Multimodal Generative Models: Unification, Planning Agents, Evaluation

Mohit Bansal

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Talk Outline

A journey of multimodal generative models 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 [CVPR 2024]
- **Interpretable Multimodal Generation via LLM Planning/Programming Agents** (for Understanding, Control, Faithfulness, OOD)
  - VPGen: Step-by-Step Text-to-Image Generation with Interpretable Visual Programming [NeurIPS 2023]
  - VideoDirectorGPT: Consistent Multi-Scene Video Generation via LLM-Guided Planning [COLM 2024]
  - DiagrammerGPT: Generating Diagrams via LLM Planning [COLM 2024]; EnvGen: Adapting Environments via LLMs for Training Embodied Agents [COLM 2024]
- **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 [ICLR 2024]
- **Next Big Challenges**: trade-offs, structure, non-verbal, interaction, reasoning, causality, long-distance fine-grained evaluation, efficiencies

Talk Outline

A journey of multimodal generative models 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 [CVPR 2024]
- **Interpretable Multimodal Generation via LLM Planning/Programming Agents** (for Understanding, Control, Faithfulness, OOD)
  - VPGen: Step-by-Step Text-to-Image Generation with Interpretable Visual Programming [NeurIPS 2023]
  - VideoDirectorGPT: Consistent Multi-Scene Video Generation via LLM-Guided Planning [COLM 2024]
  - DiagrammerGPT: Generating Diagrams via LLM Planning [COLM 2024]; EnvGen: Adapting Environments via LLMs for Training Embodied Agents [COLM 2024]
- **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 [ICLR 2024]
- **Next Big Challenges**: trade-offs, structure, non-verbal, interaction, reasoning, causality, long-distance fine-grained evaluation, efficiencies

Language: Pre-training → Fine-tuning

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

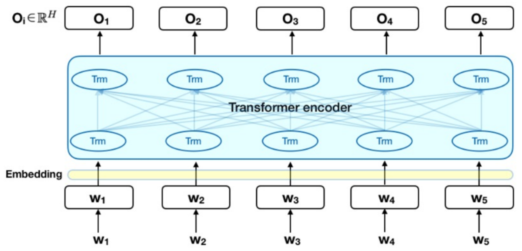
Language Pre-training:

Text in Wikipedia

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

Language Model

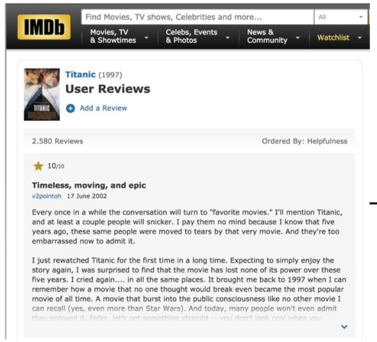
[Peters et al., NAACL 2018], [Devlin et al., NAACL 2019]



Transformer

[Vaswani, NeurIPS 2017]

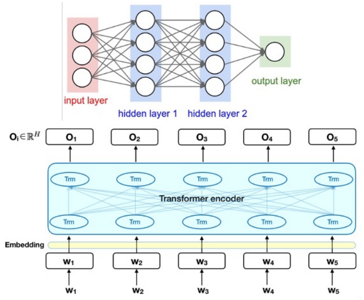
Language Fine-tuning



Movie Review [Maas et al., ACL 2011]

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

Sentiment Analysis



Transformer + Linear Layers

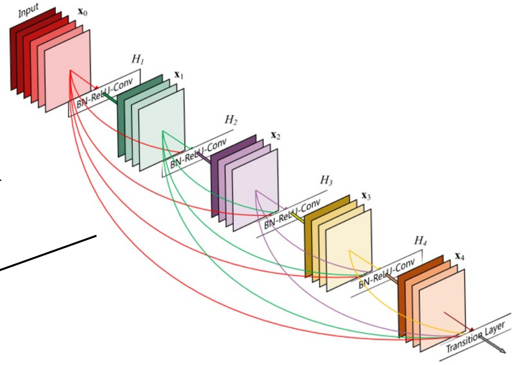
Vision: Pre-training → Fine-tuning

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 →



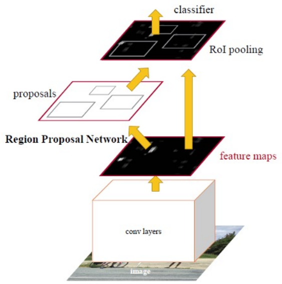
DenseNet
[Huang, CVPR 2017]

Visual Fine-tuning:

COCO Common Objects in Context

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

→ Object Detection →



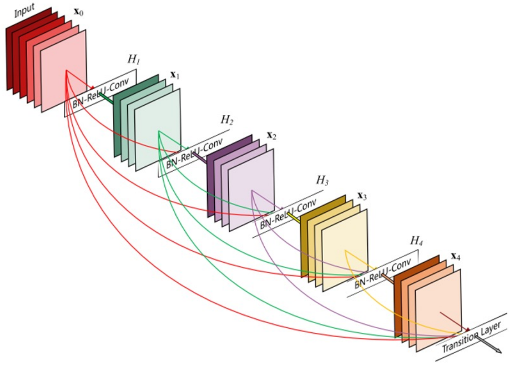
Faster RCNN
[Ren, NeurIPS 2015]

Pre-training of Single Modality Tasks

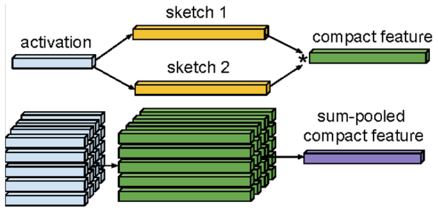
Limitation: Single-modality pre-trained models are not aware of the interactions between vision and language

Visual Pre-training:

Image Classification



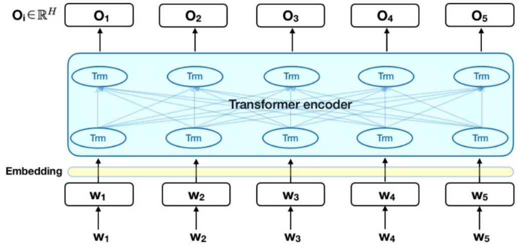
Multimodal Fusion Layers



Visual Question Answering, Navigation, Grounding, ...

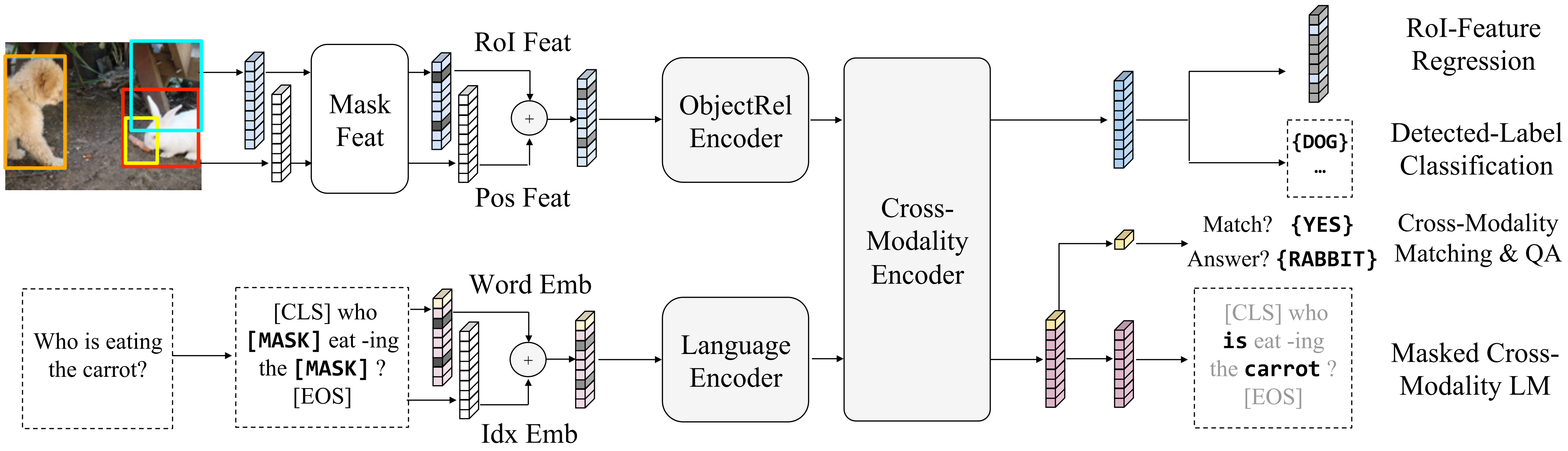
Language Pre-training:

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.

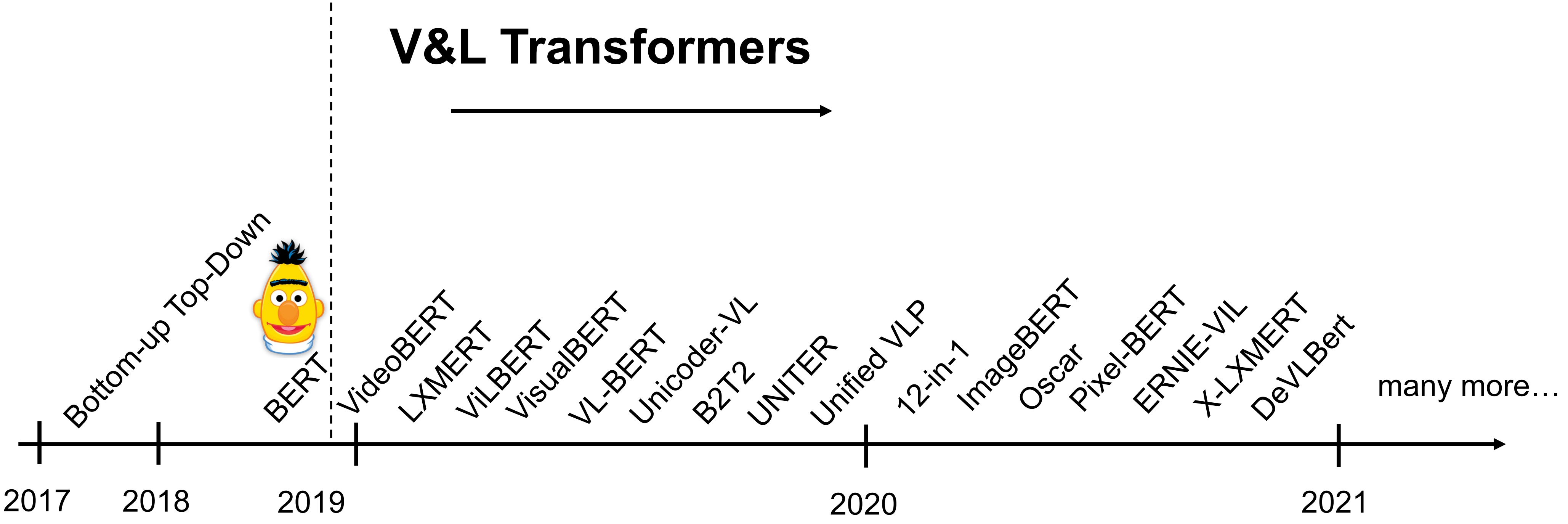


- Achieved big gains + sota on several VL tasks such as VQA, GQA, NLVR2, VizWiz, etc.

[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.!

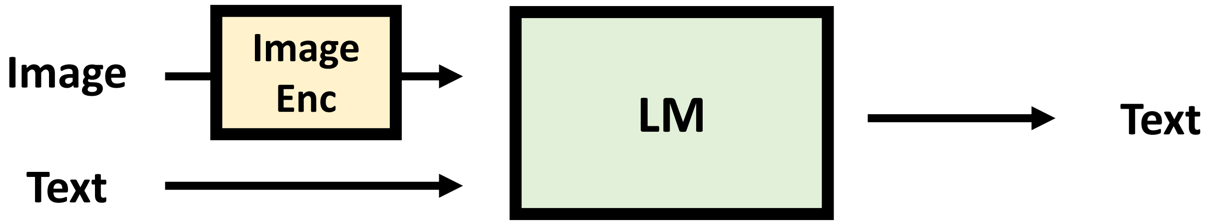


[Sun et al., 2019; Tan and Bansal, 2019; Lu et al., 2019; Li et al., 2019; Su et al., 2019; Chen et al., 2020; Zhou et al., 2020; Li et al., 2020; inter alia]

Part 1: Unified/Universal Multimodal Learning

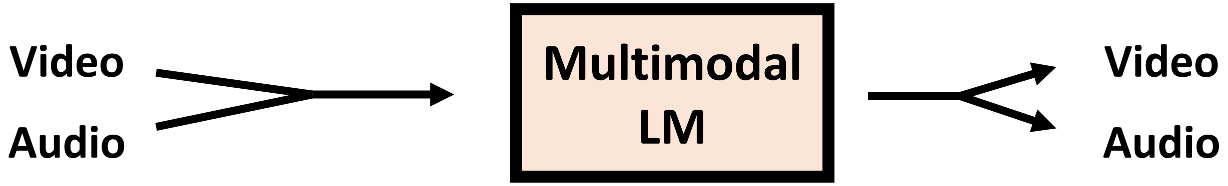
VL-T5 (ICML 2021)

all multimodal tasks via text generation



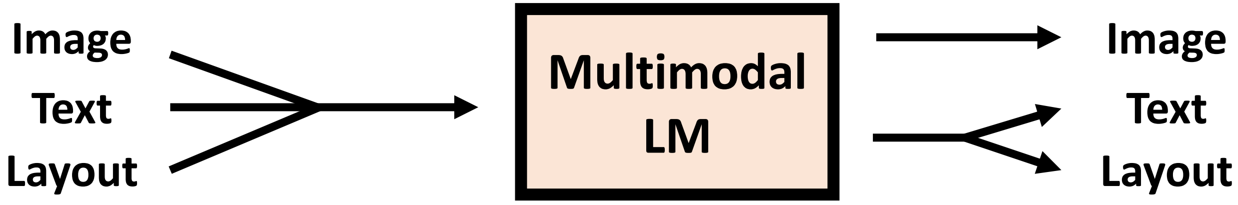
TVLT (NeurIPS 2022)

video modeling without text (audio as images)



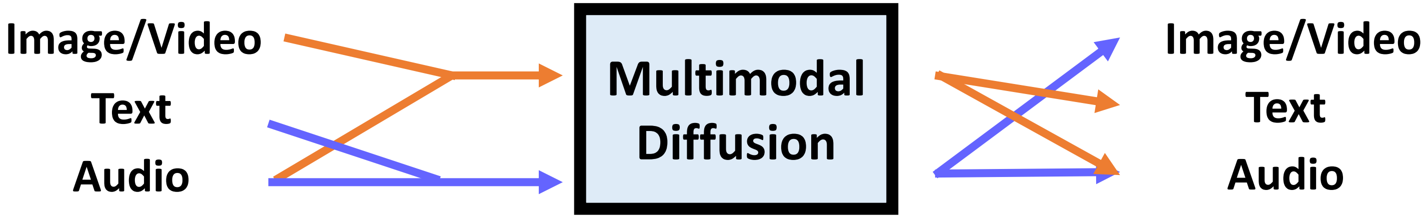
UDOP (CVPR 2023)

document image/text/layout with single architecture



CoDi (NeurIPS 2023)

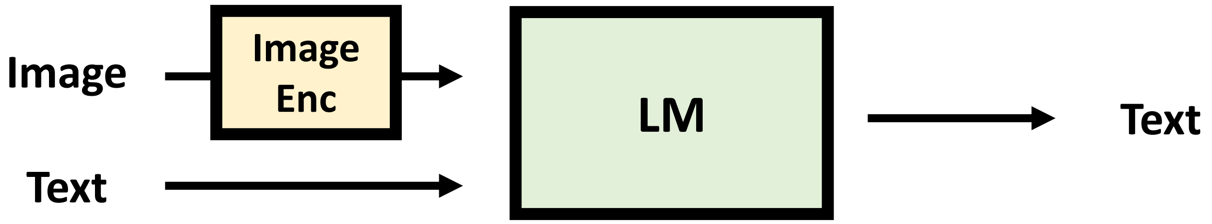
generating any-to-any input-output modality combination



Part 1: Unified/Universal Multimodal Learning

VL-T5 (ICML 2021)

all multimodal tasks via text generation



TVLT (NeurIPS 2022)

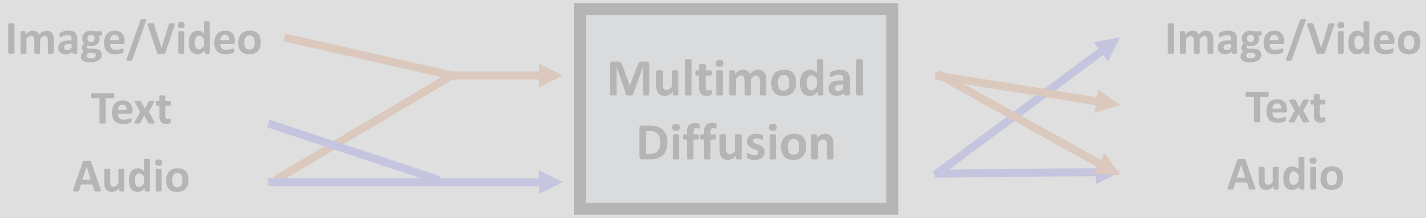
video modeling without text (audio as images)

UDOP (CVPR 2023)

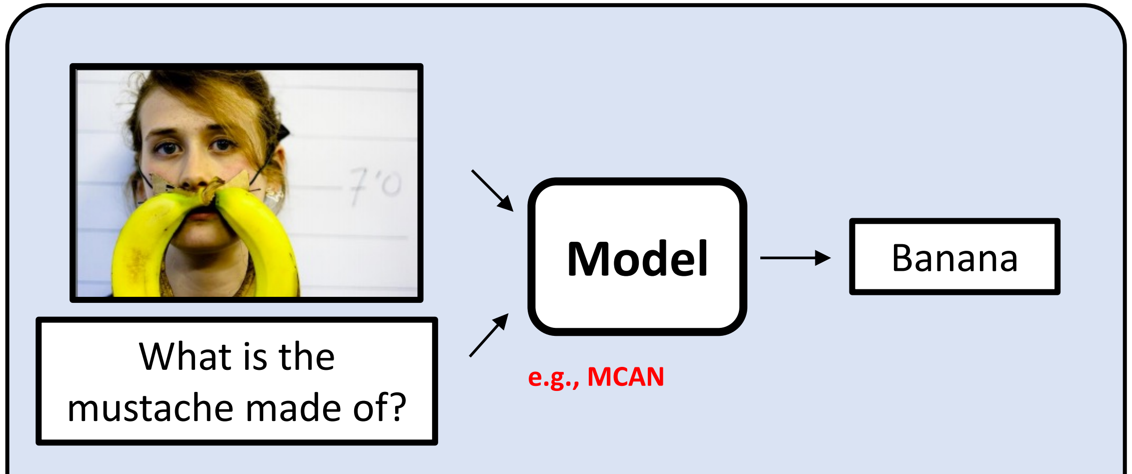
document image/text/layout with single architecture

CoDi (NeurIPS 2023)

generating any-to-any input-output modality combination



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



Visual Question Answering

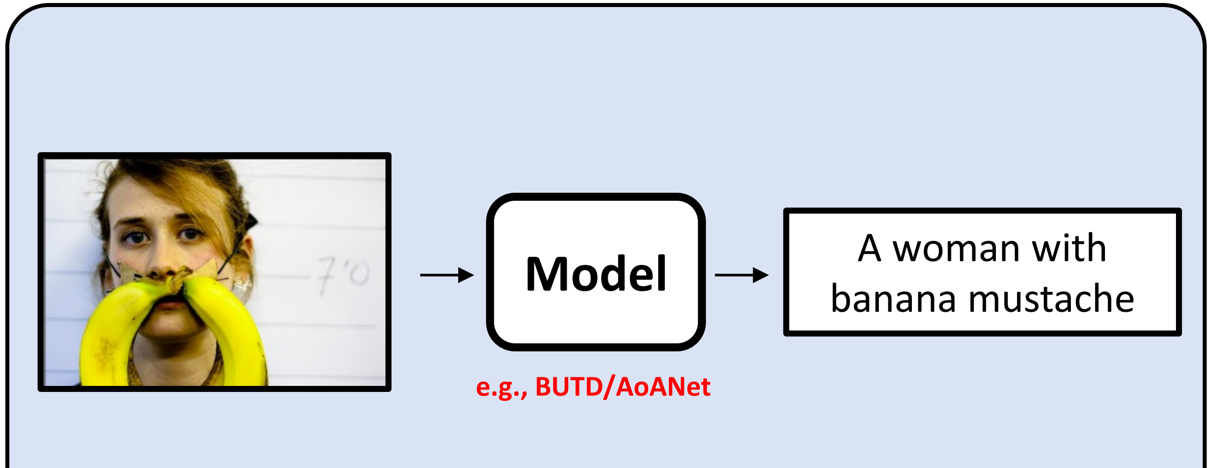
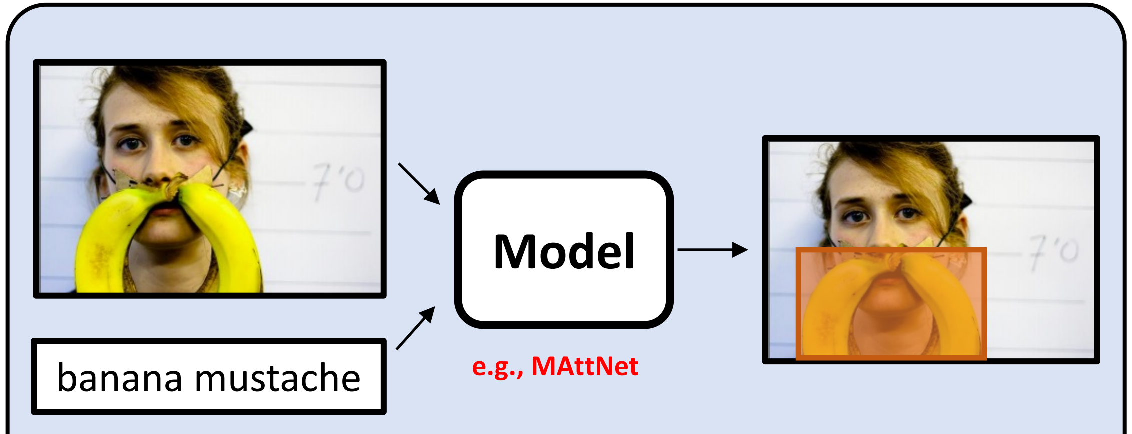
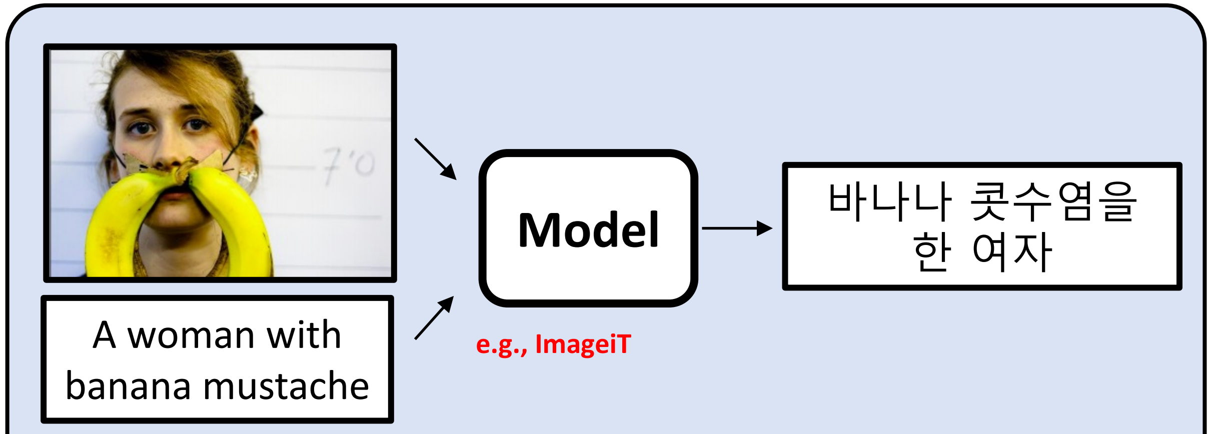


Image Captioning



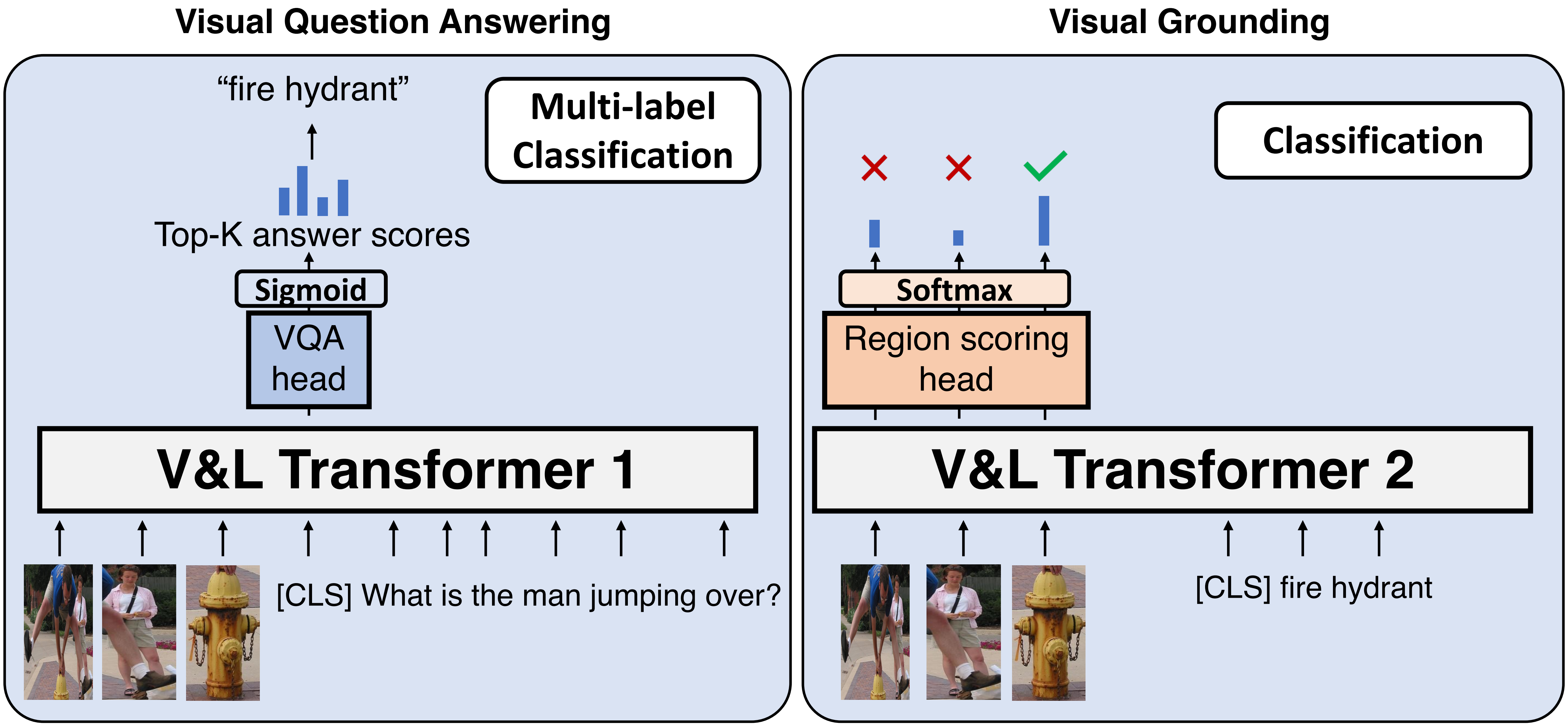
Visual Grounding



Multimodal Machine Translation (En-Kr)

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

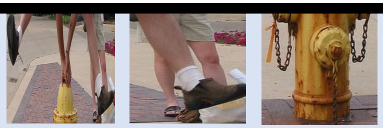


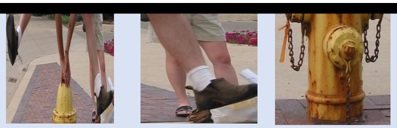
Task-specific Architectures / Objectives

Visual Question Answering

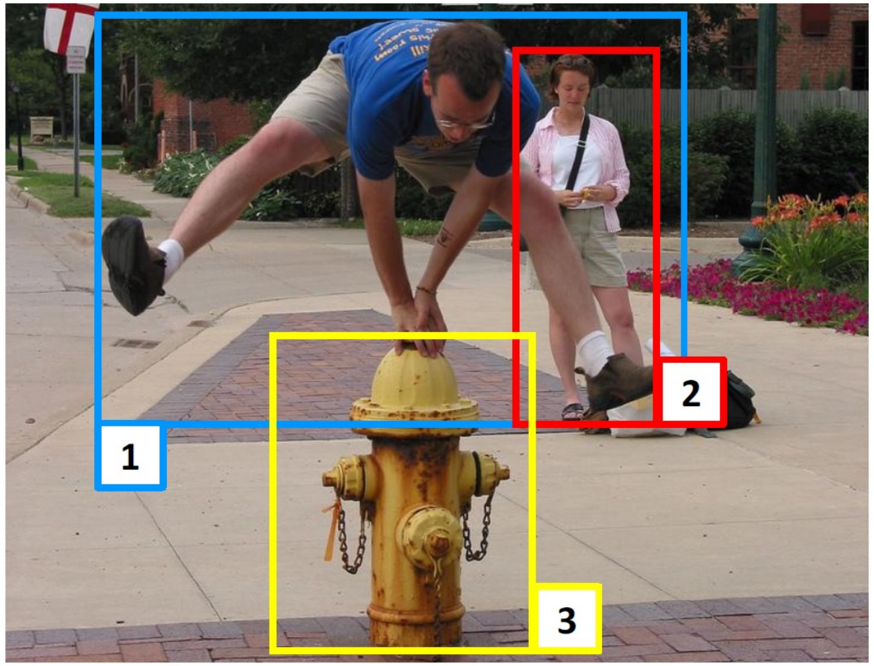
Visual Grounding

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





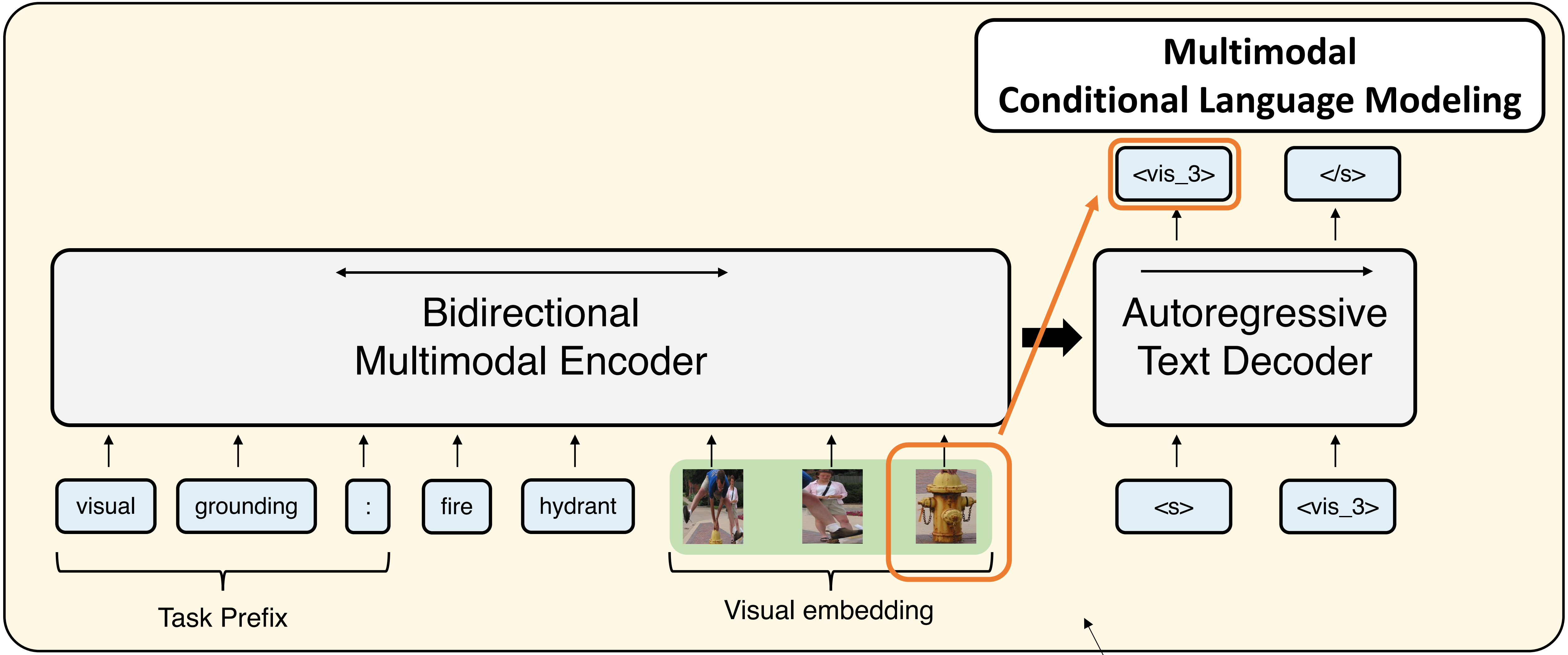
VL-T5: Many Multimodal Tasks as Text Generation



	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]

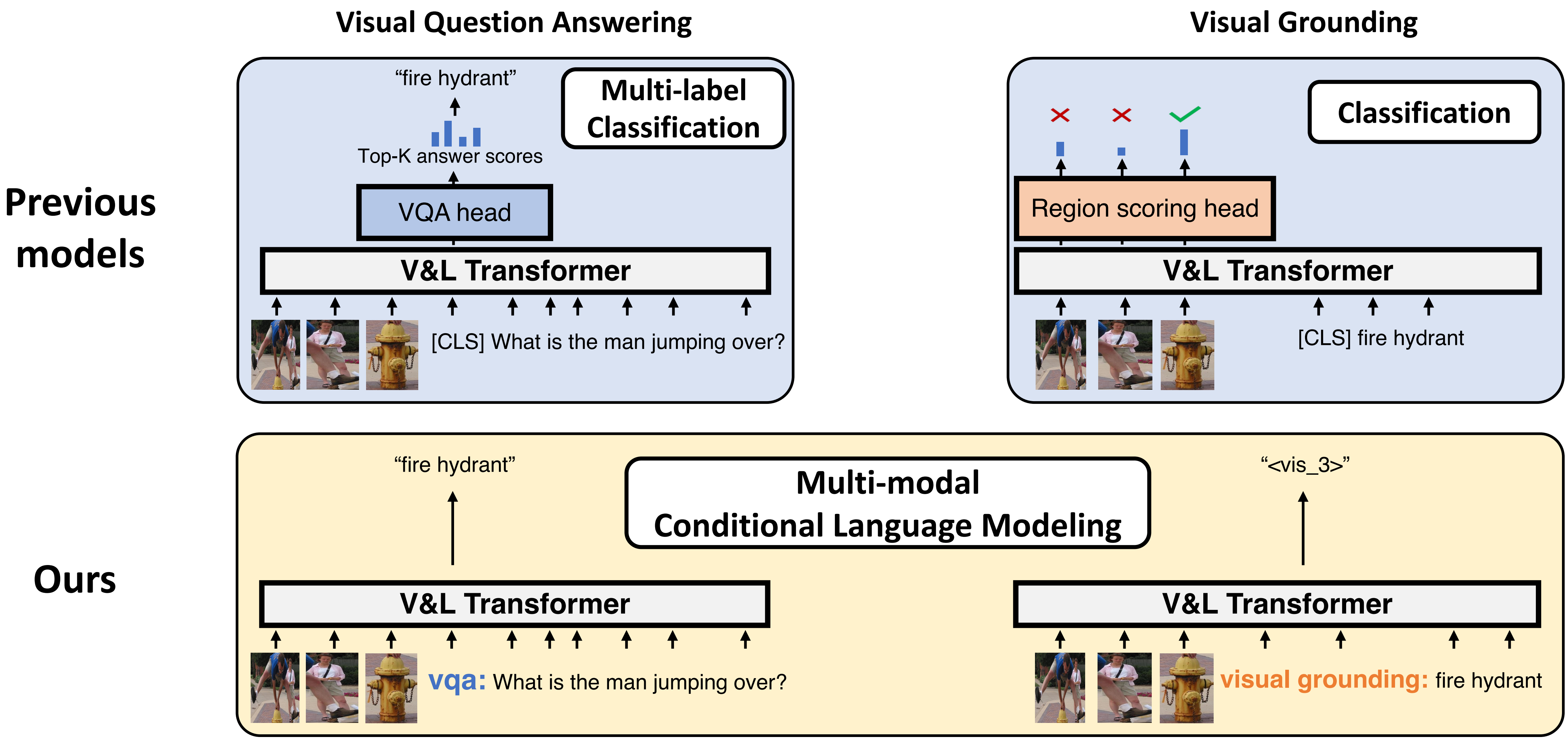
VL-T5: Many Multimodal Tasks as Text Generation



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

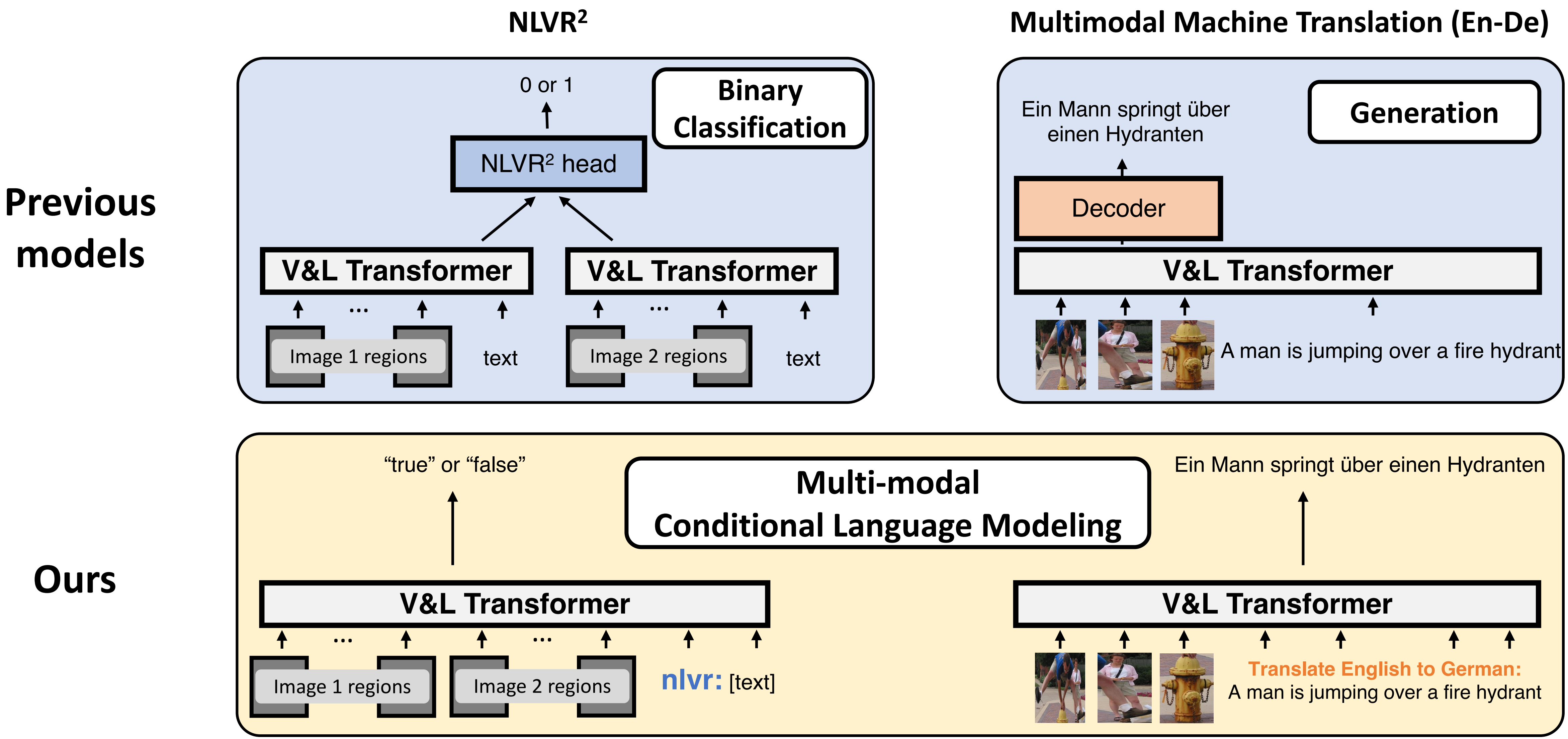
[Cho et al., ICML 2021]

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					Generative tasks	
		VQA test-std Acc	GQA test-std Acc	$NLVR^2$ test-P Acc	RefCOCOg $test^d$ Acc	VCR Q→ AR test Acc	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	$22.4^{\star}$	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 1/7th = 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)

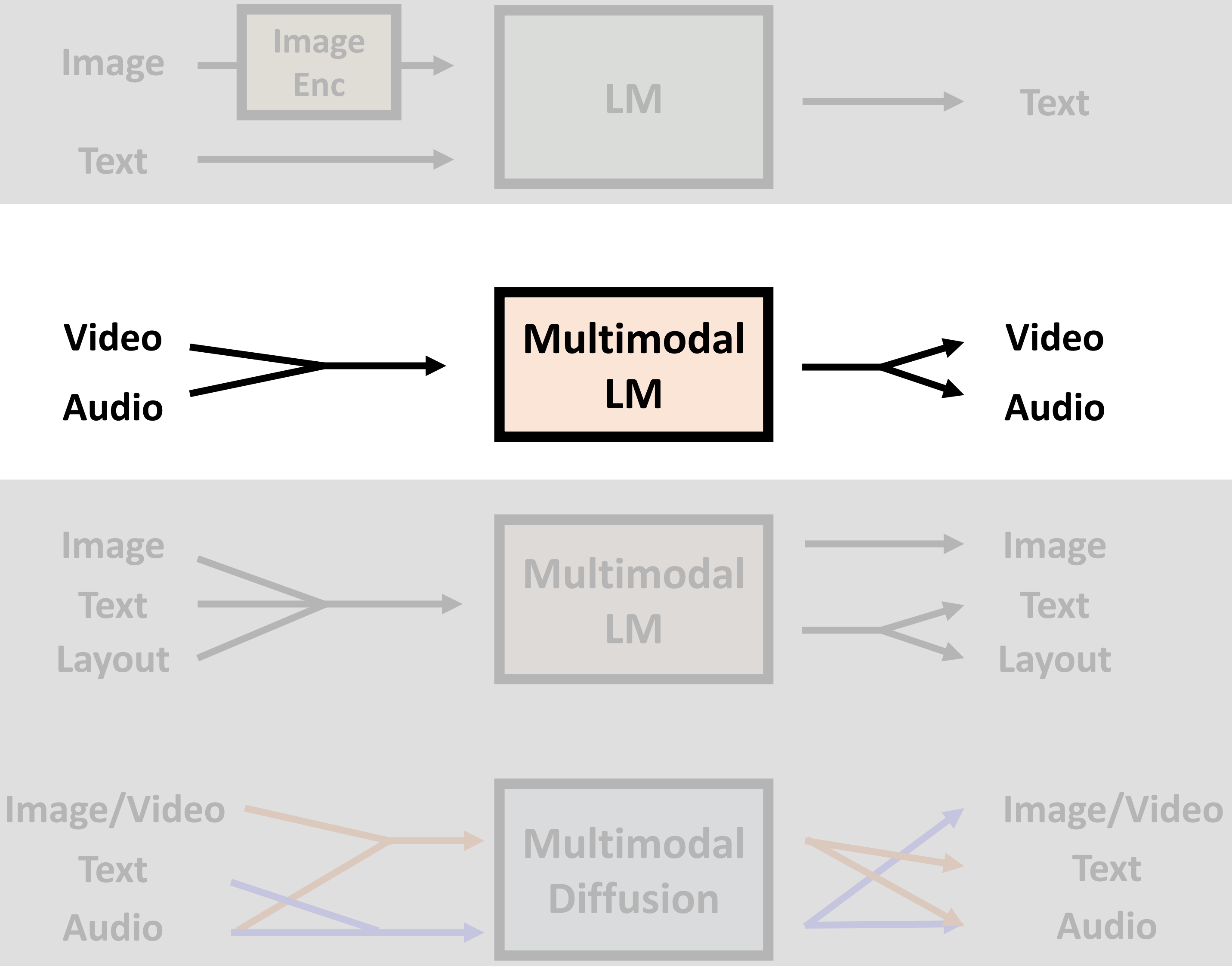
video modeling without text (audio as images)

UDOP (CVPR 2023)

document image/text/layout with single architecture

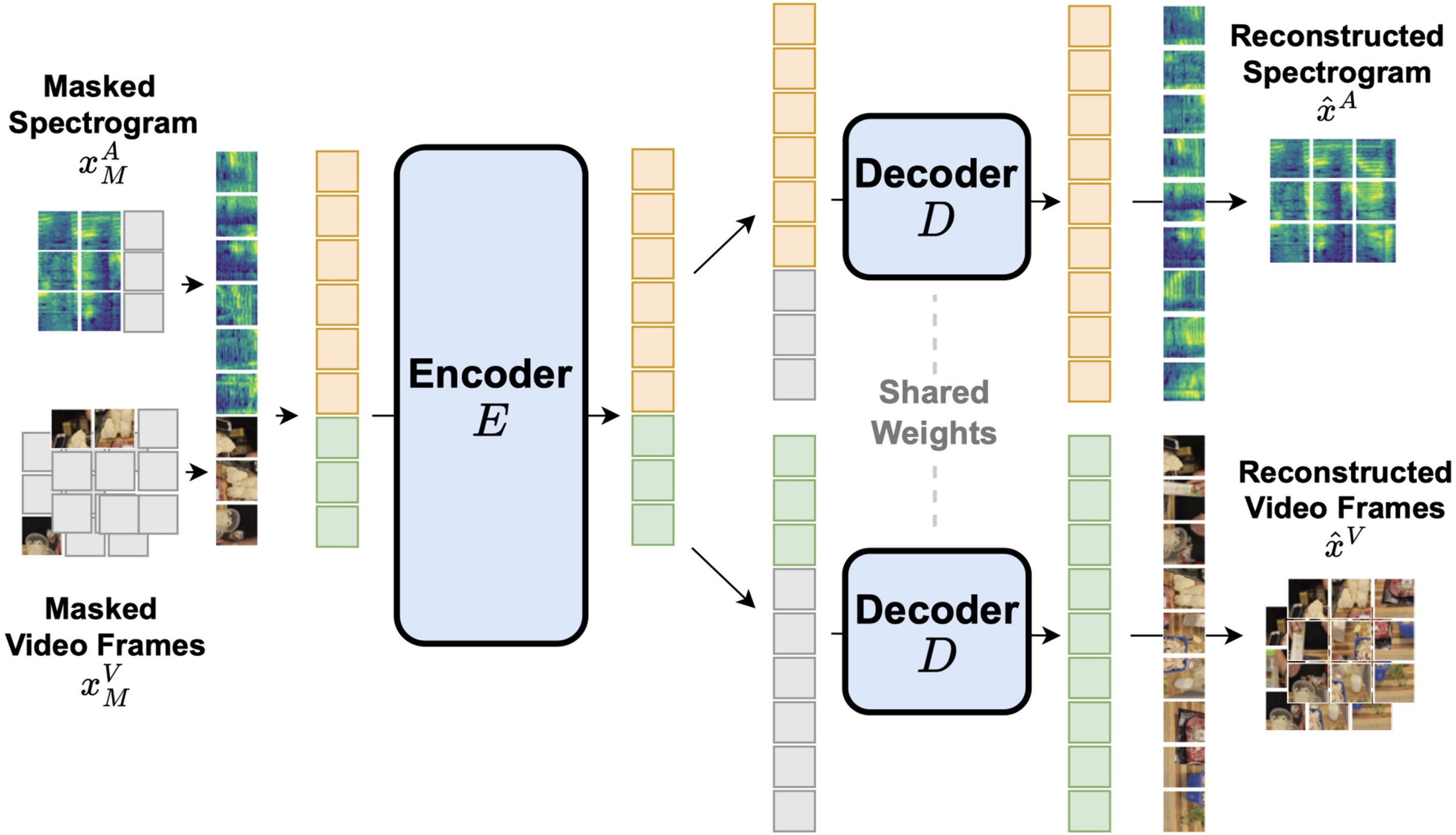
CoDi (NeurIPS 2023)

generating any-to-any input-output modality combination



TVLT: Textless Vision-Language Transformer

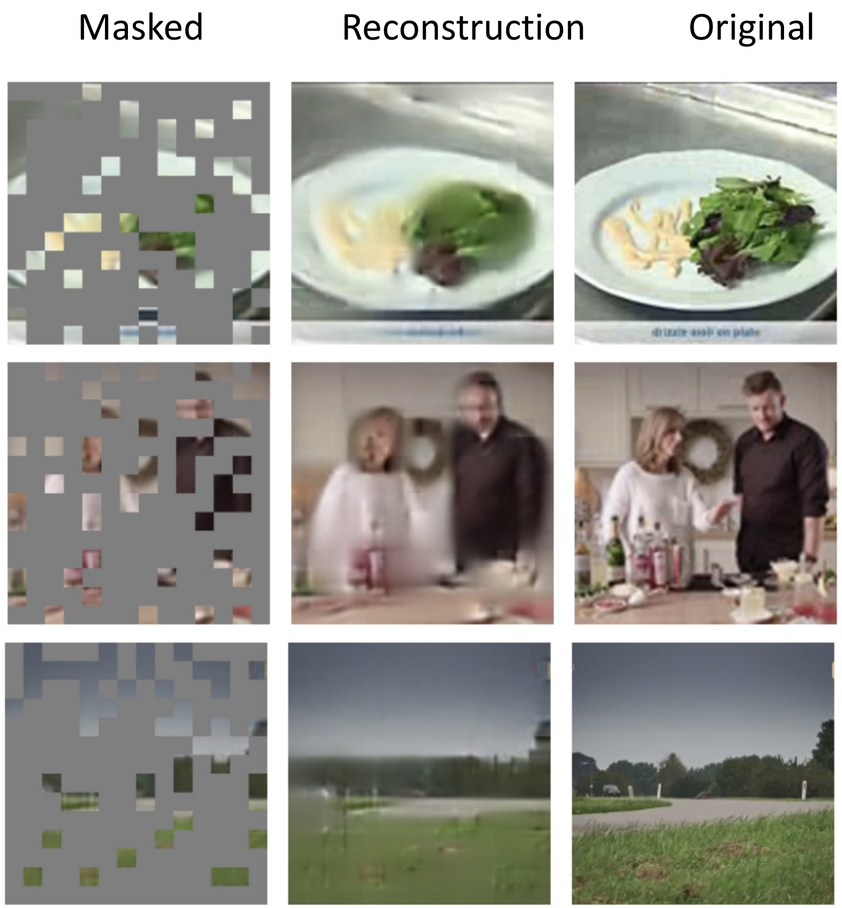
- Unified ViT-style patch embeddings for both video and audio inputs
- MAE-style enc-dec: multimodal joint encoder; decoder weights are shared for video & audio decoding
- Two objectives: (1) masked autoencoding, (2) contrastive learning

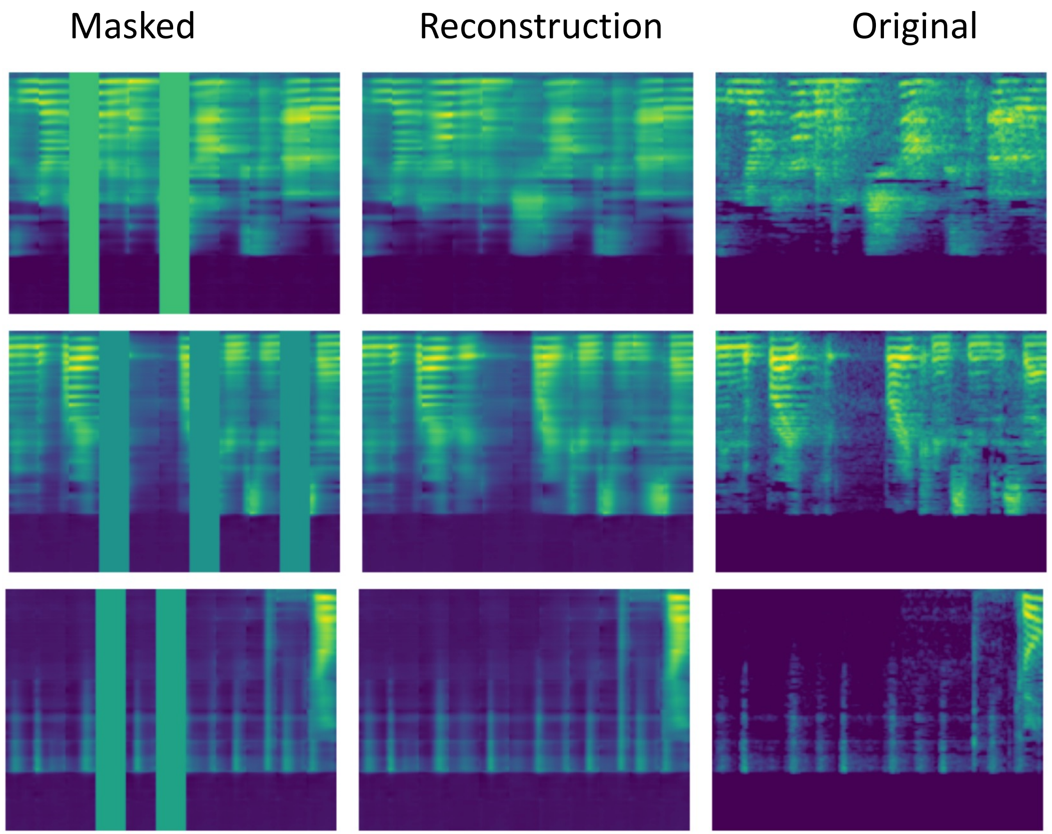


[Tang et al., NeurIPS 2022]

TVLT: Textless Vision-Language Transformer

- Results: Audio-based TVLT (**w/o any ASR/tokenization/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 inference, 1/3 #parameters)!





[Tang et al., NeurIPS 2022]

Part 1: Unified/Universal Multimodal Learning

VL-T5 (ICML 2021)

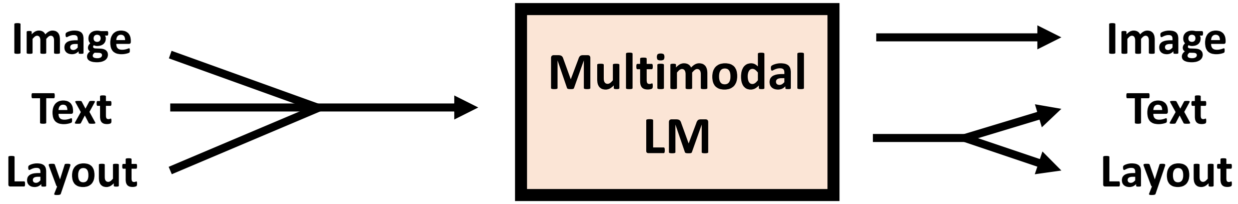
all multimodal tasks via text generation

TVLT (NeurIPS 2022)

video modeling without text (audio as images)

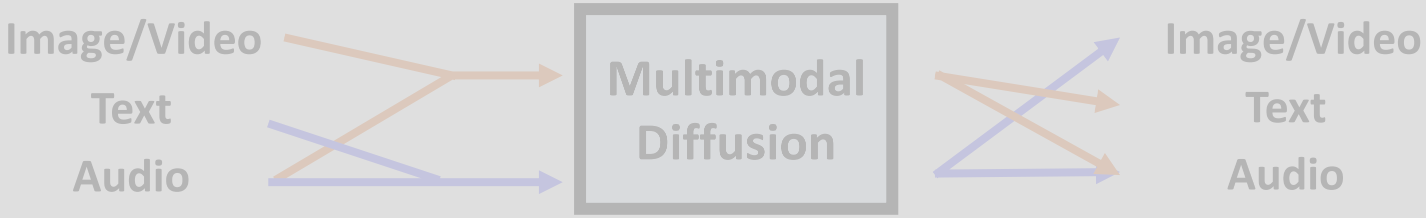
UDOP (CVPR 2023)

document image/text/layout with single architecture



CoDi (NeurIPS 2023)

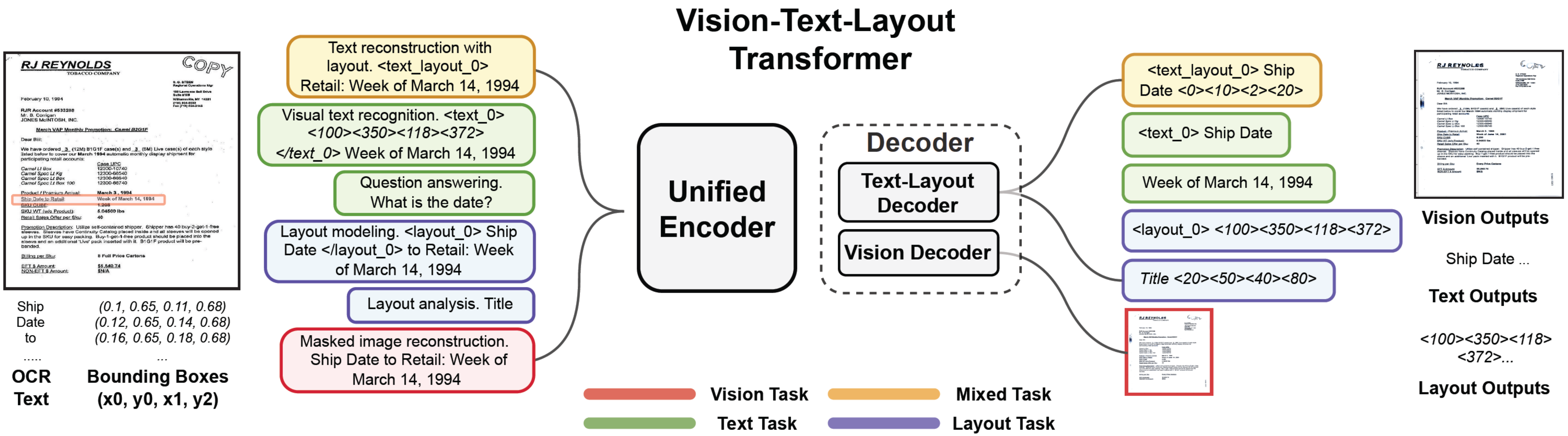
generating any-to-any input-output modality combination



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

(even different multi-image / image+text layouts hard to handle!!)

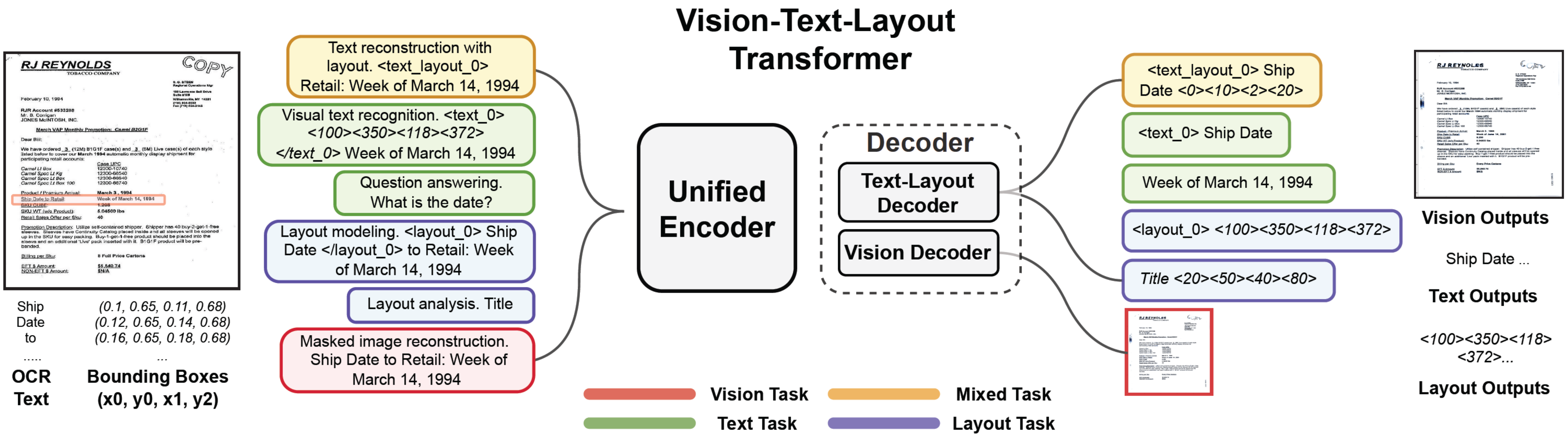
- Unifies text, image, layout modalities (*w/o specialized modules incl. OCR or layout-specific architectures*) with varied task formats, doing document understanding + generation/editing via masked image reconstruction.



[Tang et al., CVPR 2023]

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

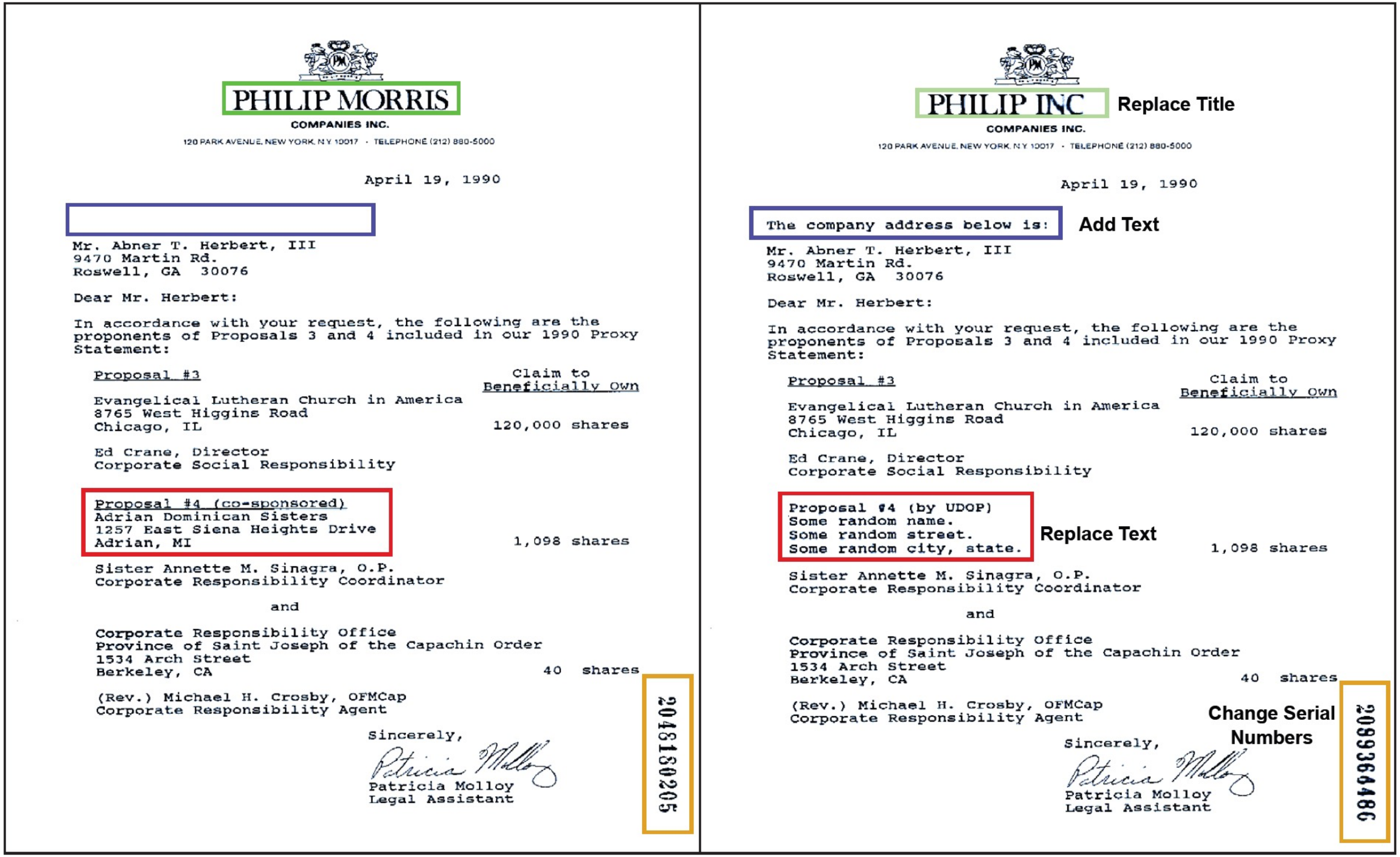
- Unifies text, image, layout modalities (*w/o specialized modules incl. OCR or layout-specific architectures*) with varied task formats, doing document understanding + generation/editing via masked image reconstruction.



- 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)

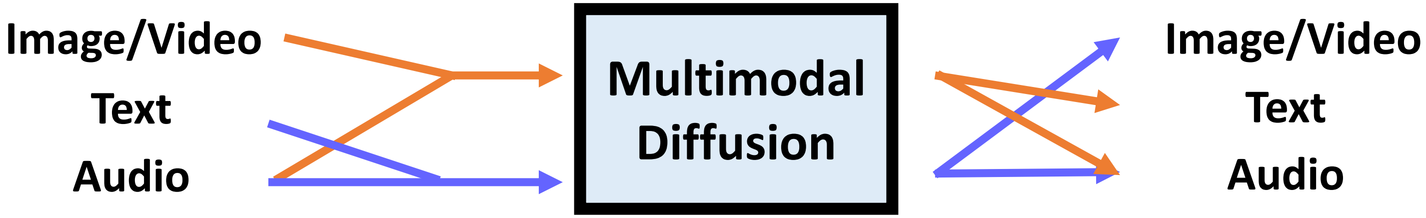
video modeling without text (audio as images)

UDOP (CVPR 2023)

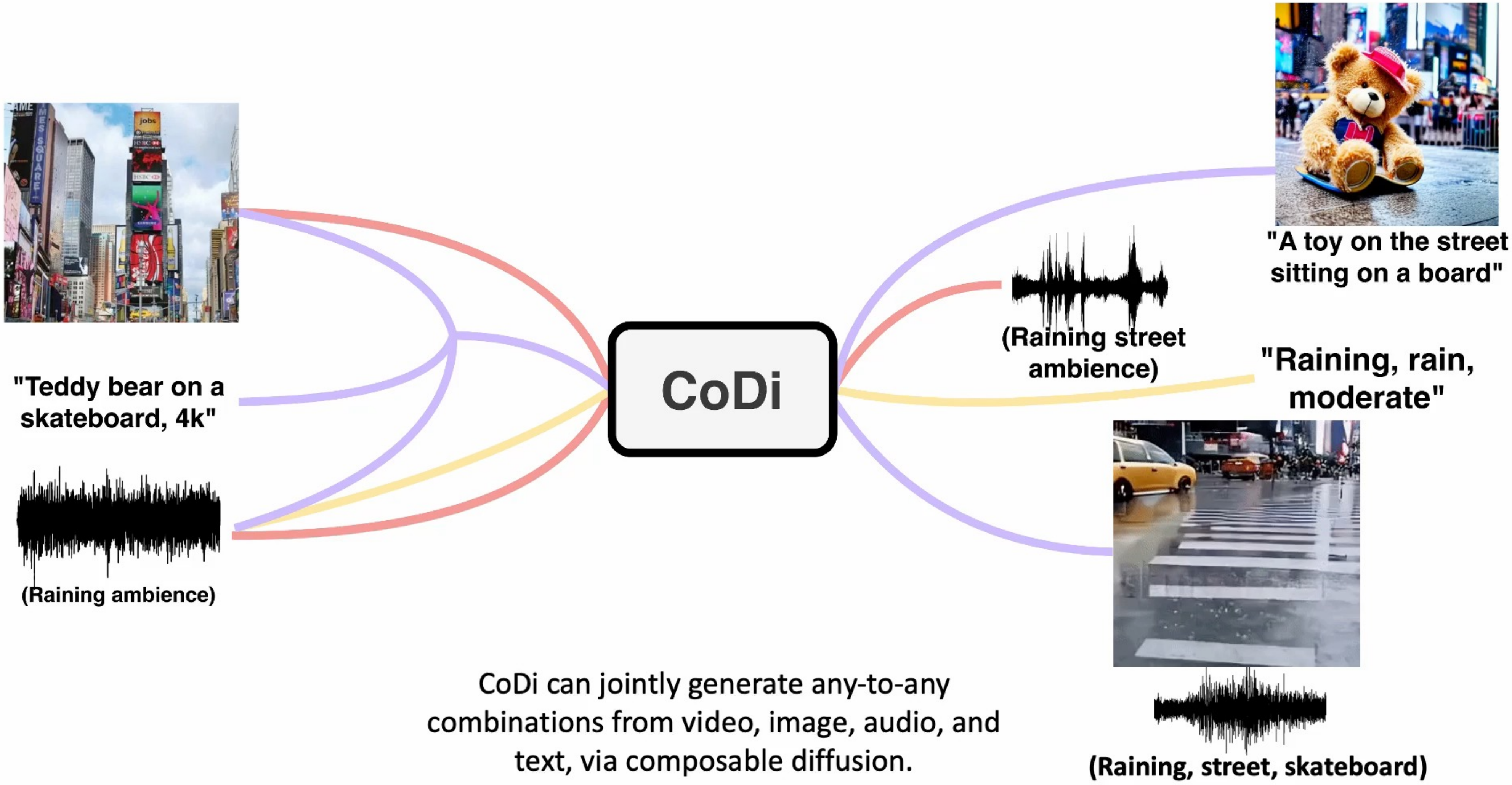
document image/text/layout with single architecture

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!

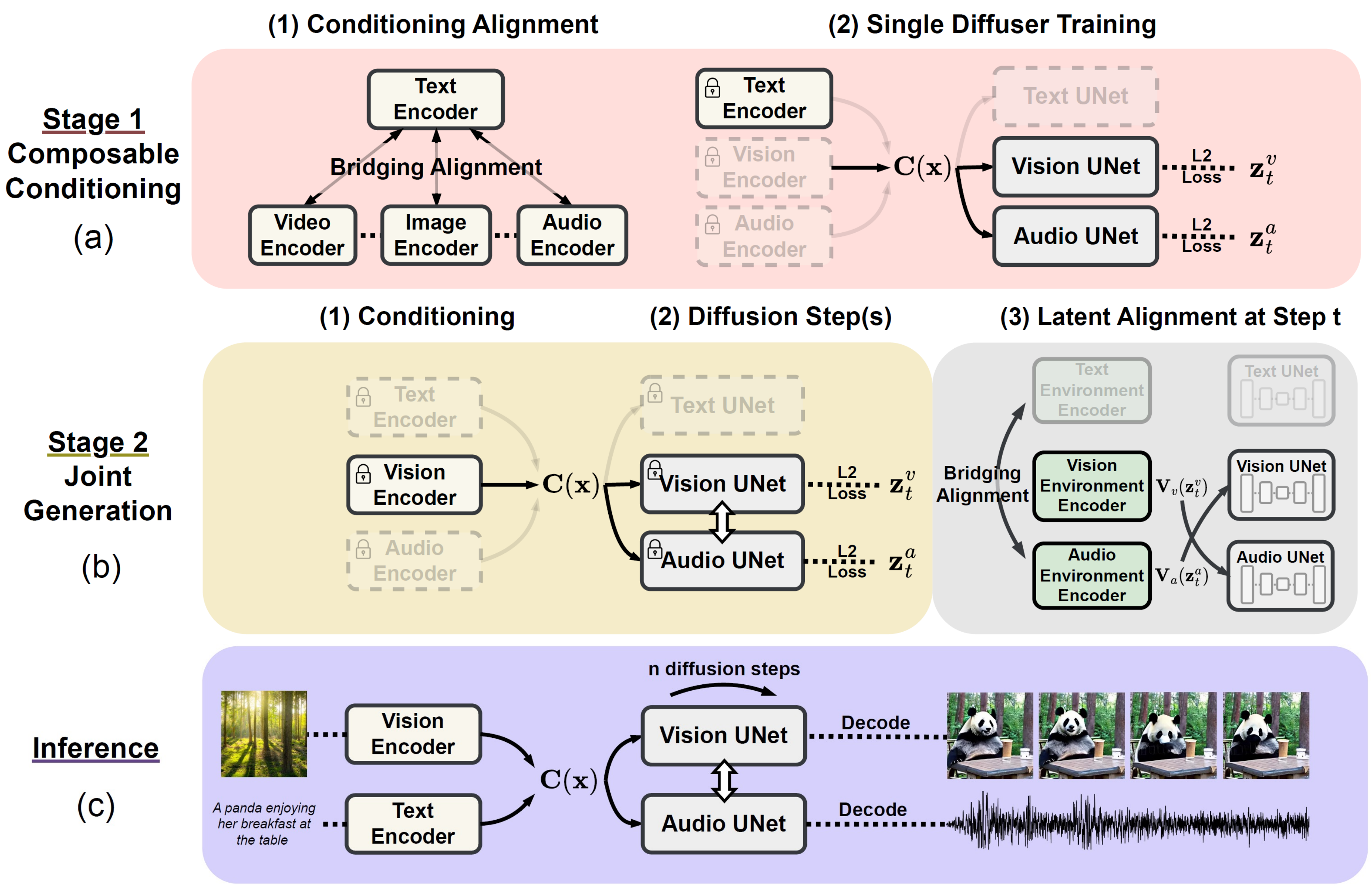
- **BUT** training such a model presents **significant costs**, as the # combinations for input and output modalities scales **exponentially** & training datasets **missing** for many combinations of modalities.

- 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]

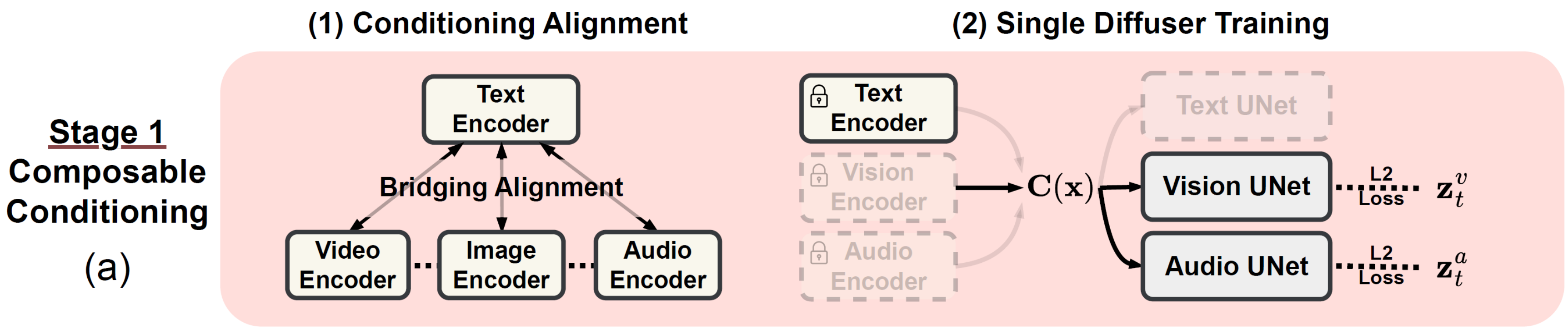
CoDi: Any-to-Any Multimodal Generation



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

[Tang et al., NeurIPS 2023]

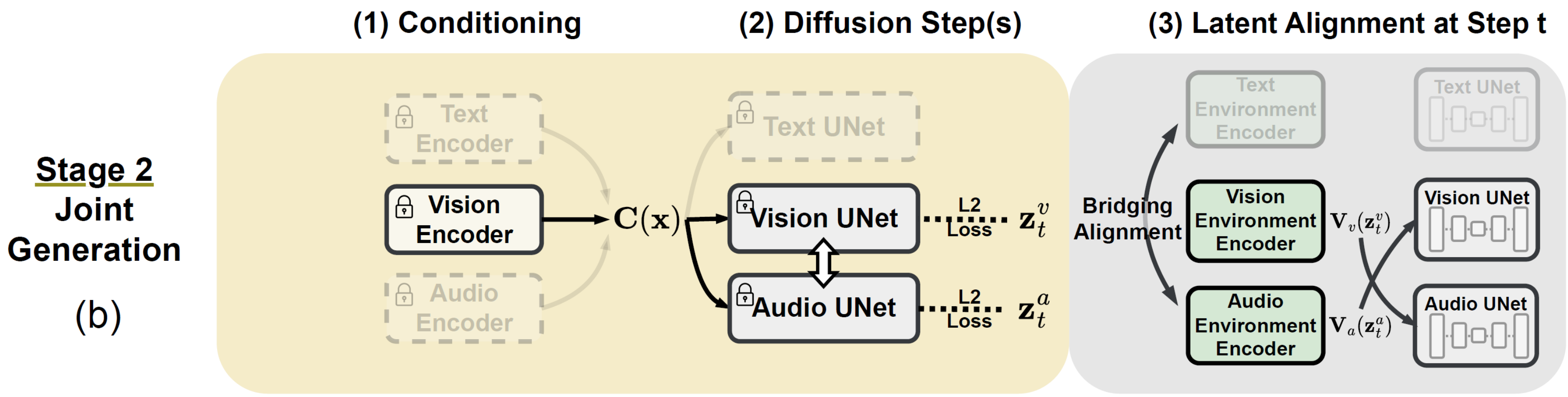
CoDi: Any-to-Any Multimodal Generation



- **Stage 1**: 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.

CoDi: Any-to-Any Multimodal Generation



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

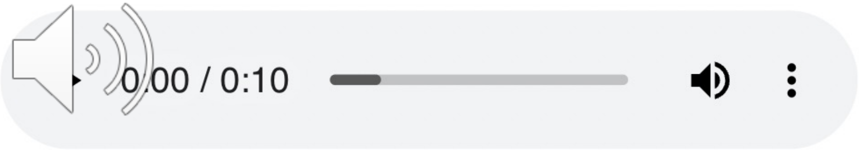
- This enables CoDi to seamlessly **mix/generate any group of output modalities, w/o training** on all generation combinations (with linear # training objectives).

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

[Tang et al., NeurIPS 2023]

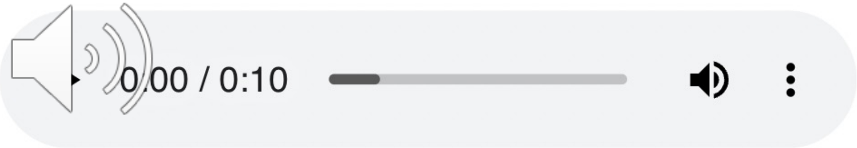
CoDi: Any-to-Any Multimodal Generation

Audio + Image → Text + Image



CoDi: Any-to-Any Multimodal Generation

Audio + Image → Text + Image



"Playing piano in a forest."

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: Any-to-Any Multimodal Generation

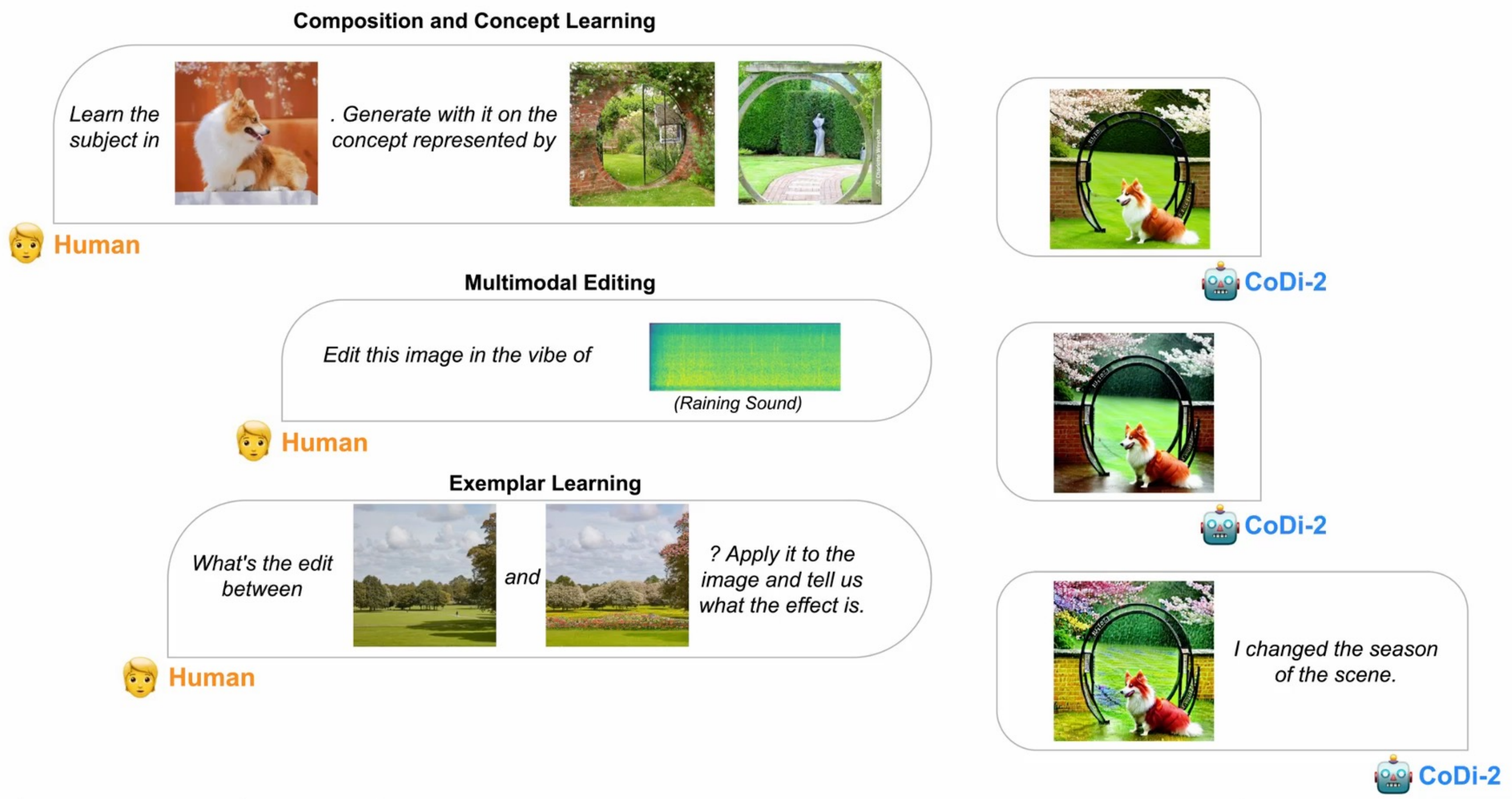
Text + Image → Video

"Eating on a coffee table."

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

[Tang et al., NeurIPS 2023]

CoDi-2: Interleaved & Interactive Any-to-Any Generation (allows Reasoning)

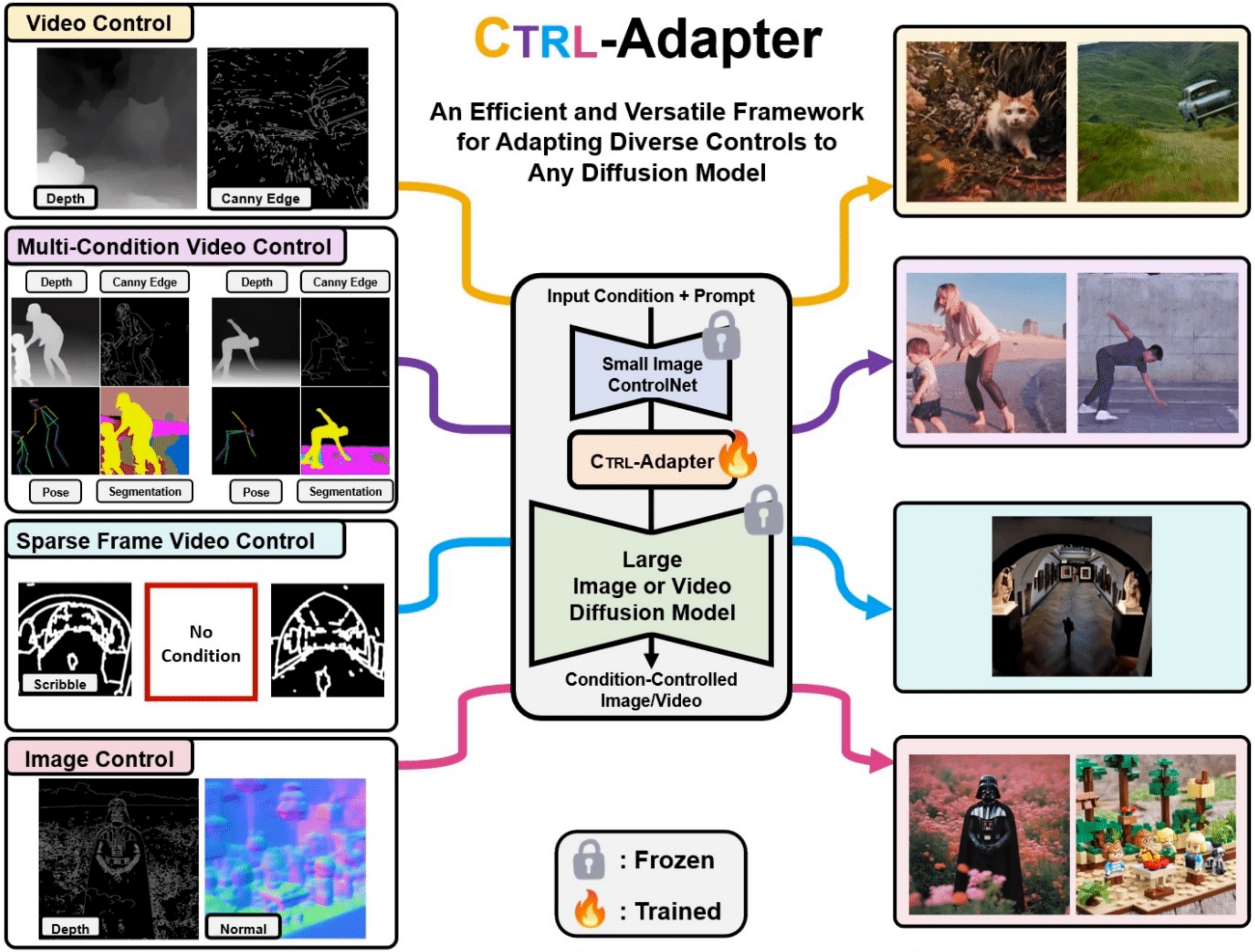


CoDi-2: In-Context, Interleaved, and Interactive Any-to-Any Generation

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

[Tang et al., CVPR 2024]

CTRL-Adapter: Efficient+Versatile Adaptation of Any Control to Any Diffusion



https://ctrl-adapter.github.io/

[Lin et al., 2024]

Talk Outline

A journey of multimodal generative models 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 [CVPR 2024]
- **Interpretable Multimodal Generation via LLM Planning/Programming Agents** (for Understanding, Control, Faithfulness, OOD)
  - VPGen: Step-by-Step Text-to-Image Generation with Interpretable Visual Programming [NeurIPS 2023]
  - VideoDirectorGPT: Consistent Multi-Scene Video Generation via LLM-Guided Planning [COLM 2024]
  - DiagrammerGPT: Generating Diagrams via LLM Planning [COLM 2024]; EnvGen: Adapting Environments via LLMs for Training Embodied Agents [COLM 2024]
- **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 [ICLR 2024]
- **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/Reasoning

VPGen (NeurIPS 2023)

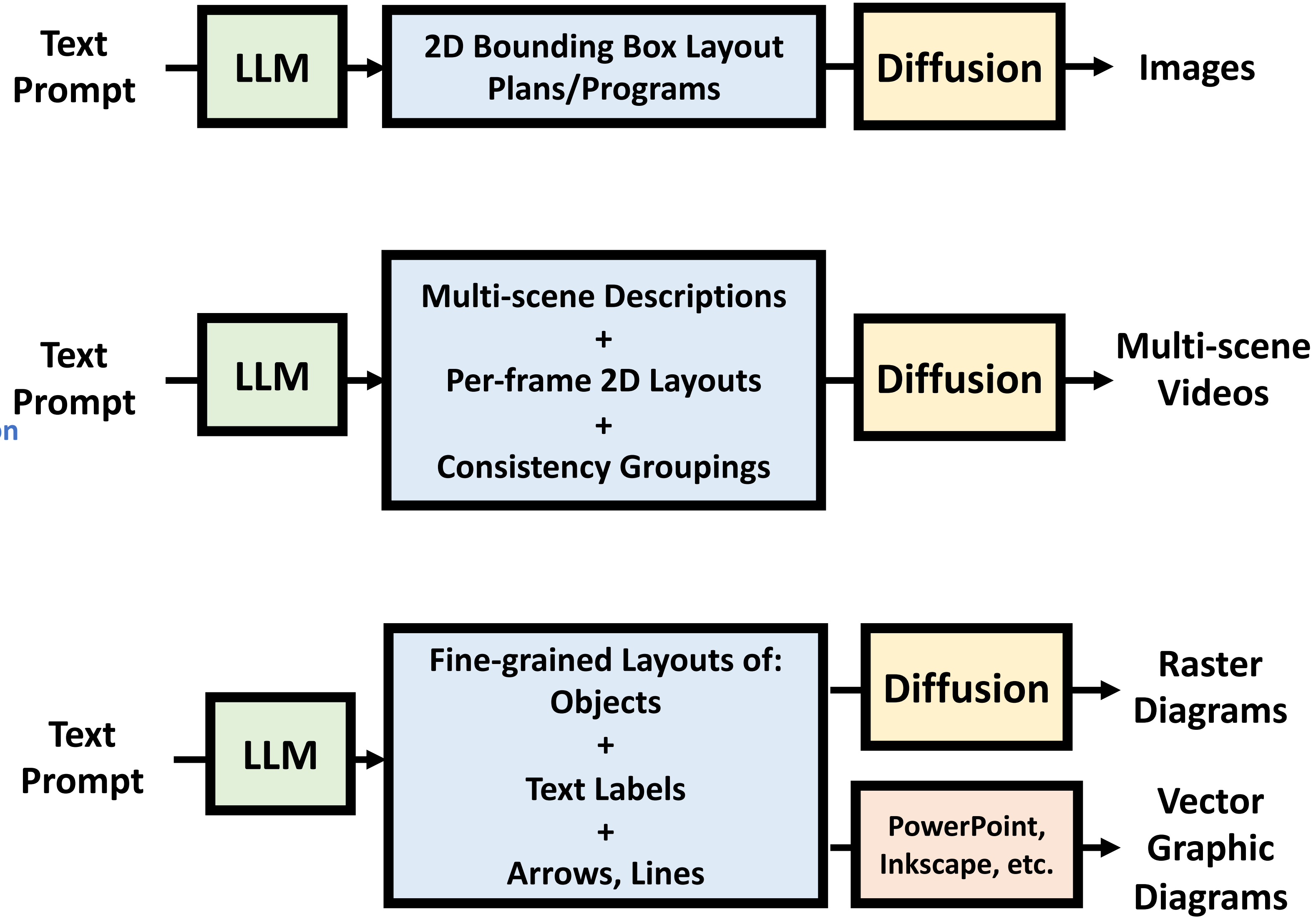
LLM Planning for Image Generation

VideoDirectorGPT (COLM 2024)

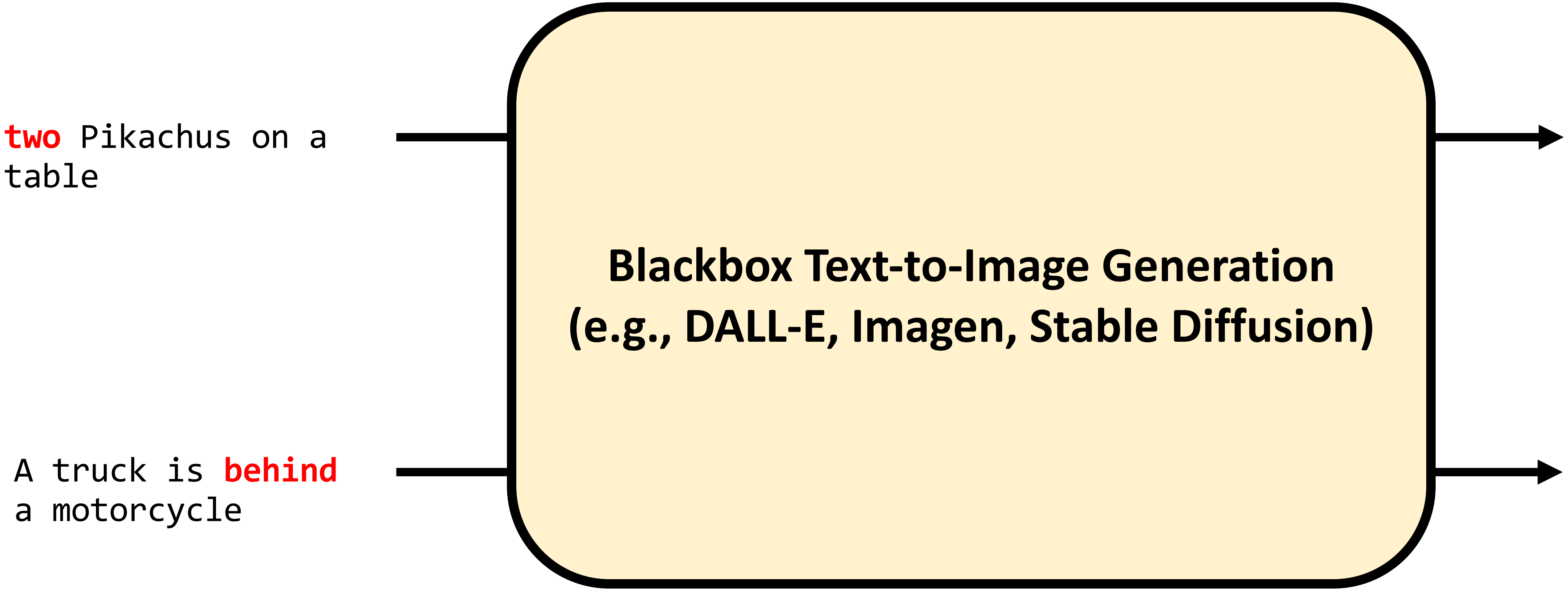
LLM Planning for Multi-Scene, Consistent Video Generation

DiagrammerGPT (COLM 2024)

LLM Planning for Open-Domain Diagram Generation



Background: Text-to-Image Generation with Blackbox Models



[Cho et al., NeurIPS 2023]

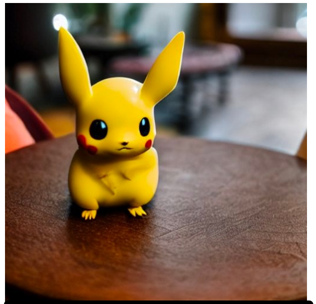
Background: Text-to-Image Generation with Blackbox Models

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 and OOD scenarios)**



one Pikachu ✗

truck is below motorcycle ✗

[Cho et al., NeurIPS 2023]

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

```
two Pikachus
on a table
```

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
```

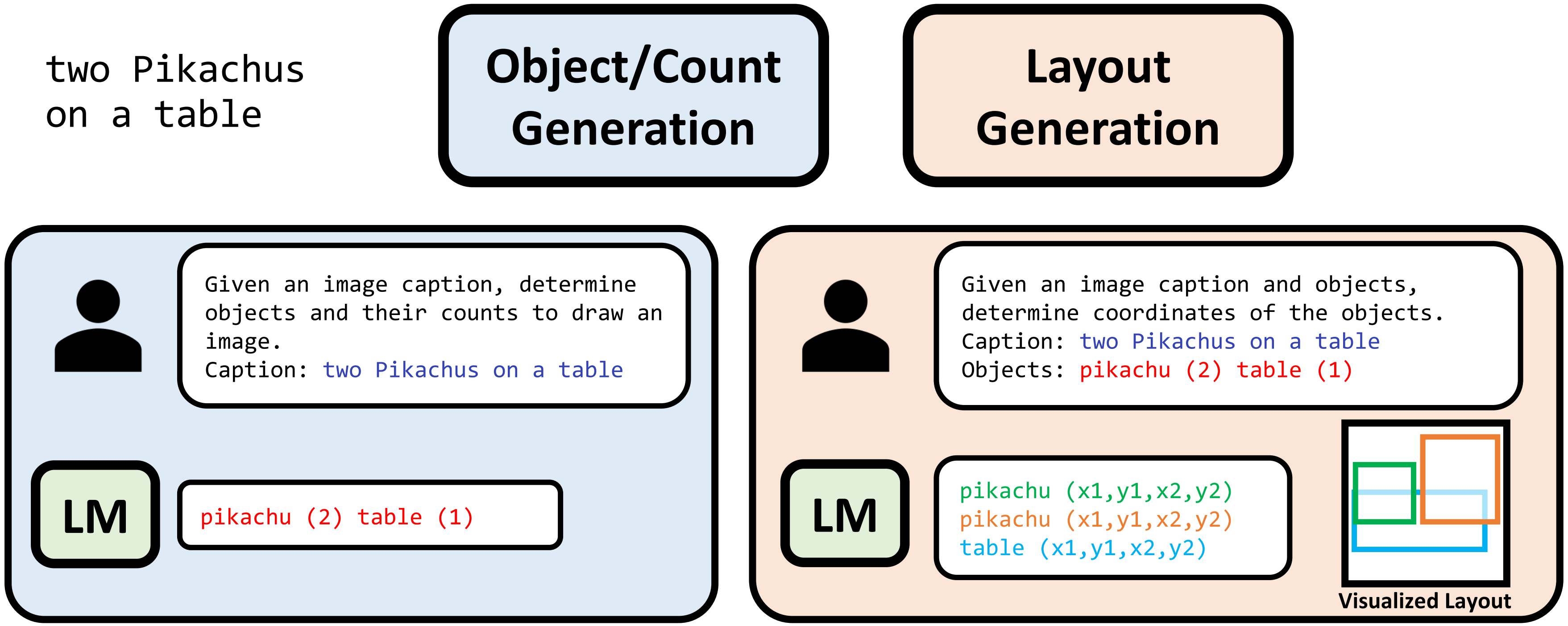
LM

```
pikachu (2) table (1)
```

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

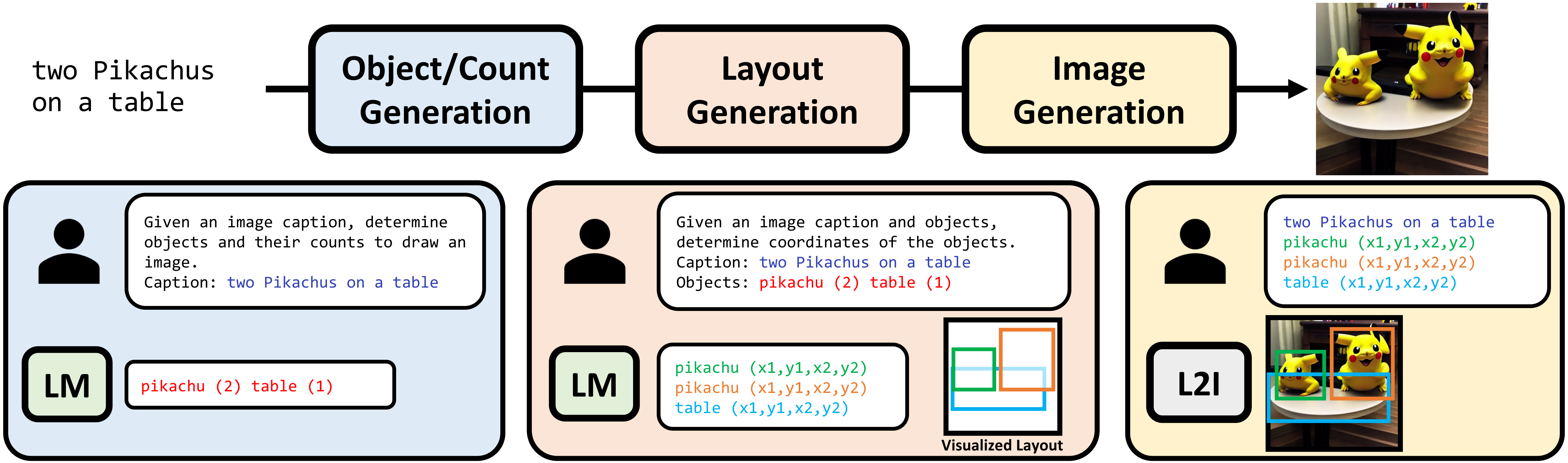
[Cho et al., NeurIPS 2023]

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



[Cho et al., NeurIPS 2023]

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



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

[Cho et al., NeurIPS 2023]

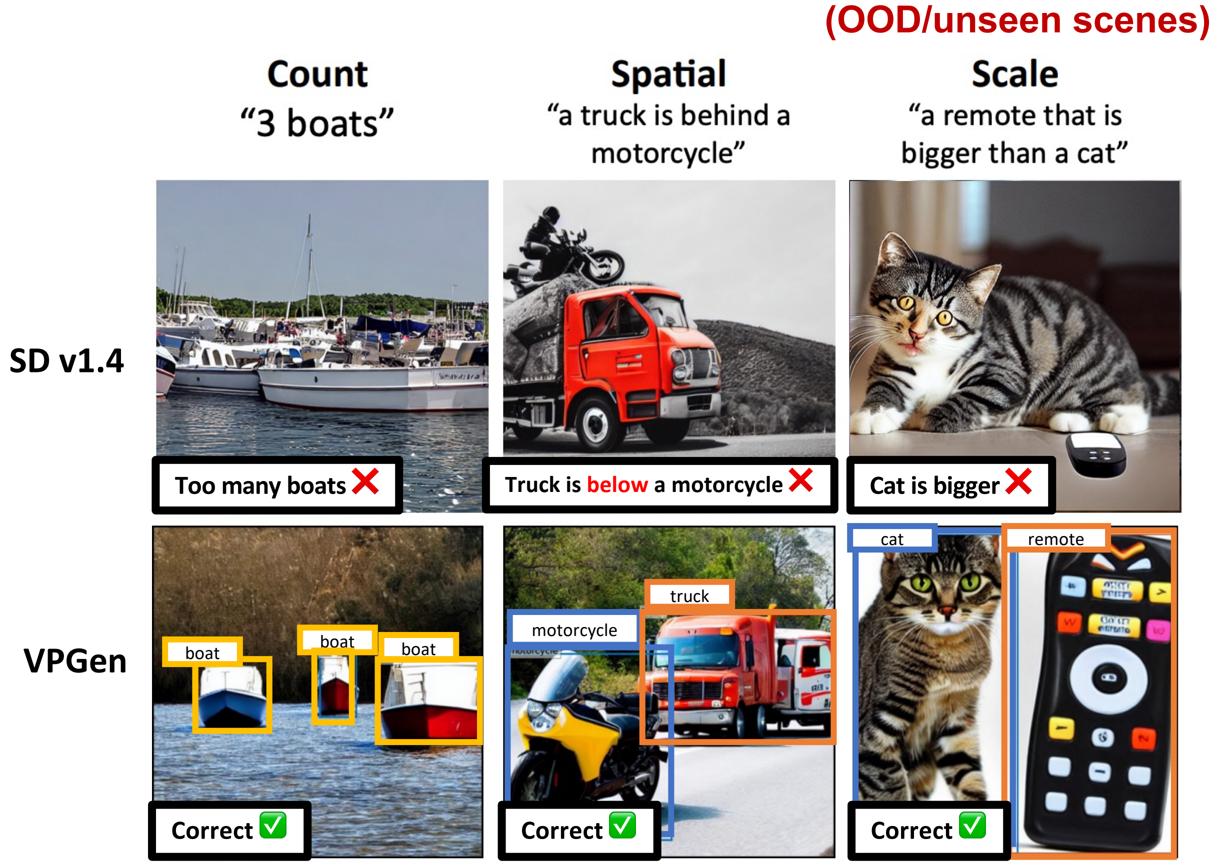
Skill-based Results

Our VPGen shows improved spatial control

- Generation via layout programs promotes better **understanding+planning** of structure/scale/spatial relations, including **out-of-distribution/unseen** cases (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**

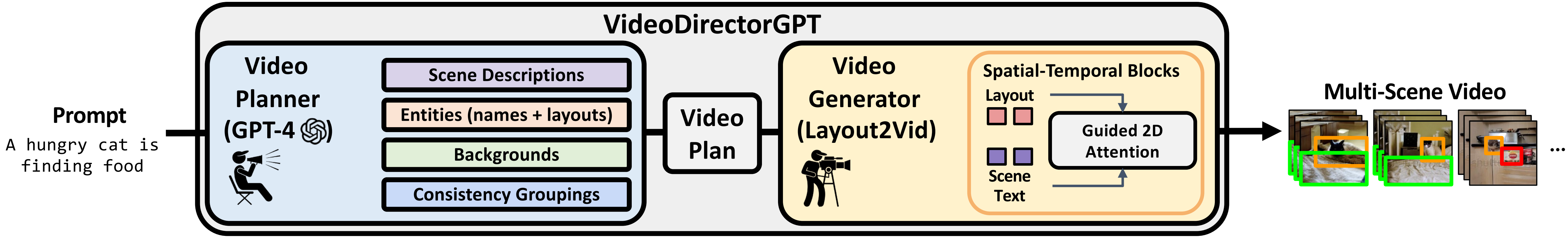
Large improvements on structural control:
- Counting
- Spatial relation
- Relative size/scale comparison



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

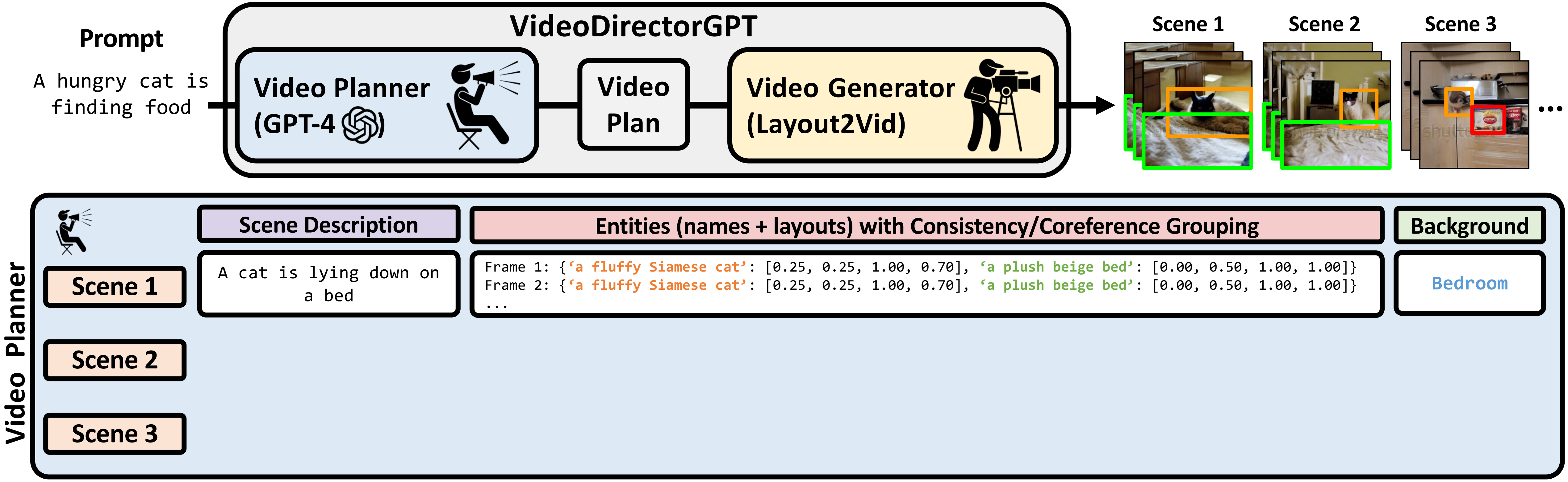
[Cho et al., NeurIPS 2023]

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



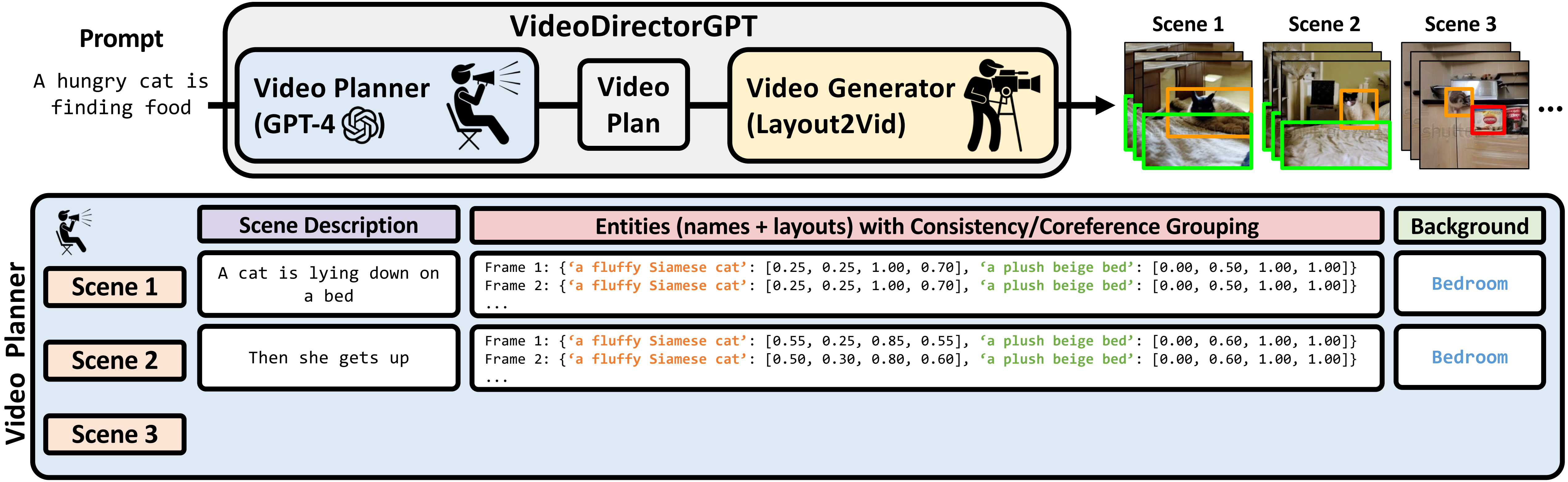
https://videodirectorgpt.github.io/

[Lin et al., COLM 2024]



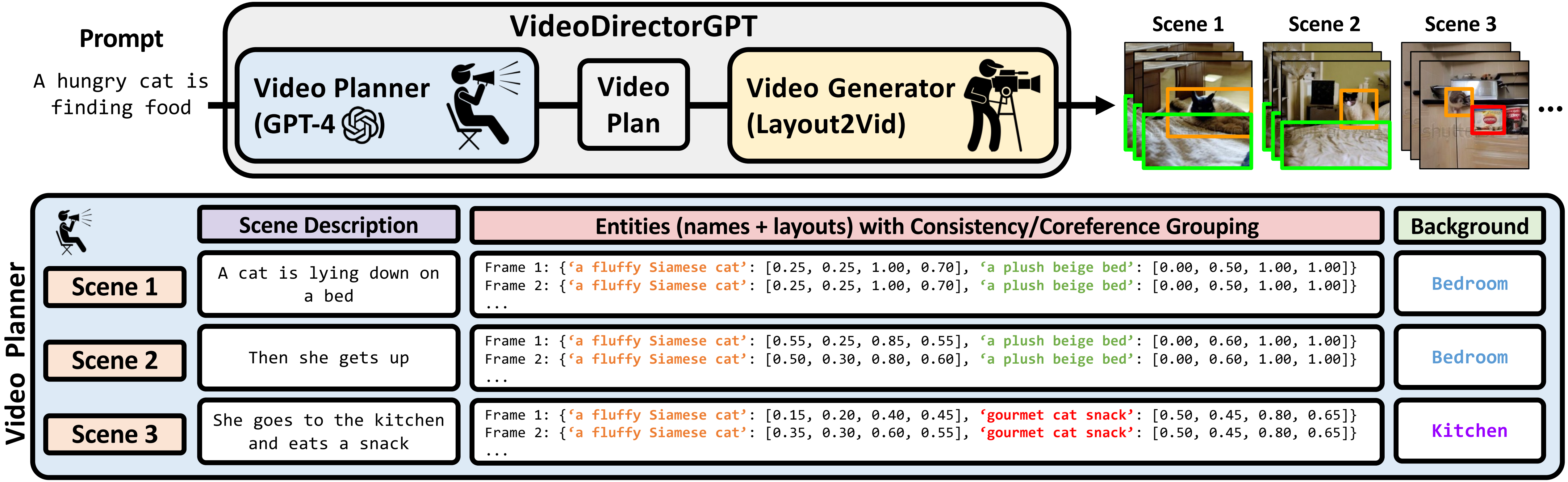
https://videodirectorgpt.github.io/

[Lin et al., COLM 2024]



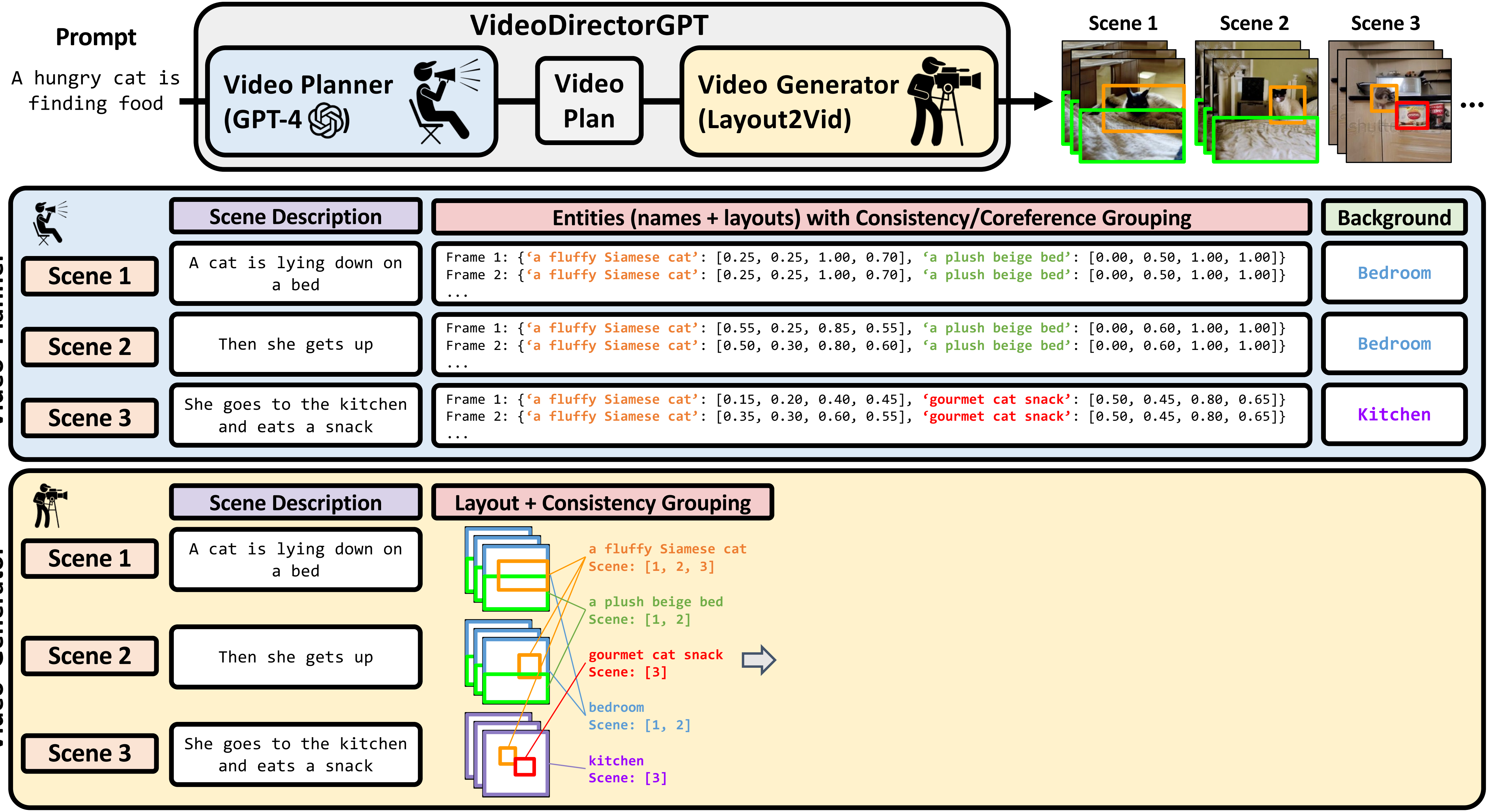
https://videodirectorgpt.github.io/

[Lin et al., COLM 2024]



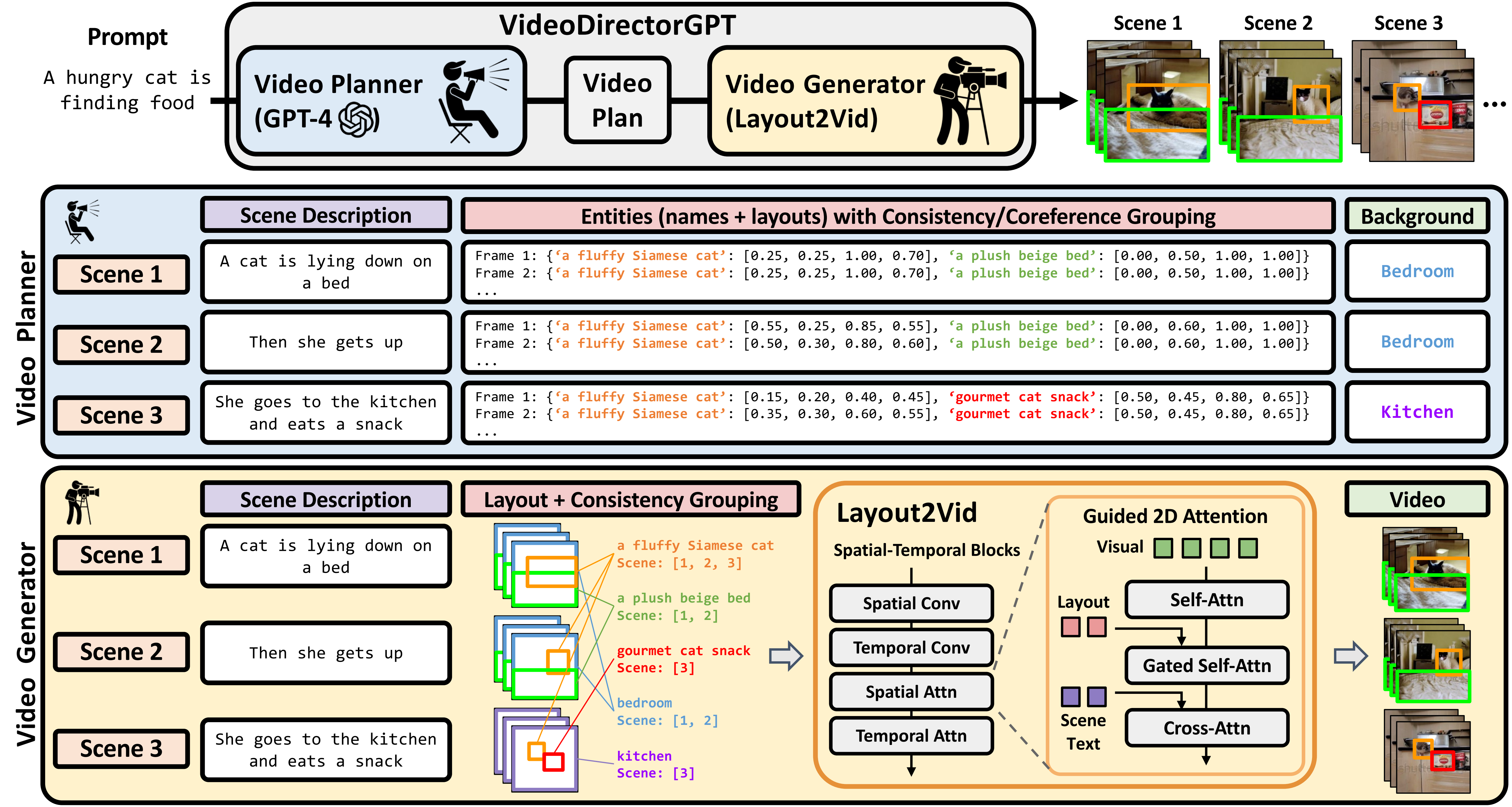
https://videodirectorgpt.github.io/

[Lin et al., COLM 2024]



https://videodirectorgpt.github.io/

[Lin et al., COLM 2024]

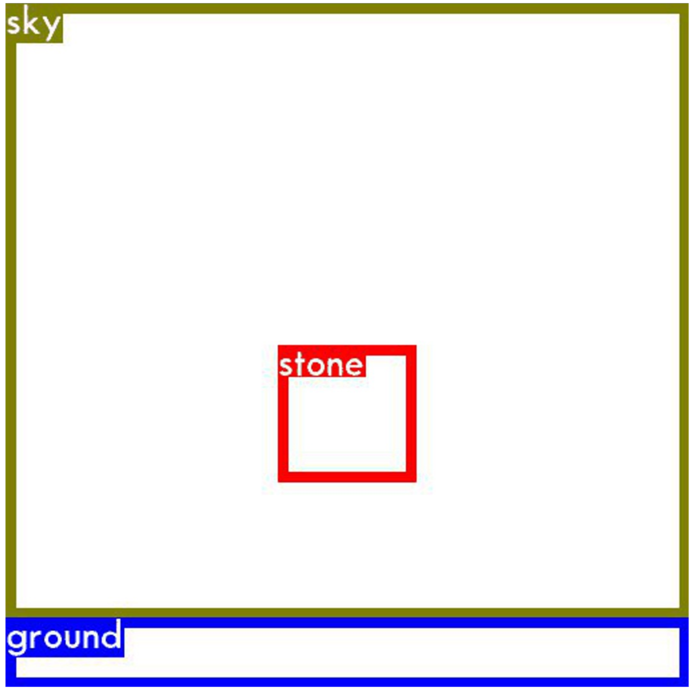


https://videodirectorgpt.github.io/ [Lin et al., COLM 2024]

Understanding of Basic Physics

Gravity

A stone thrown into the sky



Perspective

A car is approaching from a distance

Movement of Static Objects vs. Dynamic Objects

"A {**bottle/airplane**} moving from **left to right**."

Static objects
-> Movements of Camera

Objects that can move
-> Movements of Object (+ Camera)

https://videodirectorgpt.github.io/

[Lin et al., COLM 2024]

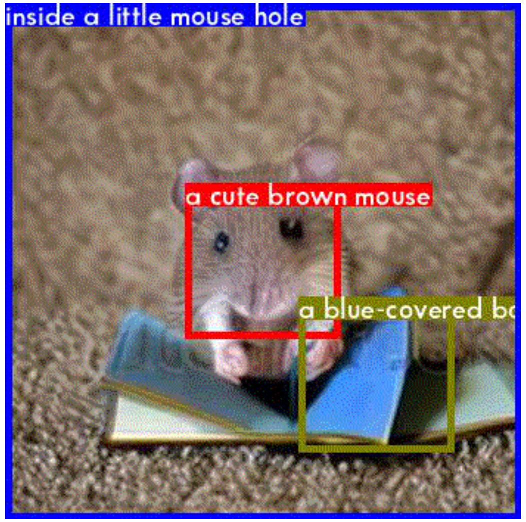
Multi-Sentence to Multi-Scene Video (Coref-SV)

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

(also helps plan+generate OOD/unseen affordances/scenes)

https://videodirectorgpt.github.io/

[Lin et al., COLM 2024]

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

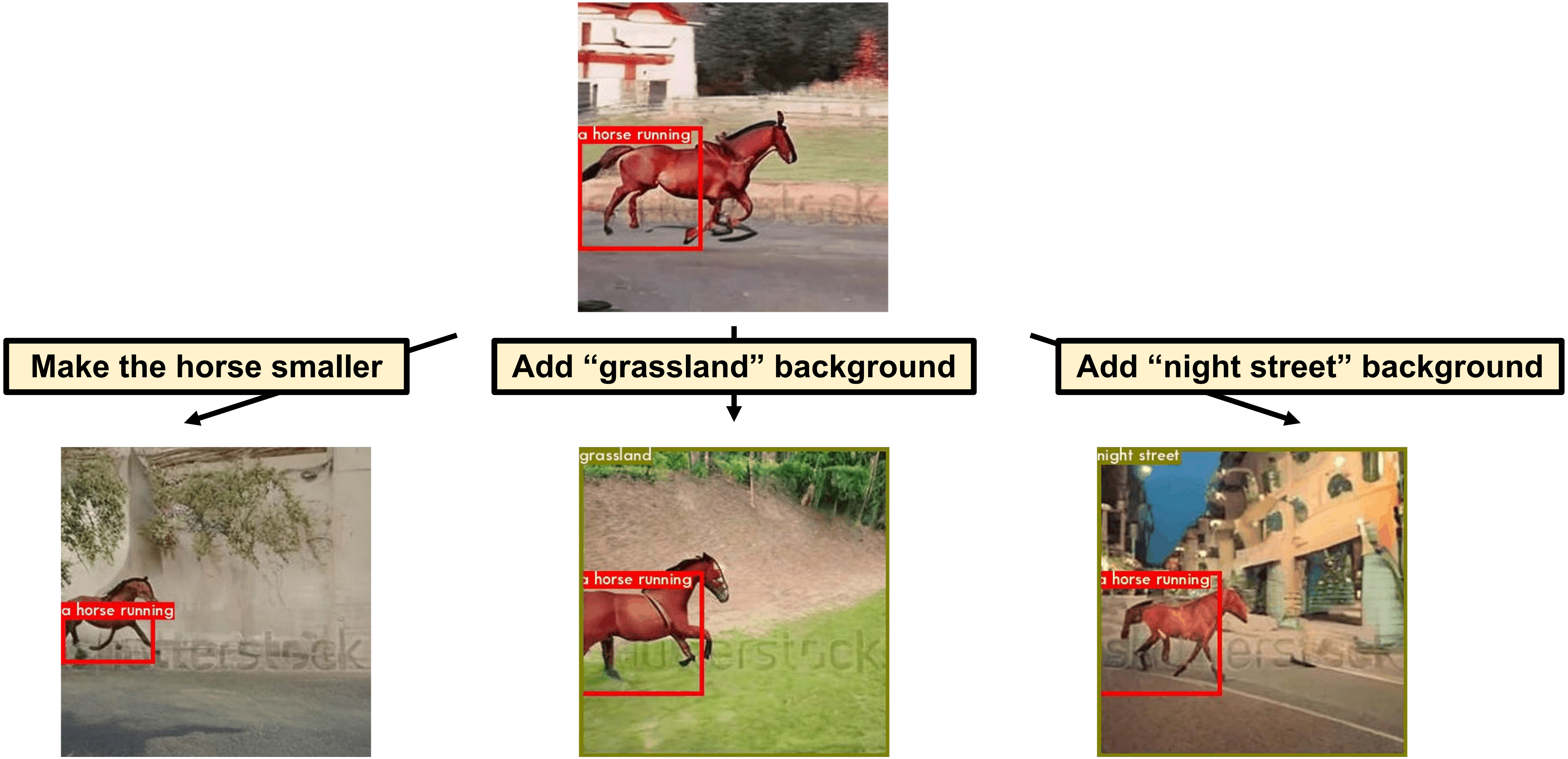
VideoDirectorGPT (Ours)

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

https://videodirectorgpt.github.io/

[Lin et al., COLM 2024]

Human-in-the-Loop Video Editing+Control



https://videodirectorgpt.github.io/

[Lin et al., COLM 2024]

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., COLM 2024]

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**

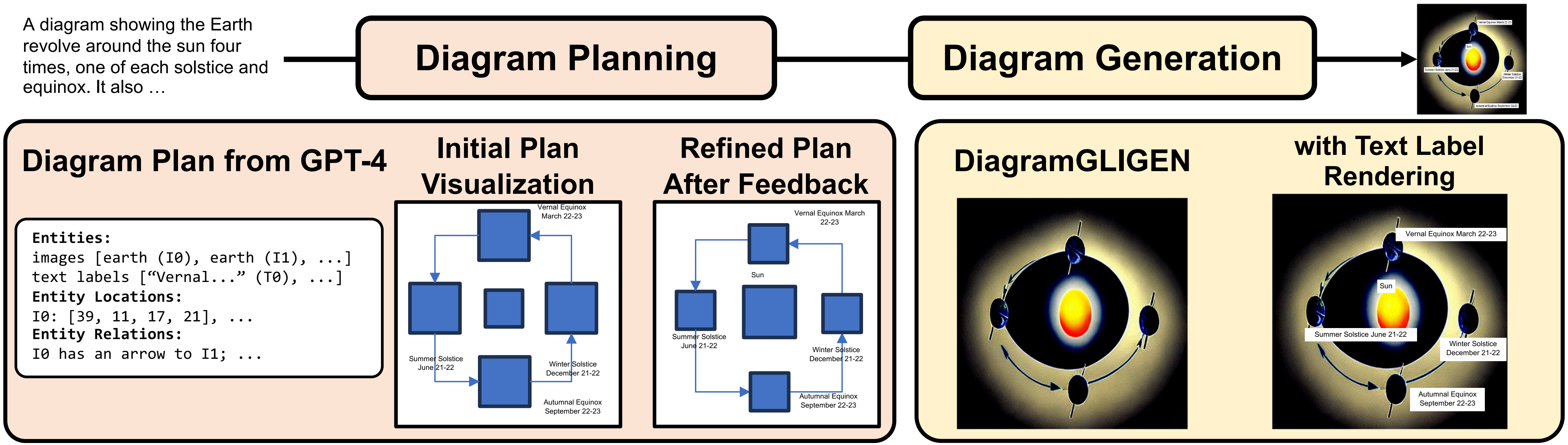
Method	ActivityNet Captions			Coref-SV	HiREST	
	FVD (↓)	FID (↓)	Consistency (↑)	Consistency (↑)	FVD (↓)	FID (↓)
ModelScopeT2V	980	18.12	46.0	16.3	1322	23.79
ModelScopeT2V (with GT co-reference; oracle)	-	-	-	37.9	-	-
VideoDirectorGPT (Ours)	**805**	**16.50**	**64.8**	**42.8**	**733**	**18.54**

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/

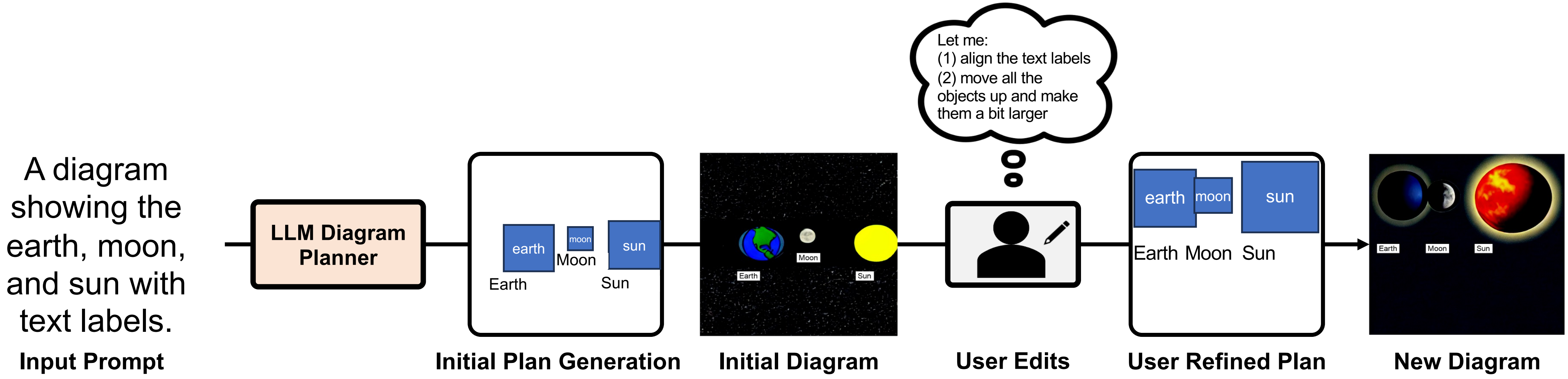
[Lin et al., COLM 2024]

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



[Zala et al., COLM 2024]

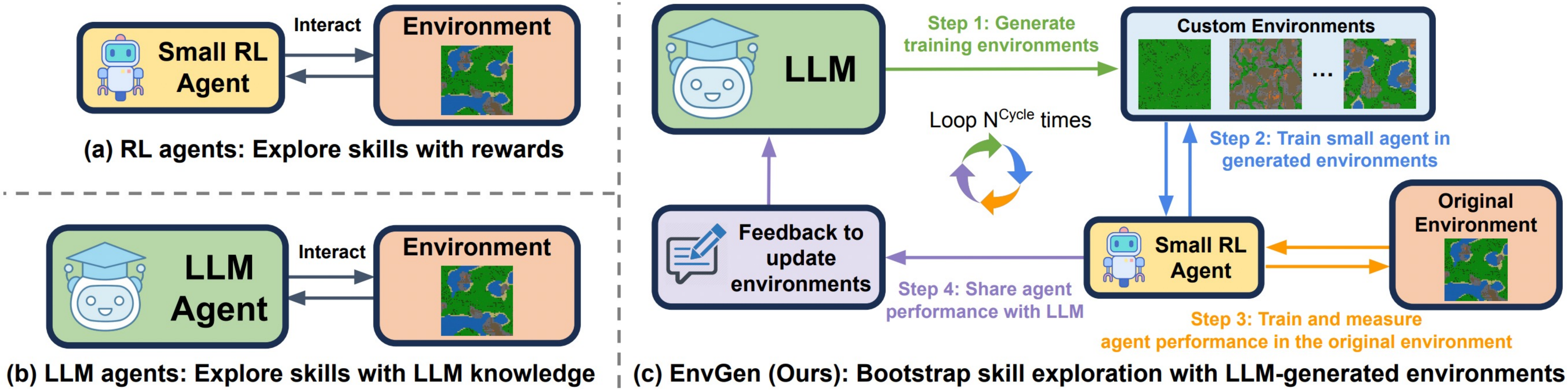
Human-in-the-Loop Diagram Editing



https://diagrammergpt.github.io/

[Zala et al., COLM 2024]

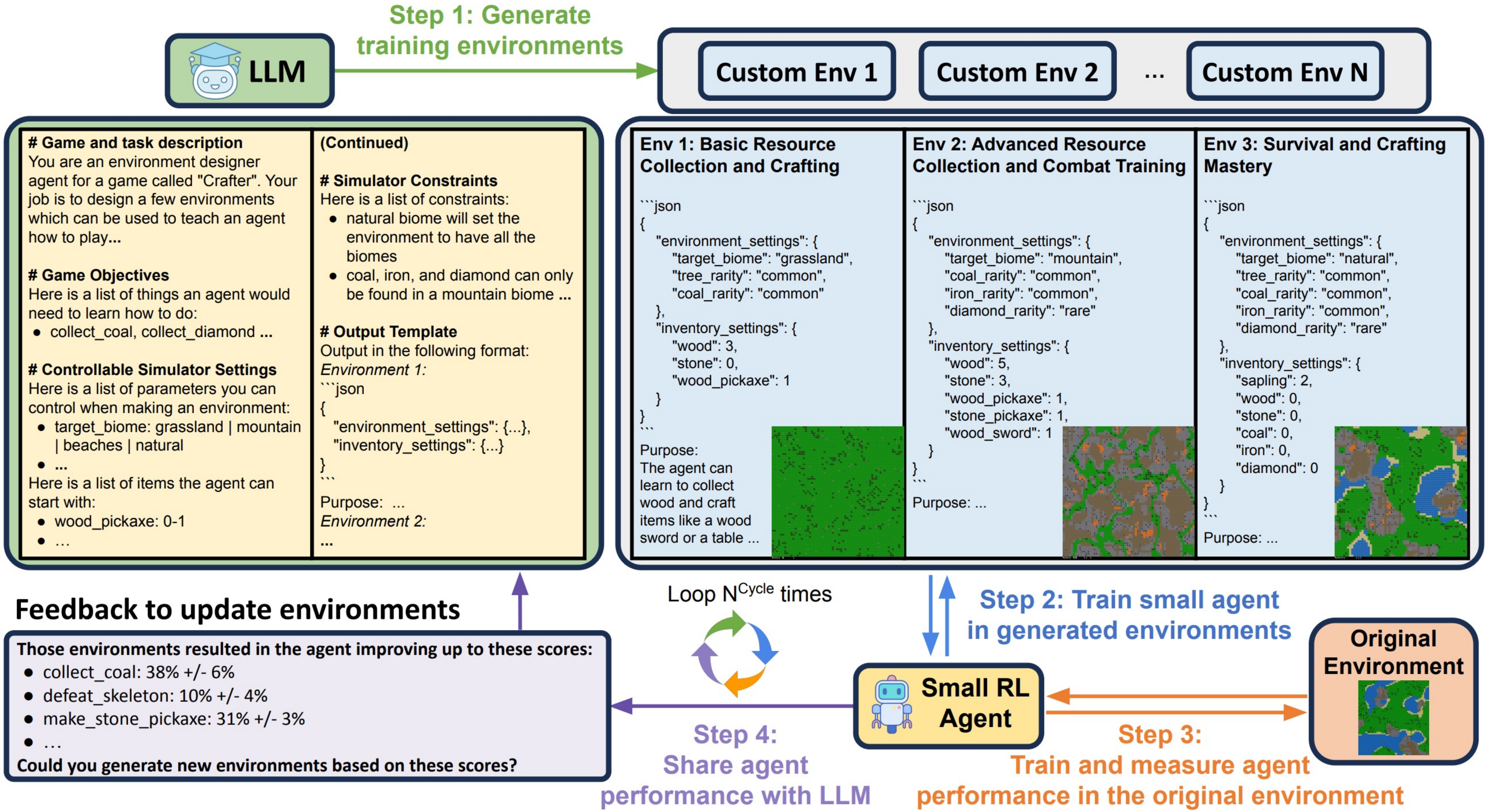
EnvGen: LLM-Planned Adaptive Environment Generation for Training Agents



https://envgen-llm.github.io/

[Zala et al., COLM 2024]

EnvGen: LLM-Planned Adaptive Environment Generation for Training Agents



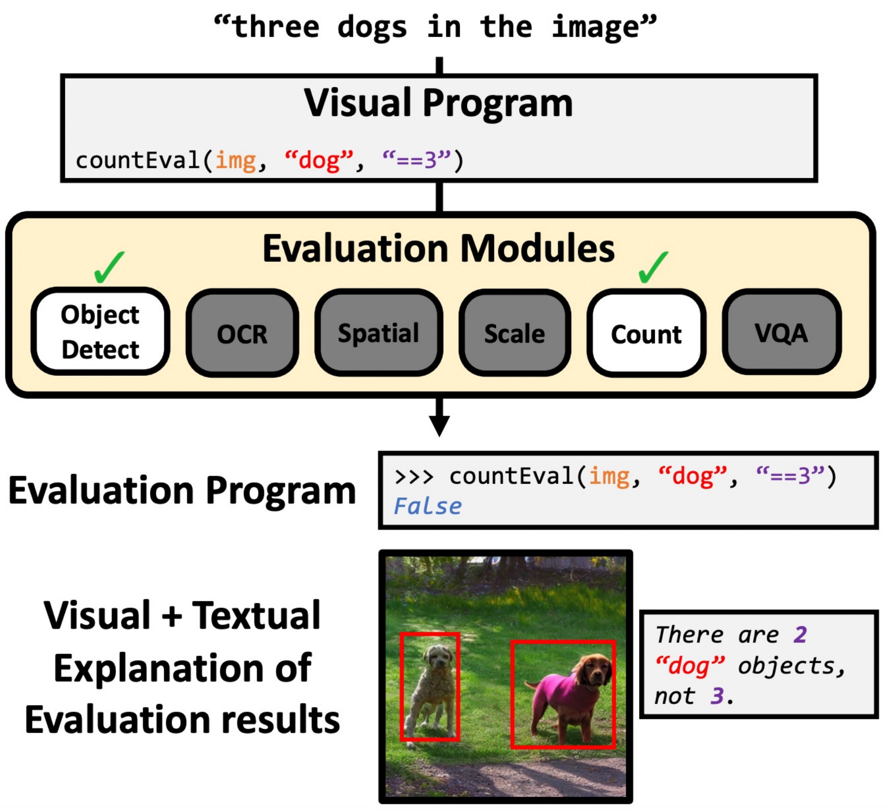
https://envgen-llm.github.io/ [Zala et al., COLM 2024]

Talk Outline

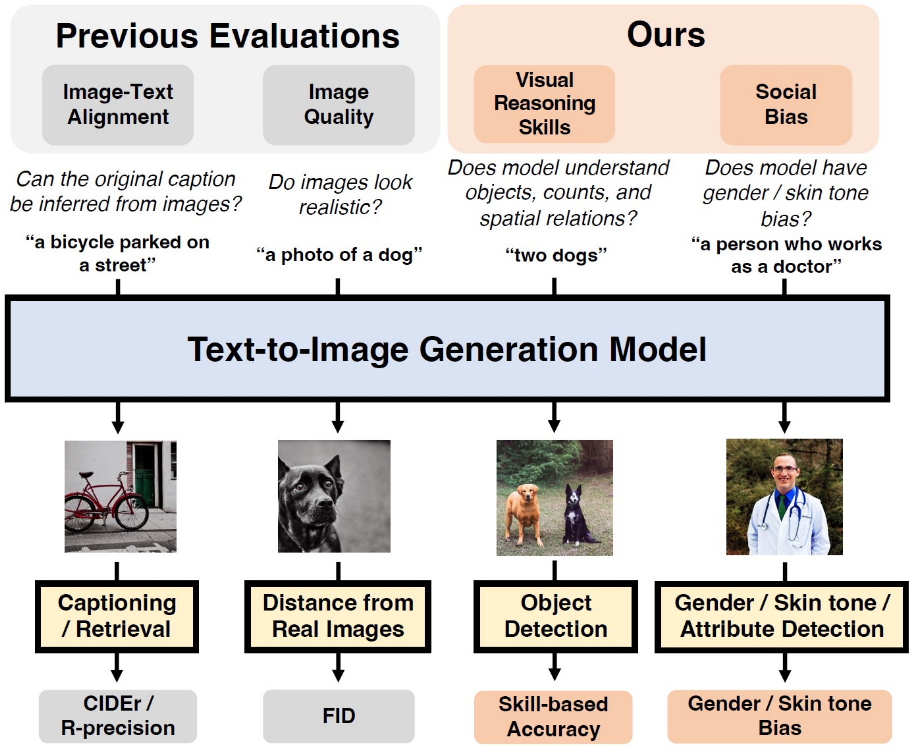
A journey of multimodal generative models for enhancing their unification, interpretable planning/programming, evaluation:

- **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 [CVPR 2024]
- **Interpretable Multimodal Generation via LLM Planning/Programming Agents** (for Understanding, Control, Faithfulness, OOD)
  - VPGen: Step-by-Step Text-to-Image Generation with Interpretable Visual Programming [NeurIPS 2023]
  - VideoDirectorGPT: Consistent Multi-Scene Video Generation via LLM-Guided Planning [COLM 2024]
  - DiagrammerGPT: Generating Diagrams via LLM Planning [COLM 2024]; EnvGen: Adapting Environments via LLMs for Training Embodied Agents [COLM 2024]
- **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 [ICLR 2024]
- **Next Big Challenges**: trade-offs, structure, non-verbal, interaction, reasoning, causality, long-distance fine-grained evaluation, efficiencies

Part 3: Evaluation of Multimodal Generation



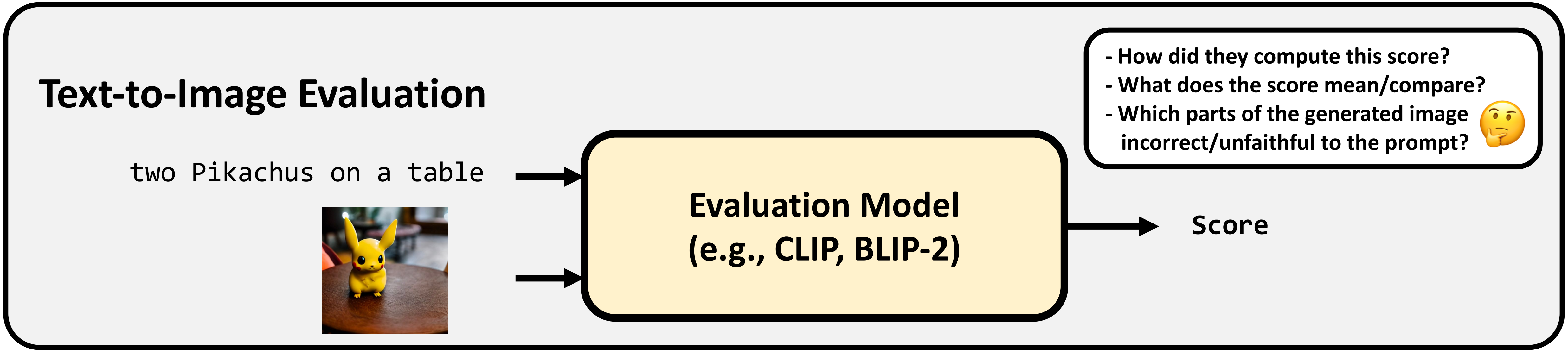
VPEval (NeurIPS 2023)



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

VPEval: Visual Programming for Explainable T2I Evaluation



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

[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)
  ...
```

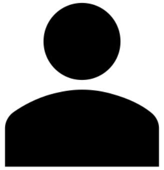
VQA Eval

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

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

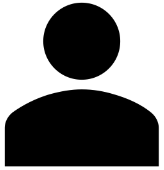
https://vp-t2i.github.io/

[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');
...
```

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

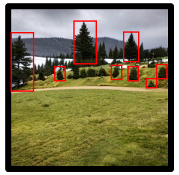
```
# Generated Program
objectEval(image, 'ram');
objectEval(image, 'evergreen');
countEval(image, 'ram', '>1');
countEval(image, 'evergreen', '==1');
vqa(image, 'what is in the foreground?', 'grassy
slope,beach,field,forest', 'grassy slope');
...
```

Visual + Textual Explanations of Errors/Hallucinations

Incorrect ❌

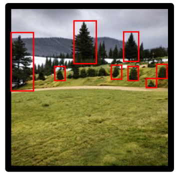
no “ram” object found.

Correct ✅



“evergreen” object found.

Incorrect ❌



there are 8 “evergreen” objects, not 1.

Correct ✅

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

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

[Cho et al., NeurIPS 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., ICLR 2024]

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

Questions w/ desired properties (following Davidsonian formal semantics):

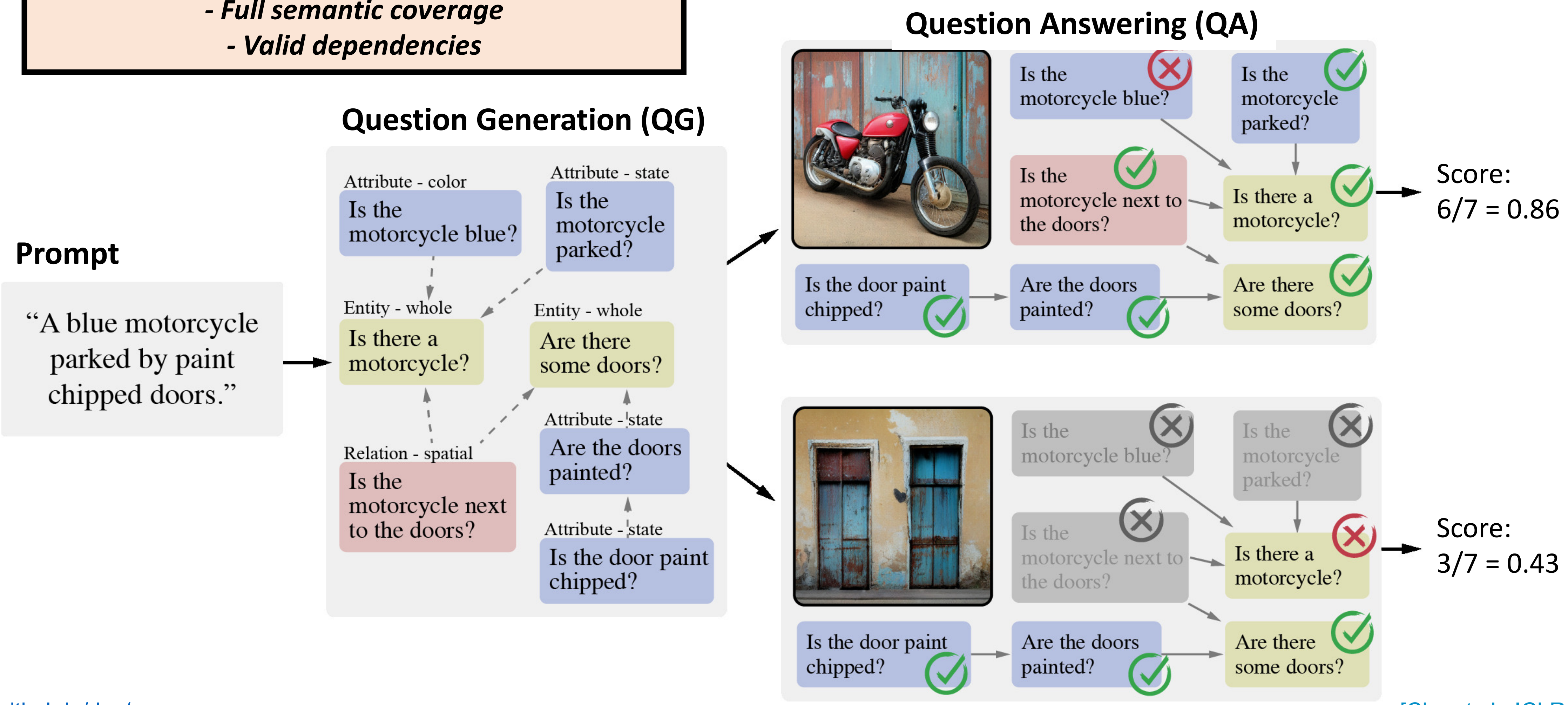
- Atomic

- Unique

- Full semantic coverage

- Valid dependencies

Answering Questions, while avoiding answering the invalid questions



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

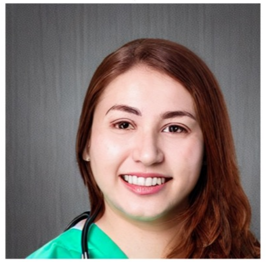
[Cho et al., ICLR 2024]

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

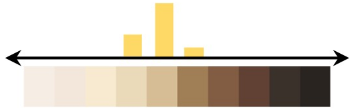
"a **person** who works as a **nurse**" diagnostic prompts

Text-to-Image Generation Model



Automated Detection + Human Evaluation

Gender | **Skin Tone** | **Attribute**



Measure Average and Variance of Distribution

How skewed are the distributions?

[Cho et al., ICCV 2023]

Conclusion + Big Challenges / Research Directions

- **Trade-off** of blackbox **pretraining** vs. **modular structure** (incl. faithfulness, efficiency, interpretability/understanding, human-in-loop/control, OOD, 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** when new/unseen information keeps coming?
- **Unlearning** of outdated/wrong/unsafe/private information?
- **Efficiency** w.r.t. many axes: 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!