Scalable and Faithful Multimodal Reasoning Frameworks

Multimodal AI models must be intelligent and trustworthy. Recent multimodal AI models such as ChatGPT and Gemini are powerful in many tasks (e.g., generating a high-quality image) based on the scaling of model parameters and training data. However, these models still struggle with tasks that require complex reasoning (e.g., generating informative presentation slides) and suffer from hallucinations. To address this challenge, my research aims to **improve the faithfulness in complex reasoning, in both understanding and generation tasks, while maintaining scalability**. My research has earned spotlight/oral awards at top AI conferences (e.g., NeurIPS) and been recognized through a Bloomberg PhD Fellowship and media coverage (e.g., MIT Tech Review, WIRED, IEEE Spectrum). In the pursuit of next-generation multimodal AI, I work on the following research topics: **Scalable Multimodal Frameworks** (Sec. 1), **Faithful Multimodal Reasoning** (Sec. 2), and **Evaluation and Refinement of Multimodal Generation** (Sec. 3).

1 Scalable Multimodal Frameworks

Modern AI models face an ever-growing variety of user queries across different modalities and tasks. Traditional approaches, which rely on specialized architectures tailored to specific tasks, now face scalability challenges. Meeting the growing demand for thousands of capabilities would require developing and maintaining an equally vast number of models. My research has addressed this challenge by introducing (a) **unified generative frameworks** that flexibly accommodate diverse modalities and tasks, enabling users to train a wide range of capabilities using a single architecture and generative objective [1; 2], and (b) **efficient finetuning frameworks** that significantly reduce parameter and memory requirements for creating task-specific models [3; 4; 5].

Unified generative frameworks. In **VL-T5** [1], I introduced a unified framework for multimodal tasks, enabling direct handling of diverse tasks through text generation (Fig. 1). By representing task descriptions and outputs as text, this framework simplifies task formulation and facilitates the seamless adoption of large language models for multimodal tasks. Widely adopted as a standard training framework (**500+ citations**), it has influenced numerous multimodal AI models, including SimVLM, CoCa, Flamingo, PaLI, Parti (Google), BLIP 1 / 2 (Salesforce), Unified-IO 1 / 2 (AI2), and OFA (Alibaba). In **X-LXMERT** [2], I introduced a novel text-to-image generation framework representing images as $N \times N$ grids of 'image tokens.' Using a multimodal language model, the framework generates image tokens progressively through iterative denoising (Fig. 2). The proposed text-to-image generation framework has received wide attention (**100+ citations**) and become standard practice in state-of-the-art models, including DALL-E (OpenAI) and Parti / Muse (Google). This work was also featured in MIT Technology Review.

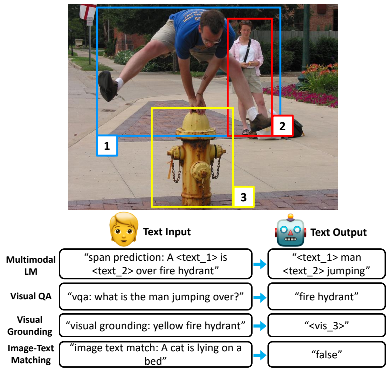

1
2
3
Text Input
Text Output
Multimodal LM
"span prediction: A <text_1> is <text_2> over fire hydrant"
"<text_1> man <text_2> jumping"
Visual QA
"vqa: what is the man jumping over?"
"fire hydrant"
Visual Grounding
"visual grounding: yellow fire hydrant"
"<vis_3>"
Image-Text Matching
"image text match: A cat is lying on a bed"
"false"

Figure 1: **VL-T5**: Unifying multimodal tasks via text generation [1]

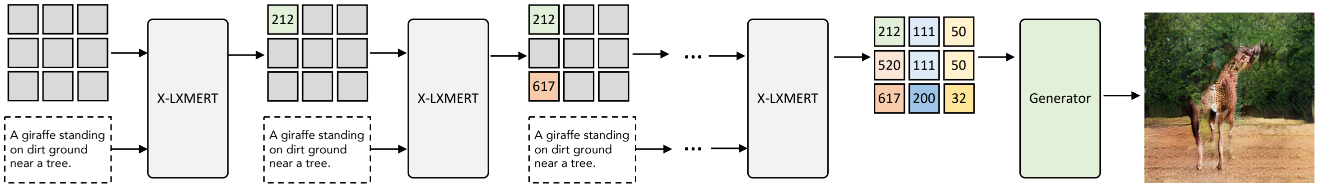


Figure 2: **X-LXMERT**: Image generation as iterative image token unmasking (i.e., denoising) [2]

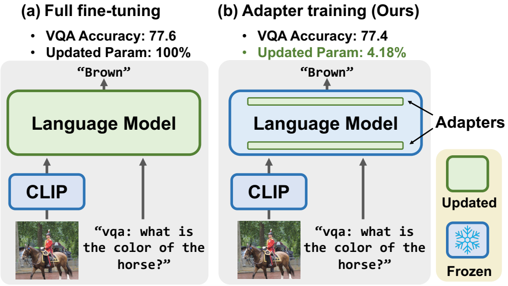
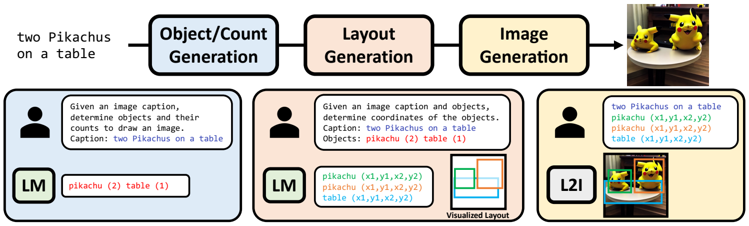


Figure 3: Efficient finetuning [3; 4; 5]

Efficient finetuning frameworks. I introduced efficient finetuning frameworks for various multimodal tasks, where we can train only small parameters to leverage the knowledge of pretrained large foundational models to obtain task-specific models. In **VL-Adapter** [3], I proposed inserting and updating a small set of adapter parameters within pretrained language models to obtain multimodal language models that can handle various multimodal tasks (Fig. 3). Building on this work, in **Ladder Side-Tuning** [4], I separated the adapter parameters from the backbone language models in a side network, which provides training memory efficiency in addition to parameter efficiency. In **Ctrl-Adapter** [5], I proposed a framework that efficiently provides diverse controls to any image/video diffusion model by adapting pretrained ControlNets, significantly reducing training time and memory requirements. These efficient finetuning frameworks have been widely adopted in various works (**500+ citations in total**), including ControlNet (Stanford), BLIP-2 (Salesforce), and Flamingo (Google).

2 Faithful Multimodal Reasoning

The current learning paradigm for multimodal models – relying on a single black-box model that encodes all knowledge within its parameters – faces significant limitations. Despite the vast amounts of data and computation used during training, modern multimodal models often struggle with basic reasoning tasks and frequently produce factually incorrect content. My research addresses these challenges by introducing **(a) planning-based visual generation frameworks** that decompose complex problems into faithful, human-interpretable step-by-step reasoning processes [6; 7; 8] and **(b) retrieval-augmented generation frameworks** that enhance accuracy by retrieving relevant information before generating outputs [9; 10].



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

pikachu (2) table (1)

Given an image caption and objects, determine coordinates of the objects.
Caption: two Pikachus on a table
Objects: pikachu (2) table (1)

pikachu (x1,y1,x2,y2)
pikachu (x1,y1,x2,y2)
table (x1,y1,x2,y2)

two Pikachus on a table
pikachu (x1,y1,x2,y2)
pikachu (x1,y1,x2,y2)
table (x1,y1,x2,y2)

Figure 4: Planning-based visual generation [6; 7; 8]

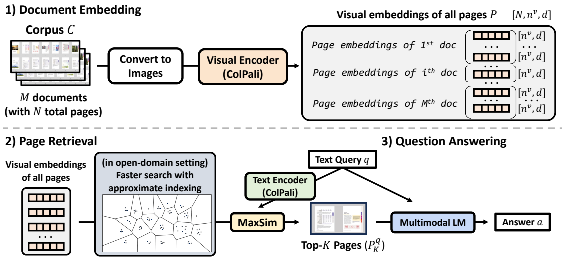

1) Document Embedding
Corpus C
M documents (with N total pages)
Convert to Images
Visual Encoder (ColPali)
Visual embeddings of all pages P $[N, n^v, d]$
Page embeddings of 1st doc
Page embeddings of ith doc
Page embeddings of Mth doc
$[n^v, d]$
2) Page Retrieval
Visual embeddings of all pages
(in open-domain setting) Faster search with approximate indexing
MaxSim
3) Question Answering
Text Query q
Text Encoder (ColPali)
Top-K Pages (P_K^q)
Multimodal LM
Answer a

Figure 5: Retrieval augmentation [9; 10]

Planning-based visual generation. I introduced novel frameworks that break the bottleneck of using a single model for complex visual generation tasks and explicitly perform step-by-step multimodal planning to improve the faithfulness of the generation process. In **VPGen** [6], I introduced a two-stage framework for text-to-image generation (Fig. 4), where a model first predicts a concrete plan (e.g., what objects to generate / how many objects to generate / where to generate them) then creates an actual image based on the plan, significantly improving the semantic layouts of the generated images. In the follow-up works, I demonstrated that the planning-based framework could also improve the generation of long videos (**VideoDirectorGPT** [7]) and scientific diagrams (**DiagrammerGPT** [8]). These planning-based generation frameworks have received wide attention (**90+ citations in total**), including many extensions in the image, video, and motion generation research communities.

Multimodal retrieval-augmented generation. Relying solely on large model parameters for reasoning increases both computational demands and the likelihood of generating inaccurate information. To this end, I introduced **HiREST** [9] and **M3DocRAG** [10], two novel multimodal retrieval-augmented generation (RAG) frameworks that retrieve knowledge from large multimodal data storage (e.g., thousands of videos or documents) to augment the downstream models, making the reasoning steps more faithful and efficient (Fig. 5). Since ensuring the factuality of the generated content is crucial when developing AI assistants in any domain, these frameworks have received wide attention, with inquiries and outreach from many companies in diverse domains, including hedge funds, medical institutions, and law firms.

3 Evaluation and Refinement of Multimodal Generation

With recent multimodal generative models demonstrating significant advancements, conventional evaluation metrics have been often saturated and no longer provide meaningful insights into future research direction. Thus, it is crucial to develop benchmarks that assess their new capabilities and address their limitations across diverse, practical scenarios. To this end, I introduced (a) **fine-grained evaluation frameworks** that comprehensively measure generation skills in multiple dimensions to uncover detailed strengths and weaknesses [11; 6; 12; 13; 14; 9; 10], and (b) **automatic model refinement frameworks** that use these evaluations to detect weaknesses, use the feedback to refine the reasoning process, and enhance model faithfulness [15; 16; 17; 18].

Fine-grained evaluation. As modern text-to-image (T2I) models achieve photorealistic image generation, traditional standard metrics like FID are no longer sufficient to assess their broader capabilities. In **DALL-Eval** [11], I introduced two critical evaluation criteria to the field: visual reasoning skills and social biases (Fig. 6). This work revealed that even cutting-edge models struggle with tasks like object counting and spatial reasoning while exhibiting gender and skin tone biases. Building upon this, in **VPEval** [6], **DSG** [12], and **DOCCI** [13], I expanded evaluations to include advanced skills like 3D spatial reasoning, text rendering, and understanding paragraph-level prompts (Fig. 7). Beyond text-to-image generation, I developed benchmarks for diverse multimodal tasks such as image captioning (**FineCapEval** [14]), layout-to-image generation (**LayoutBench** [19]), video information retrieval (**HiREST** [9]), and document understanding (**M3DocVQA** [10]). These evaluation frameworks have garnered significant attention (**400+ citations in total**), including Parti, Imagen 1/3 (Google), OPT2I (Meta), HEIM (Stanford), TIFA (UW), VQAScore (CMU) and media coverage (e.g., IEEE Spectrum, WIRED). My expertise also led to an invitation to participate in the red-teaming process for DALL-E 2, where I worked with OpenAI to identify and mitigate biases in the model (e.g., only generating female images for a prompt 'nurse').

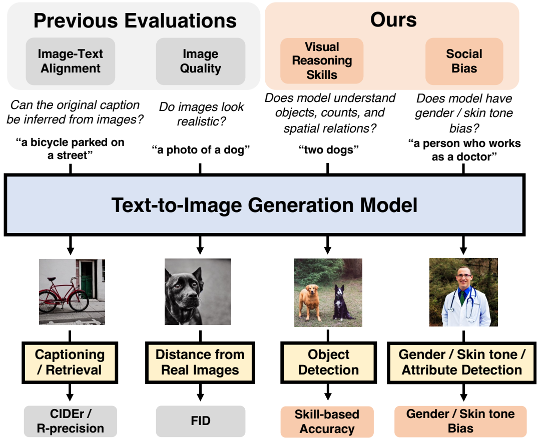
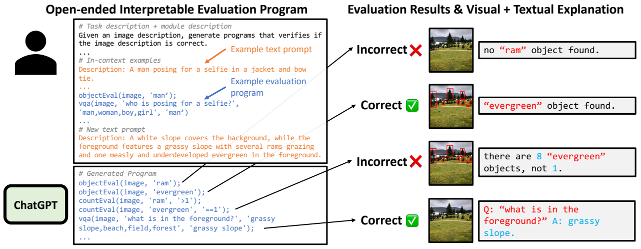

Previous Evaluations
Image-Text Alignment
Image Quality
Ours
Visual Reasoning Skills
Social Bias
Can the original caption be inferred from images?
Do images look realistic?
Does model understand objects, counts, and spatial relations?
Does model have gender / skin tone bias?
"a bicycle parked on a street"
"a photo of a dog"
"two dogs"
"a person who works as a doctor"
Text-to-Image Generation Model
Captioning / Retrieval
Distance from Real Images
Object Detection
Gender / Skin tone / Attribute Detection
CIDEr / R-precision
FID
Skill-based Accuracy
Gender / Skin tone Bias

Figure 6: **DALL-Eval**: The first T2I benchmark of reasoning skills and social biases [11]


Open-ended Interpretable Evaluation Program
Example text prompt
Example evaluation program
ChatGPT
Evaluation Results & Visual + Textual Explanation
Incorrect ❌
Correct ✅
Incorrect ❌
Correct ✅

```
# 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.
```

```
# 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');
...
```

```
no "ram" object found.
```

```
"evergreen" object found.
```

```
there are 8 "evergreen"
objects, not 1.
```

```
Q: "what is in the
foreground?" A: grassy
slope.
```

Figure 7: Fine-grained evaluation reveals detailed strengths and weaknesses of models [6; 11; 12]

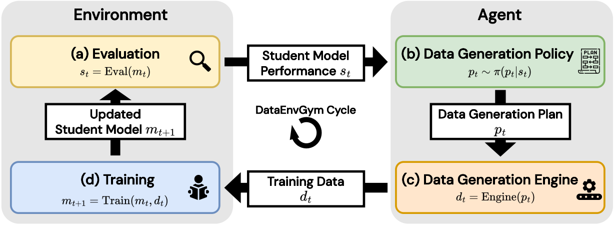

Environment
(a) Evaluation
$s_t = \text{Eval}(m_t)$
Student Model Performance s_t
Agent
(b) Data Generation Policy
$p_t \sim \pi(p_t|s_t)$
Updated Student Model m_{t+1}
DataEnvGym Cycle
Data Generation Plan
p_t
(d) Training
$m_{t+1} = \text{Train}(m_t, d_t)$
Training Data
d_t
(c) Data Generation Engine
$d_t = \text{Engine}(p_t)$

Figure 8: Closing the loop via automatic refinement using evaluation as feedback [15; 16; 17; 18]

Closing the loop via automatic refinement. Improving AI models has traditionally relied on manual efforts, where humans analyze model weaknesses, create additional training data, or modify architectures. To automate and streamline this process, I developed automated refinement frameworks that use the aforementioned fine-grained evaluations [6; 11; 12] to detect model weaknesses and take corrective actions, such as generating additional training data or re-producing outputs with detailed error feedback. I have demonstrated the usefulness of these frameworks in diverse domains, including embodied AI (**EnvGen** [15]), mathematics, visual question answering, programming (**DataEnvGym** [16]), image generation [17], and video generation (**VideoRepair** [18]).

4 Future Work

Improving reasoning is key to advancing multimodal models, making them more intelligent and trustworthy; however, many challenges remain. For example, further scaling model sizes has become prohibitively expensive, and most multimodal models rely on language models, whose text-based reasoning may be suboptimal for multimodal tasks. My lab will address current challenges and push the boundaries of reasoning capabilities in multimodal models, tackling problems that current models are far from solving. At Johns Hopkins University, I am excited about the opportunity to collaborate with exceptional faculty members with expertise in AI and its applications in diverse domains, including Benjamin Van Durme, Philipp Koehn, Daniel Khashabi, Jason Eisner, and many others.

Improving multimodal reasoning beyond model scaling. While AI research in the past decade has largely focused on scaling model sizes, further scaling of models is becoming prohibitively expensive. My future work will study effective techniques for multimodal reasoning empowering smaller models to tackle complex understanding and generation tasks – a challenge akin to enabling the brain to think longer rather than making it bigger. Specifically, I will investigate **multimodal scratchpads**, where models can concretize reasoning processes via generating auxiliary latent variables, building on my work in planning-based visual generation [6; 7; 8]. While current scratchpad approaches (e.g., 'chain-of-thought') primarily target mathematical and text-based reasoning, I aim to explore non-textual, multimodal representations that are better suited for multimodal tasks. These scratchpads will also be guided with **multimodal rewards**, building on my expertise from CLIP-reward [14], the first work to use CLIP for training multimodal language models. Furthermore, I will develop models capable of **reflecting and refining their own previous reasoning processes**, drawing on my prior work on iterative refinement [2; 19; 18]. Lastly, I will explore novel methods for **combining the expertise of multiple models**, building on my research about using external guidance models [20; 18], visual programming [6; 7; 8], and model merging [17].

Scalable visual representation for truly multimodal world model. While parameter scaling has driven significant advances in language models, equivalent progress has not been achieved for visual representations. Current multimodal models are often described as 'language models that can see," relying on a large language model architecture paired with a shallow visual encoder. To pave the way for truly multimodal models capable of seamlessly understanding and generating content across different modalities, we need **a new class of visual representations that scale more effectively than existing approaches (e.g., 2D grids for images)**. Building on my expertise in developing compact latent vision-language representations [21], I aim to explore more scalable visual representation for the transition toward a truly multimodal world model.

References

[1] **Jaemin Cho**, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In ***ICML***, 2021.

[2] **Jaemin Cho**, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, and Aniruddha Kembhavi. X-lxmert: Paint, caption and answer questions with multi-modal transformers. In ***EMNLP*** *(long)*, 2020.

[3] Yi-Lin Sung, **Jaemin Cho**, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In ***CVPR***, 2022.

[4] Yi-Lin Sung, **Jaemin Cho**, and Mohit Bansal. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. In ***NeurIPS***, 2022.

[5] Han Lin*, **Jaemin Cho***, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. 2024.

[6] **Jaemin Cho**, Abhay Zala, and Mohit Bansal. Visual programming for text-to-image generation and evaluation. In ***NeurIPS***, 2023.

[7] Han Lin, Abhay Zala, **Jaemin Cho**, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. In ***COLM***, 2024.

[8] Abhay Zala, Han Lin, **Jaemin Cho**, and Mohit Bansal. Diagrammergpt: Generating open-domain, open-platform diagrams via llm planning. In ***COLM***, 2024.

[9] Abhay Zala*, **Jaemin Cho***, Satwik Kottur, Xilun Chen, Barlas Oğuz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In ***CVPR***, 2023.

[10] **Jaemin Cho**, Ozan İrsoy, Debanjan Mahata, Yujie He, and Mohit Bansal. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. 2024.

[11] **Jaemin Cho**, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In ***ICCV***, 2023.

[12] **Jaemin Cho**, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In ***ICLR***, 2024.

[13] Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, **Jaemin Cho**, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. DOCCI: Descriptions of Connected and Contrasting Images. In *ECCV*, 2024.

[14] **Jaemin Cho**, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with clip reward. In ***Findings of NAACL*** *(short)*, 2022.

[15] Abhay Zala*, **Jaemin Cho***, Han Lin, Jaehong Yoon, and Mohit Bansal. Envgen: Generating and adapting environments via llms for training embodied agents. In ***COLM***, 2024.

[16] Zaid Khan, Elias Stengel-Eskin, **Jaemin Cho**, and Mohit Bansal. Dataenvgym: Data generation agents in teacher environments with student feedback. 2024.

[17] Jialu Li*, **Jaemin Cho***, Yi-Lin Sung, Jaehong Yoon, and Mohit Bansal. Selma: Learning and merging skill-specific text-to-image experts with auto-generated data. In ***NeurIPS***, 2024.

[18] Daeun Lee, Jaehong Yoon, **Jaemin Cho**, and Mohit Bansal. Videorepair: Improving text-to-video generation via misalignment evaluation and localized refinement. 2024.

[19] **Jaemin Cho**, Linjie Li, Zhengyuan Yang, Zhe Gan, Lijuan Wang, and Mohit Bansal. Diagnostic benchmark and iterative inpainting for layout-guided image generation. In *The First Workshop on the Evaluation of Generative Foundation Models (**oral**)*, 2024.

[20] David Wan, **Jaemin Cho**, Elias Stengel-Eskin, and Mohit Bansal. Contrastive region guidance: Improving grounding in vision-language models without training. In ***ECCV***, 2024.

[21] Zineng Tang*, **Jaemin Cho***, Jie Lei, and Mohit Bansal. Perceiver-vl: Efficient vision-and-language modeling with iterative latent attention. In ***WACV***, 2023.