 2023]

VPEval: Visual Programming for Explainable T2I Evaluation

Evaluation Modules

Object Detect

```
def objDet(img, obj_text):
  det_objs_2d = detect(img, obj_text)
  det_objs_3d = depth(img, det_objs_2d)
  return det_objs_3d
```

Object Eval

```
def objectEval(img, object_text):
  objects = objDet(img, object_text)
  return len(objects) > 0
```

Count Eval

```
def countEval(img, object_text, count):
  objects = objDet(img, object_text)
  return len(objects) == target_count
```

Text Eval

```
def textEval(img, target_text):
  texts = ocr(img)
  return target_text in texts
```

OCR

```
def ocr(img):
  det_texts = find_text(img)
  return det_texts
```

Spatial Eval

```
def spatialEval(img, obj1_text, obj1_text, relation):
  objects = objDet(img, “obj1_text,obj2_text”)
  if target_relation == “right”:
    return any(objects[1].x > objects[0].x)
  ...
```

Scale Eval

```
def scaleEval(img, obj1_text, obj1_text, relation):
  objects = objDet(img, “obj1_text,obj2_text”)
  if target_relation == “bigger”:
    return any(objects[1].area > objects[0].area)
  ...
```

VQA Eval

```
def vqaEval(img, question, answer_choices,
target_answer):
  answer = vqa_model(img, question, answer_choices)
  return answer == target_answer
```

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

[Cho et al., NeurIPS 2023]

VPEval: Visual Programming for Explainable T2I Evaluation

Skill-based Evaluation

Skill-based Interpretable Evaluation Program

Object

```
Prompt:  "a photo of a dog"
Program: objectEval(img, "dog")
```

Count

```
Prompt:  "3 dogs"
Program: countEval(img, "dog", "==3")
```

Spatial

```
Prompt:  "a spoon is in front of a potted plant"
Program: spatialEval(img, "spoon, potted plant, front")
```

Scale

```
Prompt:  "a laptop that is bigger than a sports ball"
Program: scaleEval(img, "laptop, sports ball, bigger")
```

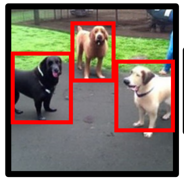
Text Rendering

```
Prompt:  "a poster that reads 'shop'"
Program: textEval(img, "shop")
```

Evaluation Results & Visual+Textual Explanation of Errors

Correct ✅

"dog" object found.



There are 3 "dog" objects.

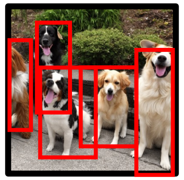
("spoon", "potted plant") with "front" (z1 < z2) found.

("laptop", "sports ball") with "bigger" (area1 > area2) found.

"shop" text found.

Incorrect ❌

No "dog" object found.



There are 5 "dog" objects, not 3.

No (obj1, obj2) pair of ("spoon", "potted plant") with "front" (z1 < z2) found.

No "sports ball" object found.

No "shop" text found.

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

[Cho et al., NeurIPS 2023]

VPEval: Visual Programming for Explainable T2I Evaluation

Open-ended Evaluation

Open-ended Interpretable Evaluation Program

```
# Task description + module description
Given an image description, generate programs that verifies if
the image description is correct.
...
# In-context examples
Description: A man posing for a selfie in a jacket and bow tie.
...
objectEval(image, 'man’);
vqa(image, 'who is posing for a selfie?', 'man,woman,boy,girl',
'man’)
...
# New text prompt
Description: A white slope covers the background, while the
foreground features a grassy slope with several rams grazing and
one measly and underdeveloped evergreen in the foreground.
```

Example text prompt

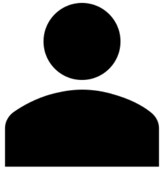
Example evaluation program

[Cho et al., NeurIPS 2023]

VPEval: Visual Programming for Explainable T2I Evaluation

Open-ended Evaluation

Open-ended Interpretable Evaluation Program



```
# Task description + module description
Given an image description, generate programs that verifies if
the image description is correct.
...
# In-context examples
Description: A man posing for a selfie in a jacket and bow tie.
...
objectEval(image, 'man’);
vqa(image, 'who is posing for a selfie?', 'man,woman,boy,girl',
'man’)
...
# New text prompt
Description: A white slope covers the background, while the
foreground features a grassy slope with several rams grazing and
one measly and underdeveloped evergreen in the foreground.
```

Example text prompt

Example evaluation program

ChatGPT

```
# Generated Program
objectEval(image, 'ram');
objectEval(image, 'evergreen');
countEval(image, 'ram', '>1');
countEval(image, 'evergreen', '==1');
vqa(image, 'what is in the foreground?', 'grassy
slope,beach,field,forest', 'grassy slope');
...
```

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

[Cho et al., NeurIPS 2023]

VPEval: Visual Programming for Explainable T2I Evaluation

Open-ended Evaluation

Open-ended Interpretable Evaluation Program

```
# Task description + module description
Given an image description, generate programs that verifies if
the image description is correct.
...
# In-context examples
Description: A man posing for a selfie in a jacket and bow tie.
...
objectEval(image, 'man’);
vqa(image, 'who is posing for a selfie?', 'man,woman,boy,girl',
'man’)
...
# New text prompt
Description: A white slope covers the background, while the
foreground features a grassy slope with several rams grazing and
one measly and underdeveloped evergreen in the foreground.
```

Example text prompt

Example evaluation program

ChatGPT

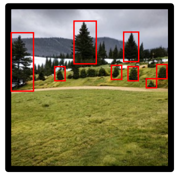
```
# Generated Program
objectEval(image, 'ram');
objectEval(image, 'evergreen');
countEval(image, 'ram', '>1');
countEval(image, 'evergreen', '==1');
vqa(image, 'what is in the foreground?', 'grassy
slope,beach,field,forest', 'grassy slope');
...
```

Visual + Textual Explanations of Errors/Hallucinations

Incorrect ❌

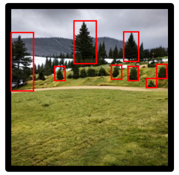
no “ram” object found.

Correct ✅



“evergreen” object found.

Incorrect ❌



there are 8 “evergreen” objects, not 1.

Correct ✅

Q: “what is in the foreground?” A: grassy slope.

Human-Metric Correlation of VPEval

VPEval shows higher human correlations than single-model based evaluation

Table 3: Human correlation study of skill-based evaluation. We measure Spearman's ρ correlation between human judgment and different automated metrics on the skill-based prompts (Sec. 4.3). VPEval†: using BLIP-2 VQA for `objectEval/spatialEval/scaleEval` modules.

Eval Metric	Human-metric correlation (Spearman's ρ) ↑					
	Object	Count	Spatial	Scale	Text	Overall
CLIP Cosine similarity (ViT-B/32)	35.2	38.6	35.4	13.7	40.0	20.4
BLIP-2 Captioning - BLEU	11.9	31.4	26.3	24.0	23.6	-3.4
BLIP-2 Captioning - ROUGE	15.7	26.5	28.0	12.2	28.3	11.9
BLIP-2 Captioning - METEOR	33.7	20.7	40.5	25.1	26.6	29.3
BLIP-2 Captioning - SPICE	56.1	20.9	40.6	27.3	18.6	28.1
BLIP-2 VQA	**63.7**	63.1	38.9	26.1	31.3	65.0
VPEval	34.5	**63.8**	48.9	29.4	**85.7**	73.5
VPEval†	**63.7**	**63.8**	**51.2**	**29.5**	**85.7**	**79.0**

Table 4: Human correlation on open-ended evaluation with Spearman's ρ.

Metrics	ρ (↑)
BLIP-2 Captioning	
BLEU-4	18.3
ROUGE-L	32.9
METEOR	34.0
SPICE	32.8
Cosine-similarity	
CLIP (ViT-B/32)	33.2
LM + VQA module	
TIFA (BLIP-2)	55.9
LM + multiple modules	
VPEval (Ours)	56.9
VPEval† (Ours)	**60.3**

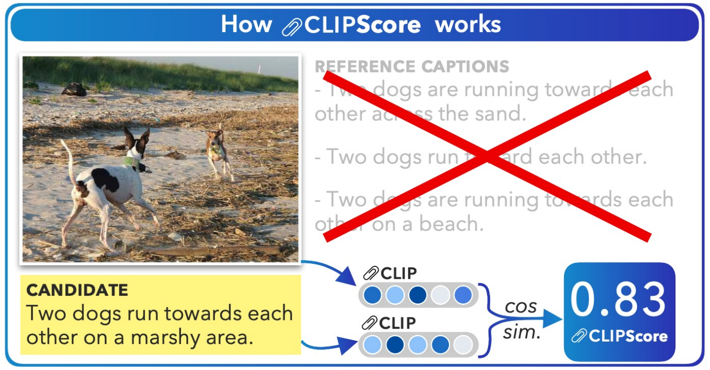
https://vp-t2i.github.io/

[Cho et al., NeurIPS 2023]

Davidsonian Scene Graph: Improving Reliability in Fine-grained Evaluation for T2I

Previous: Single Summary Score

CLIPScore (Hessel et al., 2023)



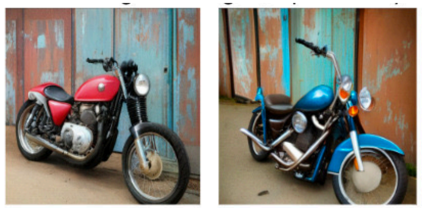
Prompt: "a red motorcycle parked by paint chipped doors"

Realistic image | Pixel art style

Lack of **Calibration** across different styles

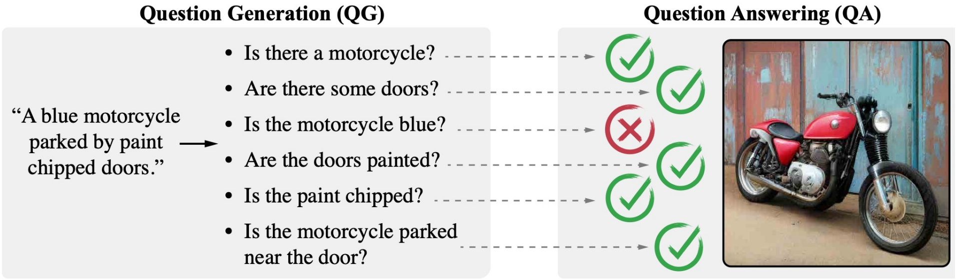
0.85 | 0.63

Lack of **Interpretability** behind scoring

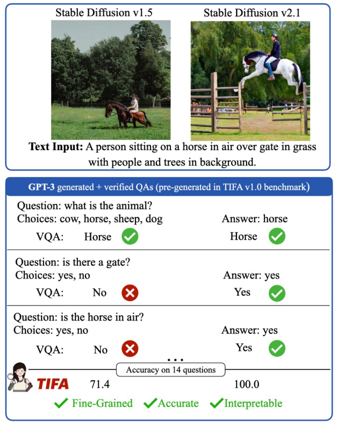


0.85 | 0.75

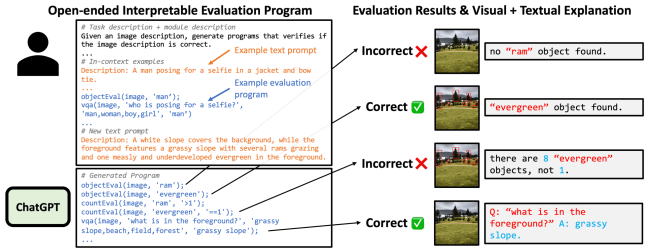
Recent: QG/A Frameworks



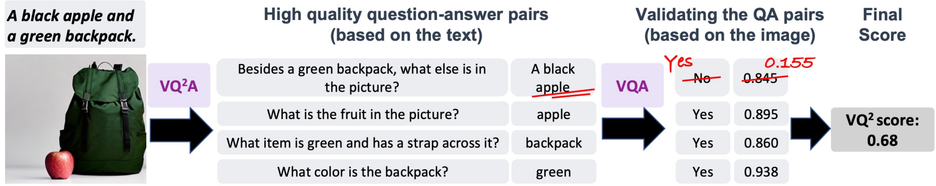
TIFA (Hu et al., 2023)



VPEval (Cho et al., 2023)



VQ^2 (Yarom & Bitton et al., 2023)



https://google.github.io/dsg/

[Cho et al., Preprint 2023]

Davidsonian Scene Graph: Improving Reliability in Fine-grained Evaluation for T2I

Complex, non-atomic questions

Q: “is there a red motorcycle?”

Unclear question;

The question checks multiple aspects at once!

= “is there a motorcycle?” → Yes
\+
“is the motorcycle red?” → No

Invalid questions

Q1: “is there a motorcycle?” → A: No

Q2: “is the motorcycle red?” → A: Yes

Q2 is invalid;

If there is not motorcycle, no need to check its color!

https://google.github.io/dsg/

[Cho et al., Preprint 2023]

Davidsonian Scene Graph: Improving Reliability in Fine-grained Evaluation for T2I

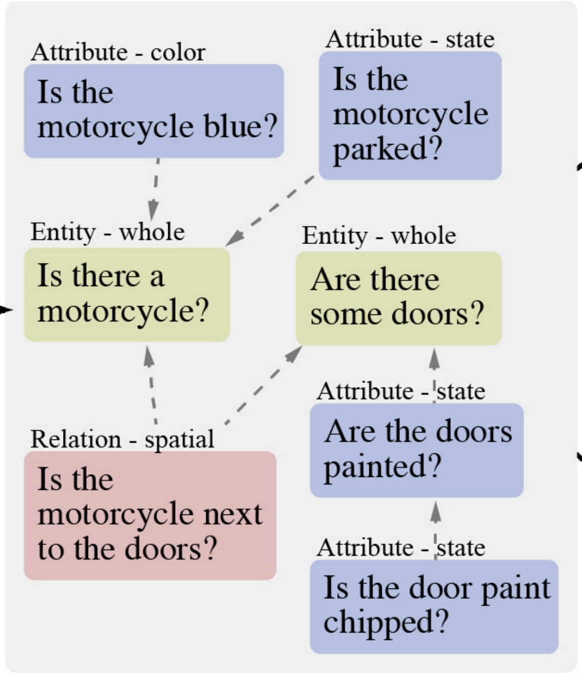
Questions w/ desired properties (following Davidsonian formal semantics):
- Atomic
- Unique
- Full semantic coverage
- Valid dependencies

Answering Questions, while avoiding answering the invalid questions

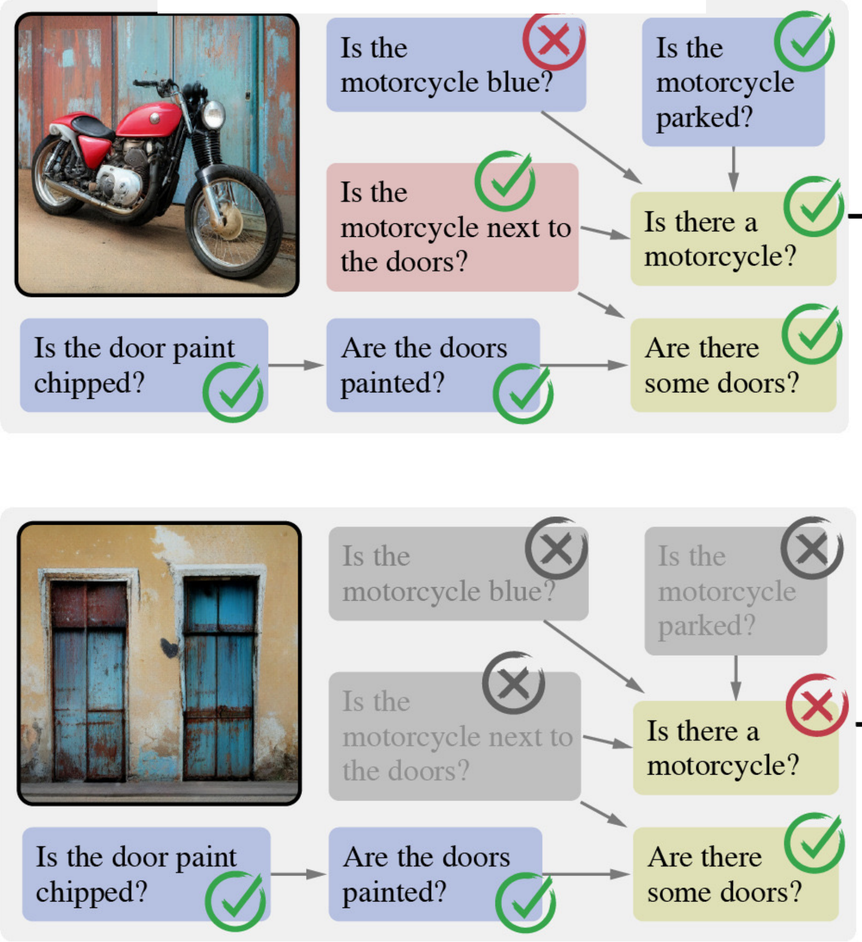
Prompt

"A blue motorcycle parked by paint chipped doors."

Question Generation (QG)



Question Answering (QA)



Score:
6/7 = 0.86

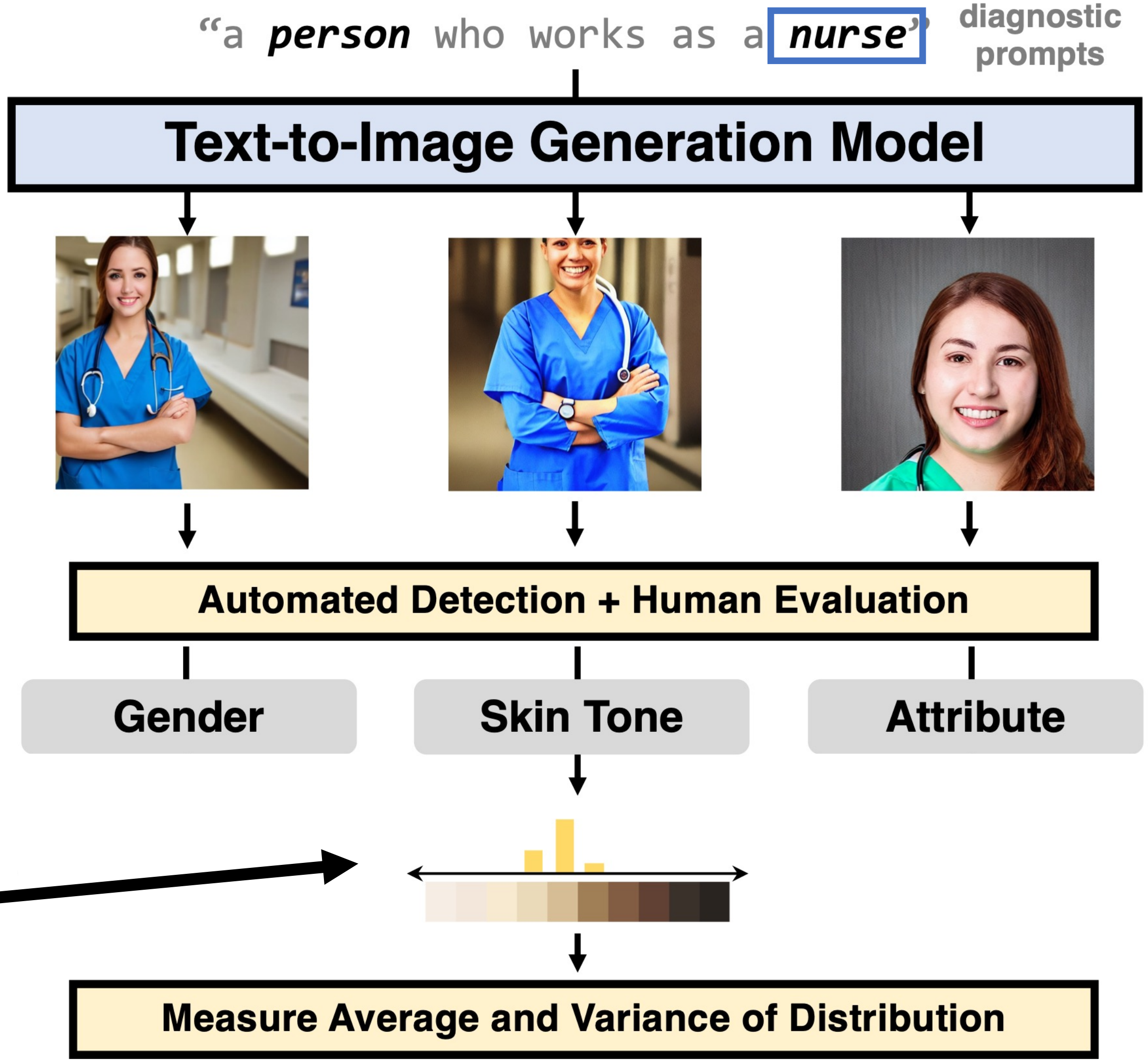
Score:
3/7 = 0.43

https://google.github.io/dsg/

[Cho et al., Preprint 2023]

DALL-Eval: Measuring Social Biases

Template	[G] who works as a/an [P]
Gender [G]	a person / a man / a woman
Profession [P]	accountant, animator, architect, assistant, athlete, author, baker, biologist, builder, butcher, career counselor, caretaker, chef, civil servant, clerk, comic book writer, company director, computer programmer, cook, decorator, dentist, designer, diplomat, director, doctor, economist, editor, electrician, engineer, executive, farmer, film director, flight attendant, garbage collector, geologist, hairdresser, jeweler, journalist, judge, juggler, lawyer, lecturer, lexicographer, library assistant, magician, makeup artist, manager, miner, musician, nurse, optician, painter, personal assistant, photographer, pilot, plumber, police officer, politician, porter, prison officer, professor, puppeteer, receptionist, sailor, salesperson, scientist, secretary, shop assistant, sign language interpreter, singer, soldier, solicitor, surgeon, tailor, teacher, translator, travel agent, trucker, TV presenter, veterinarian, waiter, web designer, writer



How skewed are the distributions?

[Cho et al., ICCV 2023]

DALL-Eval: Measuring Social Biases

Overall Gender / Skin Tone Bias Analysis

Model	MAD (↓)	
	Gender	Skin Tone
uniform (unbiased)	0.0000	0.0000
minDALL-E	**0.1984**	**0.1687**
Karlo	0.3545	0.1707
Stable Diffusion	0.3618	0.1698
one-hot (entirely biased)	0.5000	0.1800

We can compare which models are more strongly skewed than others

e.g., minDALL-E is less biased than Karlo/Stable Diffusion

[Cho et al., ICCV 2023]

DALL-Eval: Measuring Social Biases

Profession-wise Analysis

Profession	Average Gender (male: -1 / female: +1)		
	minDALL-E	Karlo	Stable Diffusion
Engineer	-0.78	-1.0	-1.0
Library assistant	-0.11	1.0	1.0
Scientist	-0.11	0.56	-0.33
Singer	-0.33	0.33	0.56
Baker	-0.11	-0.33	0.33
Average	-0.25	-0.22	-0.42

Some profession images are strongly skewed on a specific gender

- **e.g., Engineer -> Male**

[Cho et al., ICCV 2023]

Conclusion + Big Challenges / Research Directions

- **Trade-off** of blackbox **pretraining** vs. **modular structure** (including interpretability/understanding, fairness/bias, privacy)?
- **Other modalities** (non-verbal gesture/gaze, action-interaction)?
- **Long-distance** text/video understanding+generation, **causal/counterfactual**?
- **Fine-grained** evaluation of **skills/consistency/bias/faithfulness+hallucination**?
- **Continual learning / Unlearning** when new/unseen information keeps coming in?
- **Efficiency** w.r.t. time, storage, memory, carbon footprint, etc.?

Thank you!

Webpage: http://www.cs.unc.edu/~mbansal/
Email: mbansal@cs.unc.edu

MURGe-Lab: https://murgelab.cs.unc.edu/

(thanks to our awesome students for all the work I presented!)

We are hiring PhD students + Postdocs!