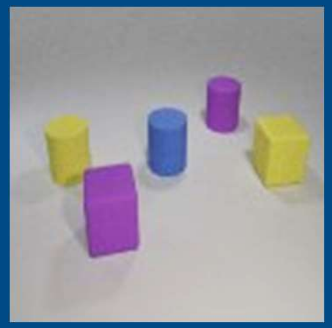
Textual-Visual Logic Challenge: Understanding and Reasoning in Text-to-Image Generation

Peixi Xiong, Mike Kozuch, Nilesh Jain
Intel Labs

Conference: ECCV-2024, 10/2024

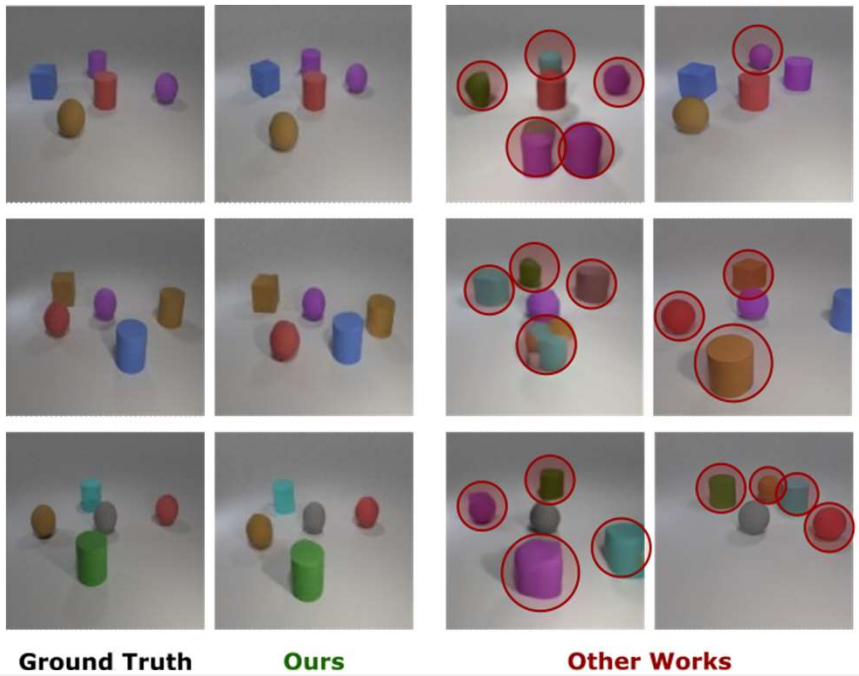


Add a **blue cylinder** at the center . Add a **purple cylinder** behind **it** on the right . Add a **yellow cylinder** in front of **it** on the left and in front of the **blue cylinder** on the left . Add a **yellow cube** in front of the **purple cylinder** on the right and in front of the **blue cylinder** on the right . Add a **purple cube** in front of the **purple cylinder** on the left and in front of the **blue cylinder** on the left .

Introduction

Current Challenges - 1

- **Methods: Cannot understand complex input text prompt**
  - SOTAs simply stack entities based on their labels and attributes, **struggling to capture the relationships embedded in the text**.



[Text] Add a red cylinder at the center . Add a brown sphere in front of it on the left . Add a blue cube behind it on the left and behind the red cylinder on the left . Add a purple cylinder behind it on the right and behind the brown sphere on the right . Add a purple sphere in front of it on the right and behind the red cylinder on the right .

[Text] Add a purple sphere at the center . Add a brown cylinder in front of it on the right . Add a red sphere in front of it on the left and in front of the purple sphere on the left . Add a brown cube behind the brown cylinder on the left and behind the purple sphere on the left . Add a blue cylinder in front of it on the right and in front of the purple sphere on the right .

[Text] Add a gray sphere at the center . Add a cyan cylinder behind it on the left . Add a green cylinder in front of it on the right and in front of the gray sphere on the left . Add a brown sphere behind it on the left and in front of the gray sphere on the left . Add a red sphere behind it on the right and behind the green cylinder on the right .

Figure 2. Structural information is important in the domain of text-to-image generation. Current research often lacks the capacity to represent the intricate relations present in text prompts, particularly when these prompts are lengthy and laden with detailed content.

Introduction

Current Challenges - 2

- **Evaluation Metrics: Lacks structural assessment**
  - Current metrics might not always align with human judgments of image quality.

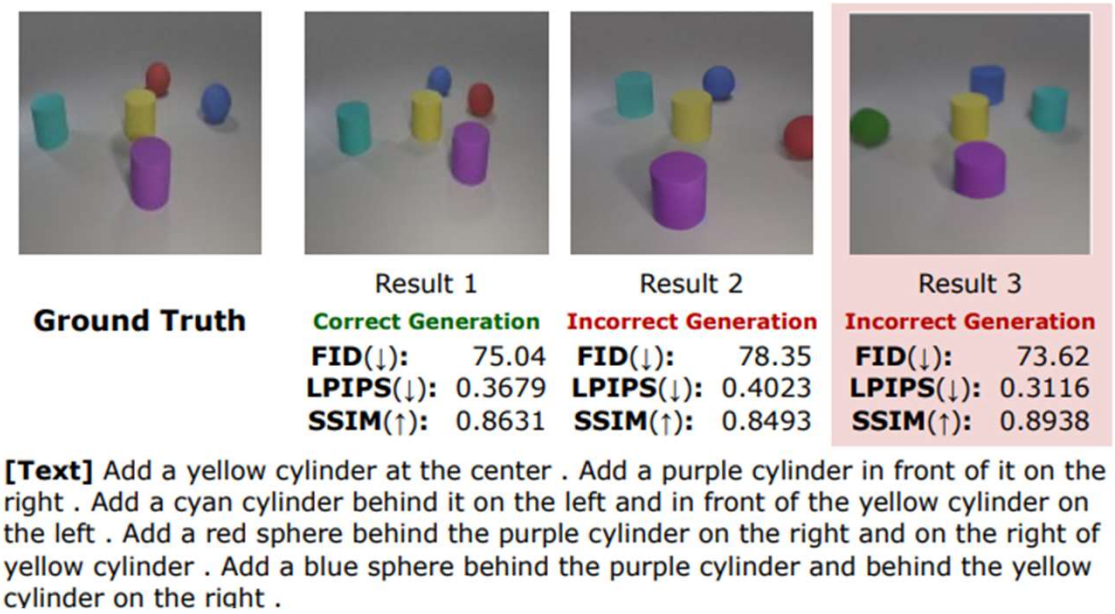


Figure 3. Limitations of existing evaluation metrics in accurately assessing model performance. From an evaluative aspect, the absence of a comprehensive structural assessment can result in evaluations that diverge from human assessments of image quality and relevance.

Approach

Problem Statement

- **We introduce a task named Logic-Rich Text-to-Image Generation (LRT2I):**
  - Diverges from conventional text-to-image tasks, highlighting the limitations in semantic understanding.
  - Addresses the complex relational challenges in text prompts.
    - Emphasizes reasoning in text-to-image generation, focusing on inferring relations and entities.
    - Concentrates on spatial relations and entity inference.
    - Defines "logic" as a structured and systematic reasoning process.

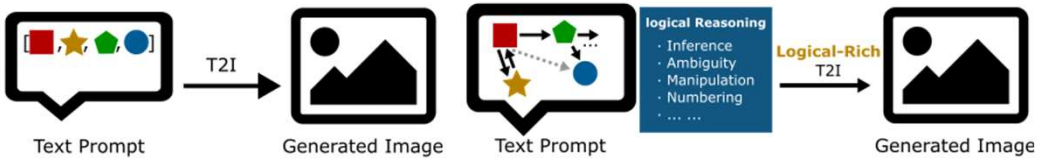


Figure 5. Comparison between conventional text-to-image generation and our LRT2I task.

Add a **blue cylinder** at the center . Add a **purple cylinder** behind **it** on the right . Add a **yellow cylinder** in front of **it** on the left and in front of the **blue cylinder** on the left . Add a **yellow cube** in front of the **purple cylinder** on the right and in front of the **blue cylinder** on the right . Add a **purple cube** in front of the **purple cylinder** on the left and in front of the **blue cylinder** on the left .

Figure 4. Logic-Rich Text-to-Image Generation, a variant that diverges from traditional task where models generate RGB images from short, semantically simple text prompts. In our refined task, the model handles text inputs that are not only lengthy but also semantically rich, featuring multiple interconnected entities with annotated attributes.

Approach

Novel Dataset

- **We collected a novel dataset (TV-Logic) comprising 15,213 samples.**
  - Each sample includes a **long, content-rich text prompt** and its corresponding images.
  - To assess the degree of reasoning required, we have established six categories for the logical-rich text-to-image generation (LRT2I) task.

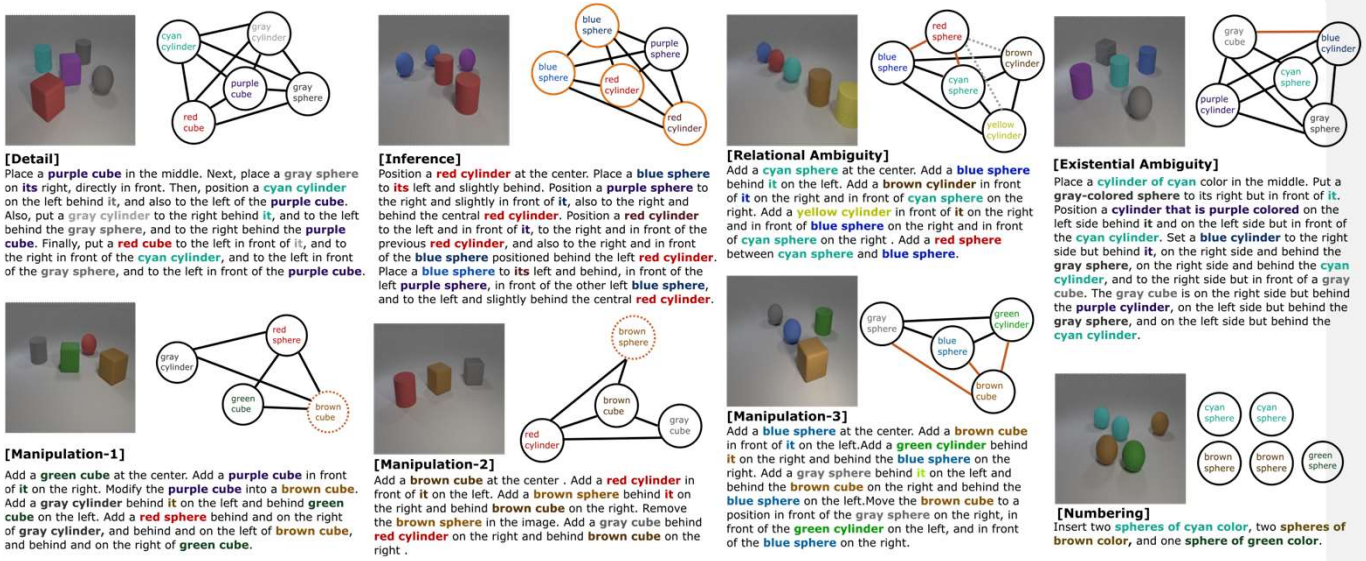


Figure 6. Overview of TV-Logic Dataset Categories. This composite image illustrates the diverse challenges in text-to-image generation, showing six categories for model evaluation. These categories demonstrate the diverse challenges in text-to-image generation. Within each category, the graph depicts scene information derived from text prompts. Solid edges signify relations explicitly mentioned in the prompts, while dashed edges indicate unmentioned relations. Orange lines represent case-related information.

Approach

Novel Dataset

- **More diverse entities in images.**
- **Longer text prompt.**
- **More reasoning.**

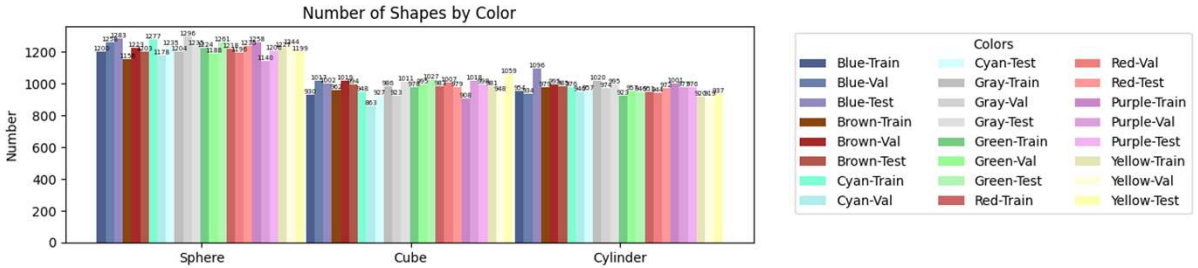


Figure 7. Statistical Overview of Shape-Attribute Composition. The colors of the bars correspond to the entity colors, with different shades representing the training, validation, and test splits.

Detail				Inference				Relational Ambiguity			
Max	Min	Avg	Std	Max	Min	Avg	Std	Max	Min	Avg	Std
14/30	4/6	4.66/16.74	1.31/3.75	13/35	4/6	5.92/19.85	1.58/3.97	10/30	4/5	4.22/13.03	0.72/4.90
Existential Ambiguity				**Manipulation**				**Numerical Rep**			
Max	Min	Avg	Std	Max	Min	Avg	Std	Max	Min	Avg	Std
13/30	4/6	5.12/17.09	1.62/3.56	10/21	4/4	4.05/13.27	0.37/3.37	3/0	0/0	0.01/0	0.10/0

TV-Logic			
Max	Min	Avg	Std
14/35	0/0	3.56/13.18	2.61/7.33

Table 2. Statistical Overview of Reasoning in Prompts. Two measures: cross-sentence object reference counts before slash and relation mentions, underlined after slash.

	Average	Std Dev	Max	Min
MS-COCO	10.61	2.43	179	8
SR_{2D}	6.56	1.58	10	2
TV-Logic	89.17	38.32	192	13

Table 1. Statistics of the MS-COCO, SR_{2D}, and TV-Logic datasets. It displays the average, standard deviation, maximum, and minimum values of the prompt lengths.

Approach

Network Architecture

- **Relation Understanding Module**
  - Encourage the text encoder to pay more attention on the relation-related tokens in generation phase.
- **Multimodality Fusion Module**
  - Learn a better text understanding (self-attention) and visual-text alignment (cross-attention).
- **Negative Pair Discriminator**
  - Penalize the model if it cannot handle the minor disturbance on the informative tokens.

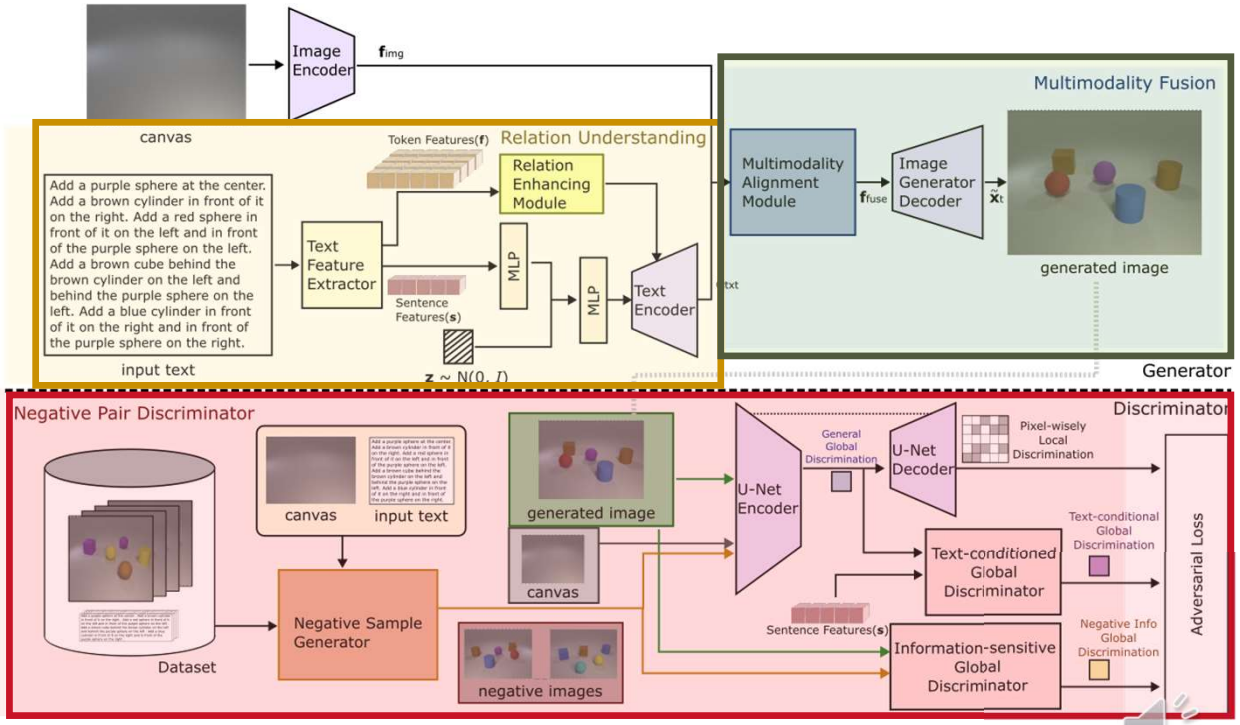


Figure 8. Overview of our approach.

Experiment Results

Figure 9. Qualitative comparisons on the TV-Logic dataset. The columns, from left to right, display the ground truth image, our model's results, results from other works, and the corresponding input text prompts.

Experiment Results

Quantitative Evaluation Metrics

- Borrow the concepts from image manipulation task.
  - Average Precision (AP), Average Recall (AR), F1 score, and relational similarity (RSIM).
  - $RSIM\left(E_{G_{\mathrm{gt}}}, E_{G_{gen}}\right) = recall \times \frac{|E_{G_{\mathrm{gt}}} \cap E_{G_{\mathrm{gt}}}|}{|E_{G_{\mathrm{gt}}}|}$
    - $E_{G_{\mathrm{gt}}}$ is the set of relational edges for the ground-truth image that correspond to vertices that are common to both images.
    - $E_{G_{gen}}$ is the set of relational edges for the generated image that correspond to the vertices in common to both images.

Experiment Results

Quantitative Results

No.	Name	Focal Info	Focal Rel	2Modal Ref	FID↓	LIPIPS↓	SSIM↑	PSNR↑	AP↑	AR↑	F1↑	RSIM↑
1	DALL-E	-	-	-	79.33	0.3963	0.8146	12.45	19.64	16.31	17.05	8.41
2	U-Vit	-	-	-	55.19	0.2908	0.8932	21.46	28.01	26.61	26.98	20.94
3	GeNeVA-GAN	-	-	-	49.06	0.3039	0.9199	22.09	16.61	13.63	14.39	7.47
4	LatteGAN	-	-	-	48.81	0.1294	0.9275	24.85	66.65	69.29	67.10	63.03
5	Composable Diff	-	-	-	53.93	0.3267	0.8756	18.84	45.45	25.55	30.75	14.29
6	GPT+Blender	-	-	-	77.38	0.3656	0.8984	18.00	37.84	21.24	26.15	10.33
7	Ours	✗	✓	✓	45.62	**0.1243**	0.9282	24.90	72.50	75.00	73.72	69.91
8	Ours	✓	✗	✓	**39.41**	0.1284	0.9282	24.90	70.53	73.65	71.24	66.12
9	Ours	✓	✓	✗	42.57	0.1254	0.9299	25.04	71.81	74.35	72.75	68.90
10	Ours	✓	✓	✓	40.60	0.1250	**0.9303**	**25.08**	**73.62**	**76.24**	**74.13**	**72.78**

Figure 10. Quantitative analysis and ablation study comparisons. Baseline and proposed methods are benchmarked on the TV-Logic dataset using eight evaluation metrics: average precision (AP), average recall (AR), F1 score, relational similarity (RSIM), Fréchet Inception Distance (FID), Learned Perceptual Image Patch Similarity (LPIPS), Structural Similarity Index Measure (SSIM), and Peak Signal-to-Noise Ratio (PSNR).

Methods	Detail				Inference				Relational Ambiguity			
	AP	AR	F1	RSIM	AP	AR	F1	RSIM	AP	AR	F1	RSIM
DALL-E	23.74	15.22	18.08	7.92	16.41	18.23	16.67	9.80	20.41	16.71	17.68	8.81
U-ViT	20.59	17.38	18.66	9.32	14.11	14.97	14.30	8.49	16.41	16.32	16.10	8.65
GeNeVA-GAN	21.23	11.98	15.06	6.57	12.06	13.25	12.34	7.47	16.96	16.61	16.41	9.01
LatteGAN	78.43	74.32	76.06	68.62	61.37	76.18	66.90	64.07	59.87	58.99	58.47	51.77
Composable Diff	55.31	25.11	33.01	12.50	47.28	34.67	37.56	17.38	43.64	23.08	28.16	11.70
GPT+Blender	45.67	20.21	27.33	9.57	37.06	26.01	29.67	13.92	30.14	16.64	20.68	8.15
Ours	88.01	85.85	86.79	81.01	70.83	83.57	75.61	71.50	67.92	66.27	66.08	58.04

Methods	Entity Ambiguity				Manipulation				Numerical Rep			
	AP	AR	F1	RSIM	AP	AR	F1	RSIM	AP	AR	F1	RSIM
DALL-E	23.24	14.73	17.56	7.42	18.45	15.67	16.40	8.16	15.47	17.41	15.89	-
U-ViT	19.06	16.23	17.36	9.03	65.48	61.65	62.94	47.95	10.97	11.79	11.17	-
GeNeVA-GAN	20.77	12.21	15.13	7.14	15.00	12.43	13.31	7.10	13.45	15.22	13.95	-
LatteGAN	77.48	73.86	75.41	67.89	71.74	69.82	70.54	64.40	54.58	68.64	59.78	-
Composable Diff	53.00	27.06	34.36	1289	49.36	34.30	38.47	17.56	25.99	11.79	15.43	-
GPT+Blender	42.59	19.30	25.90	9.88	36.31	20.22	25.14	10.42	34.17	27.60	29.72	-
Ours	87.87	85.36	86.45	80.57	78.49	75.95	76.93	71.70	57.33	70.70	62.12	-

Figure 11. Quantitative analysis of the TV-Logic dataset across different subcategories.

Conclusions

- Identify a challenge and introduce a novel task: **logic-rich text-to-image generation**.
  - Highlighting the significance of understanding and reasoning in this domain.
- Benchmark this task by collecting the **Textual-Visual Logic (TV-Logic) dataset**.
  - **The first to target logic-rich reasoning** in this domain.
  - Categorize reasoning in text-to-image generation into **six main categories** to comprehensively evaluate model performance.
- Proposed **a baseline model** with three modules.
  - Enhance text-to-image reasoning and extend the discriminator's role.

Thanks


intel