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

Peixi Xiong[1], Michael Kozuch[1], and Nilesh Jain[1]

Intel Labs, USA
{peixi.xiong, michael.a.kozuch, nilesh.jain}@intel.com

Abstract. Text-to-image generation plays a pivotal role in computer vision and natural language processing by translating textual descriptions into visual representations. However, understanding complex relations in detailed text prompts filled with rich relational content remains a significant challenge. To address this, we introduce a novel task: Logic-Rich Text-to-Image generation. Unlike conventional image generation tasks that rely on short and structurally simple natural language inputs, our task focuses on intricate text inputs abundant in relational information. To tackle these complexities, we collect the Textual-Visual Logic dataset, designed to evaluate the performance of text-to-image generation models across diverse and complex scenarios. Furthermore, we propose a baseline model as a benchmark for this task. Our model comprises three key components: a relation understanding module, a multimodality fusion module, and a negative pair discriminator. These components enhance the model's ability to handle disturbances in informative tokens and prioritize relational elements during image generation. https://github.com/IntelLabs/Textual-Visual-Logic-Challenge

Keywords: Text-to-Image Generation · Structural Reasoning · Relational Understanding

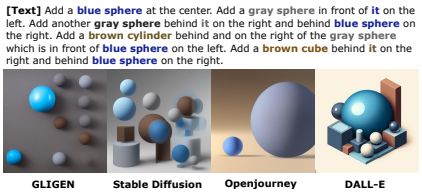
1 Introduction

Text-to-image generation, a crucial field in computer vision and natural language processing, converts textual descriptions into visual representations. The primary challenge here is reasoning through complex relations in detailed text prompts filled with relational content. Effectively tackling this complexity is vital, mirroring human perception that depends on integrating attributes, entities, and relations for precise visual interpretation.

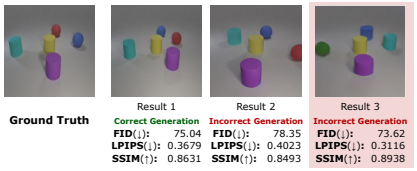
Unfortunately, state-of-the-art systems struggle with relation-rich prompts (Figure 1a), often due to inadequate structural understanding in models. Contributing challenges arise from datasets such as Visual Genome [13] and MS-COCO [15], where relation triplets, representing the relationship between subject-object pairs, often appear identical or similar. This trend restricts the evaluation of relation representation in generated images, given that 75.6% of Visual Genome's triplets consist of single relations. *For example, the relationship*

between "apple" and "plate" is frequently shown as "on". Additionally, existing metrics inadequately capture prompt-intended structures. Richer datasets still face evaluation complexities due to pixel-level ground truth ambiguity. In Figure 1b results 3, though invalid, would receive better scores in current metrics. Metrics like FID [9], SSIM [28], and LPIPS [35] may diverge from comprehensive structural evaluations.

In this paper, we introduce the "Logic-Rich Text-to-Image Generation"(LRT2I) task, focusing on the intricate relational challenges in text prompts, a departure from conventional simpler natural language image generation tasks. We also identify six categories of relational structures that may present particular challenges. To comprehensively evaluate performance across diverse challenges, we collected a new dataset as a benchmark, with well-annotated scene graphs, a set of images corresponding to the text prompts, and a diverse set of logic-rich text prompts tailored to this task. Further, we propose a baseline model adept at accurately representing complex textual relations in images by integrating a GAN framework with components designed for discerning three key factors: informative tokens, deep relational reasoning, and feature alignment across modalities. These factors, we believe, are crucial for enhancing text-to-image generation.



(a) Current works fail to generate structure-related information from text prompts, particularly when logical reasoning is involved.



(b) Limitations of existing evaluation metrics in accurately assessing model performance.

Fig. 1: Structural nuances are crucial in text-to-image generation; however, current research often fails to capture complex text relations, particularly in detailed and logic-rich prompts. The lack of comprehensive structural assessment may lead to evaluations that differ from human judgments of image quality and relevance.

In this work, our primary contributions are :

i. We identify a challenge that has not been fully addressed in discerning structural information and introduce a novel task termed "logic-rich text-to-image generation", highlighting the significance of understanding and reasoning in this domain.

ii. We benchmark this task by collecting the Textual-Visual Logic (TV-Logic) dataset, the first to target logic-rich reasoning in this domain. Additionally, we categorize reasoning in text-to-image generation into six main categories to comprehensively evaluate model performance.

iii. We proposed a baseline model with three modules—Relation Understanding, Multimodality Fusion, and Negative Pair Discriminator—that enhance text-to-image reasoning and extend the discriminator's role.

2 Related Works

2.1 Text-to-Image Generation

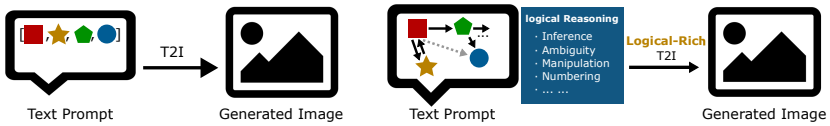
Synthesizing images from text in computer vision and natural language processing poses significant challenges. Techniques include Generative Adversarial Networks (GANs) [1, 4, 30, 34] that generate images from text using adversarial neural networks; Variational Autoencoders (VAEs) [12] that map text to latent spaces for image creation; and Diffusion models [8, 19, 21, 22] that iteratively refine text-based images. Despite advancements, semantic consistency, such as accurate depiction of color, shape, and relations aligned with text prompts, remains a challenge. Traditional evaluation metrics like the Inception Score [23] and Fréchet Inception Distance [9] also may fall short in accurately reflecting human judgment of image quality and relevance [17].

2.2 Structural Information in Computer Vision

In computer vision, especially in text-to-image synthesis, structural information plays a crucial role in handling the complexity of object interrelations, which can affect image generation quality. Researchers propose an intermediate "Scene Layout" to detail object relations and improve outcomes [11, 36]. This methods, however, often need extra resources like a pre-existing graph or additional supervision, leading to more human annotation work when such graphs aren't available [10]. In visual question answering, employing structural graph representations significantly aids in the interpretation of visual data, converting it into structured formats for better analysis [14, 25, 29]. Similarly, in image captioning, graph structures effectively encode object attributes and relations, whether through implicit scene graph representations [7, 33] or explicit ones [18, 32], enhancing the overall quality of the generated captions.

3 Novel Task: Logic-Rich Text-to-Image Generation

3.1 Task Definition



(a) Conventional Text-to-Image Generation **(b)** Logic-Rich Text-to-Image Generation

Fig. 2: Comparison between conventional text-to-image generation and logic-rich text-to-image generation.

We now formally introduce a new task, "Logic-Rich Text-to-Image Generation," which differs from conventional Text-to-Image (T2I) tasks by emphasizing

semantic understanding limitations. Reasoning in text-to-image generation involves inferring relations and entities. Our focus is on spatial relations and entity inference, while "logic" denotes a structured and systematic reasoning process.

	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, VISOR, and TV-Logic datasets. It displays the average, standard deviation, maximum, and minimum values of the prompt lengths.

In conventional tasks, the goal is to map text prompts to a 2D RGB image, optimizing parameters to minimize loss between generated and ground truth images. Our task diverges by focusing on the entities and relations within text, particularly with longer, relation-rich inputs (Figure 2). This approach aligns with human perception, emphasizing the attribute-entity-relation structure crucial for image generation. While datasets like MS-COCO provide detailed annotations for basic training, they lack emphasis on complex relational concepts and reasoning operations central to our task (Table 1). Similarly, the SR$_{2D}$ dataset focuses on spatial relations but overlooks logical operations and language diversity. These gaps highlight the need for an approach that captures a wider range of relational concepts and reasoning operations, aiming to generate images that accurately reflect the complexity of textual prompts.

3.2 Novel Dataset

To better evaluate understanding and reasoning in the text-to-image generation task, we have compiled a novel dataset comprising 15,213 samples. Each sample includes a long, content-rich text prompt and its corresponding images (more detailed information available in the Appendix). To assess the degree of reasoning required, we have established six categories for the logical-rich text-to-image generation (LRT2I) task (Figure 3):

[Detail] This category involves text prompts containing intricate scene details and comprehensive relational information within a scene, designed to test if the model can effectively visualize content-rich narratives and ensure the intricate details align with the textual description.

[Inference] In this category, prompts challenge the model with entities sharing identical attributes, like shape and color, requiring inferential reasoning to discern the target entity among others, thus assessing the model's ability to distinguish similar entities through inference.

[Relational Ambiguity] This category's prompts contain narrative ambiguities with ambiguous relations that clarify towards the end, testing the model's ability to understand context, reduce ambiguity, and generate coherent images that resolve relational ambiguities.

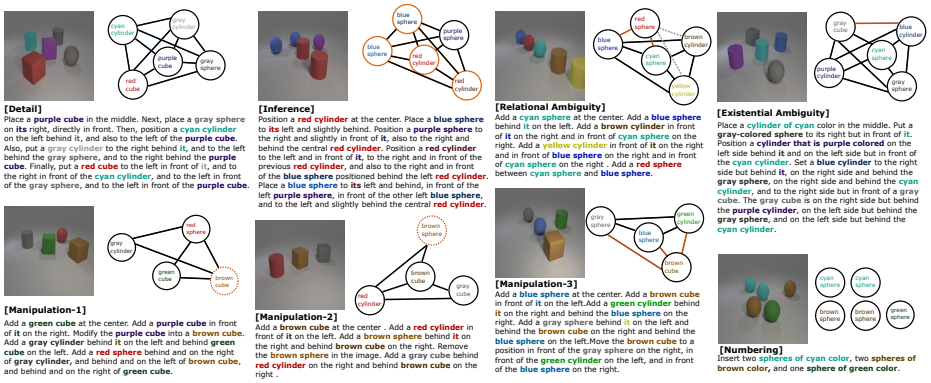


Fig. 3: Overview of TV-Logic Dataset Categories. This composite image illustrates the diverse challenges in text-to-image generation, showing six categories for model evaluation. These categories demonstrate the diverse challenges in text-to-image generation. Within each category, the graph depicts scene 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.

[Existential Ambiguity] This category assesses the model's capacity to identify and place initially undefined entities, navigating scenarios where entities' locations or attributes are not explicitly defined until later in the narrative.
[Manipulation] Prompts in this category feature textual manipulations in three subcategories: modifying the entity attributes, removing the entity, and moving an entity from one location to another. The evaluation focuses on the model's understanding of these manipulations to accurately reflect the intended changes in the generated images.
[Numerical Representation] This category addresses the challenge of visually representing numerical information from text prompts. It tests the model's precision in text-to-image generation, ensuring the quantity and attributes of entities match the textual description and accurately implementing the numerical details into the generated images.

Each category within the TV-Logic dataset is designed to challenge and quantify a model's capabilities across different dimensions of understanding and reasoning. Presenting scenarios from highly detailed to broadly ambiguous, it provides a comprehensive framework for evaluating text-to-image generation advancements. It offers textual challenges mirroring natural language complexity and nuance, enabling thorough text-to-image model evaluation. Dataset generation and annotation details are in the Appendix.

Text Prompts The vocabulary of our dataset (Figure 4) is composed of 691 substantive words, having been refined to exclude common stopwords (for instance, 'to', 'of', 'a', 'the') as well as punctuation marks, thereby ensuring an emphasis on semantically significant terms. Table 2 delineates the statistical

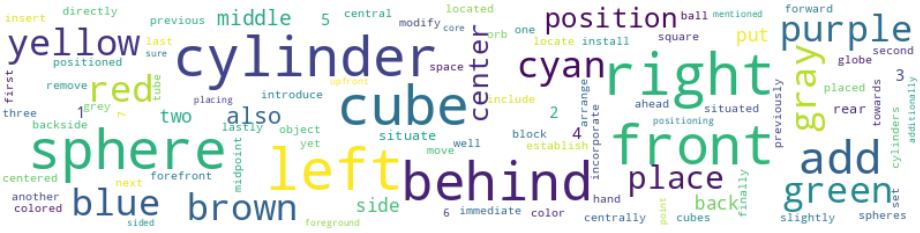


Fig. 4: Word cloud visualization representing the 100 most frequent terms within the text prompts, where font magnitude correlates directly with term occurrence frequency.

Detail					Inference					Relational Ambiguity				
Max	**Min**	**Avg**	**Std**	**Total**	**Max**	**Min**	**Avg**	**Std**	**Total**	**Max**	**Min**	**Avg**	**Std**	**Total**
149	52	104.79	9.51	2,586	192	48	130.32	14.95	2,251	142	52	90.46	16.44	2,600
Existential Ambiguity					**Manipulation**					**Numerical Rep**				
Max	**Min**	**Avg**	**Std**	**Total**	**Max**	**Min**	**Avg**	**Std**	**Total**	**Max**	**Min**	**Avg**	**Std**	**Total**
154	76	108.28	9.63	2,586	145	64	92.71	14.48	2,600	27	13	13.87	0.92	2,590

TV-Logic				
Max	**Min**	**Avg**	**Std**	**Total**
192	13	89.17	38.32	15,213

Table 2: Statistical Overview of Prompt Lengths in the TV-Logic Dataset.

attributes pertaining to the lengths of prompts within the TV-logic dataset, encompassing the minimum, maximum, average, and standard deviation for each specified category. This statistical breakdown offers a crucial insight into the dataset's composition and prompt length variability, essential for subsequent analyses.

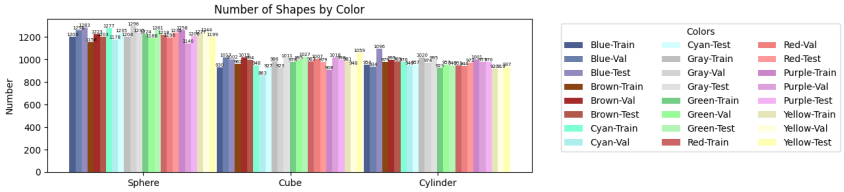


Fig. 5: Statistical Overview of Shape-Attribute Composition. The colors of the bars correspond to the entity colors, with different shades representing the training, validation, and test splits. The numbers on each bar indicate the count of each composition. There is no significant imbalance between the training and test splits.

Images Figure 5 illustrates the distribution of each shape-attribute composition. Our research focuses on balanced data with logic-rich text prompts. Additionally, for each text prompt and image pairing, we provide annotations for a scene graph and a sequence of images. These annotations serve as valuable

references for future research endeavors, particularly due to the time-sequence operations present in the text prompts. Details are available in the Appendix.

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

Textual-Visual Logical Reasoning Table 3 presents statistical data on reasoning concepts key to understanding the dataset's complexity. Two concepts are highlighted: the first is the count of cross-sentence object references within the text prompts, indicating the frequency with which an object is mentioned in subsequent sentences. *E.g., in Figure 2 [Detail], "it" in sentence 3 refers back to the "cyan cylinder" from the second sentence*, assessing the model's understanding and reasoning. The second involves the enumeration of relations within the prompts, *e.g., "in front of" between "red cube" and "cyan cylinder" in Figure 2 [Detail]*, showcasing entity interactions. For both, we report maximum, minimum, average, and standard deviations. This analysis is crucial for evaluating the depth of reasoning required to comprehend and respond to prompts, underlining its importance in text-based reasoning task complexity.

3.3 Evaluation Metrics

For task evaluation, to effectively assess the model performance in generating images that comprehend the structural information in text input, specific evaluation metrics, rather than a pixel-level measurement, are required. Given the ill-posed nature of the problem previously mentioned (Figure 1b), these metrics should concentrate on aligning entity presences and their respective relations between the ground-truth and generated images. Consequently, the objectives of the task should also prioritize these aspects.

We adapted evaluation metrics from previous work [6], which emphasizes relational information similar to our study, making its methodology applicable for our evaluation. We adopted two main metrics:

(i) Object Presence Matches: Evaluates the model's accuracy in identifying and generating objects mentioned in text prompts, comparing the presence of objects in both generated images and ground truth.

(ii) Object Position Relation Matches: Assesses spatial accuracy by comparing object positions in generated images with the ground truth, indicating the

model's understanding of spatial dynamics from the text. However, in the Numerical Representation category, this metric is omitted since relative relations aren't directly mentioned in text prompts, thus not included in overall results.

For (i), metrics include average precision (AP), average recall (AR), and F1 score, derived from detection results of both image types and computed for each scene. For (ii), the relational similarity (RSIM) is set to measure object arrangement. RSIM is articulated as:

$$RSIM(E_{G_{gt}}, E_{G_{gen}}) = recall \times \frac{|E_{G_{gt}} \cap E_{G_{gen}}|}{|E_{G_{gt}}|} \tag{1}$$

Here, *recall* represents the ratio of detected objects in the generated image in relation to those in the ground-truth. $E_{G_{gt}}$ and $E_{G_{gen}}$ denote the sets of relational edges for the ground-truth and generated images, respectively, concerning vertices shared by both images. To address concerns about the object detector's performance as an evaluation metric, we also include reference metrics, i.e., AP, AR, and F1 scores for the fully annotated image.

4 Our Baseline Model

We propose the Understanding and Reasoning Generative Adversarial Network (UnR-GAN) as a baseline model, emphasizing key aspects of the T2I task. UnR-GAN specializes in interpreting structural text information and aligning features across uni-modal and multi-modal domains, as illustrated in Figure 6.

4.1 Relation Understanding Module

The relation understanding module extracts inherent structural relations from inputs and embeds these enhanced relations into the Text Encoder (E_{text}).

First, text prompts ($\mathbf{T}$) are processed by a BERT-based encoder, extracting token ($\mathbf{f} = \{f_i\}_{i=0}^{N}$) and sentence features ($\mathbf{s}$) with context attention. The Relation Enhancing Model parses dependency information ($\mathcal{E}$), guiding self-attention to identify relation-related tokens ($\mathbf{r} = \{\mathfrak{r}_j\}_{j=0}^{D}$). Sentence features ($\mathbf{s}$) are then transformed by an Multi-Layer Perceptron(MLP) and merged with a noise vector $\mathbf{z} \sim \mathcal{N}(0, I)$, forming $\mathbf{s}_z$. E_{text} further encodes these into relation-enhanced text features ($\mathbf{e}_{txt}$), leveraging $\mathcal{E}$ and token indexes ($\mathbf{id} = \{\mathtt{id}_j\}_{j=0}^{D}$).

$$\begin{aligned} \tilde{\mathbf{f}} &= \mathcal{MHA}(\mathbf{f}, G_{guide} = \mathbf{A}_{[\mathcal{E},\mathtt{id}]}), \\ \mathbf{e}_{txt} &= Concat(\tilde{\mathbf{f}}, \mathbf{s}_z) \end{aligned} \tag{2}$$

$\mathbf{A}_{[\mathcal{E},\mathtt{id}]}$, initialized from the dependency graph's adjacency matrix $\mathcal{E}$ and marked by relation-related token indices ($\mathbf{id}$), is a learnable matrix. The multi-head attention block ($\mathcal{MHA}$) applies self-attention to token features with $\mathbf{A}_{[\mathcal{E},\mathtt{id}]}$ serving as a soft attention mask, highlighting natural language relations and relation-related tokens. These enhanced features are then combined with $\mathbf{s}_z$ to create text-enhanced features $\mathbf{e}_{txt}$ for the subsequent module.

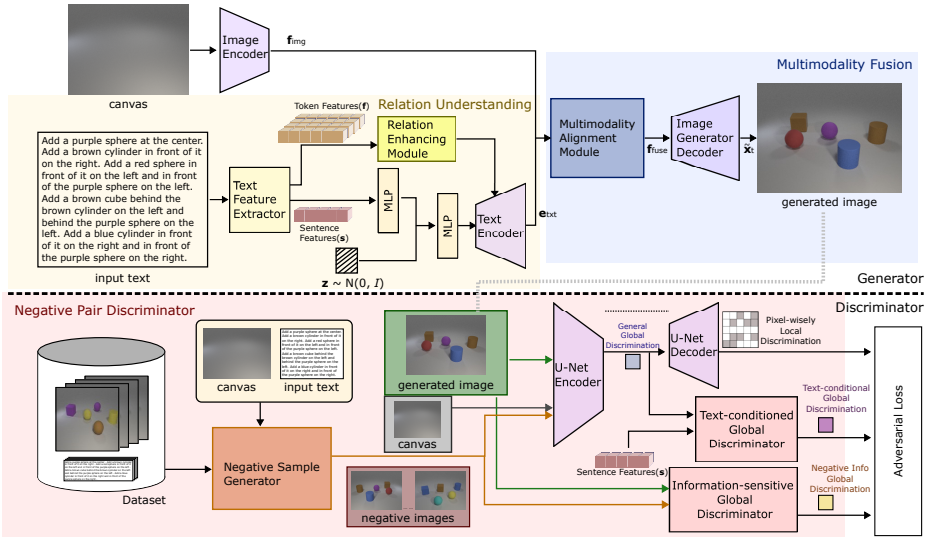


Fig. 6: Overview of our approach. The model consists of a generator to generate images from text and a discriminator for distinguishing samples. It operates in three stages: first, the Relation Understanding Module extracts structural text data; next, the Multimodality Fusion Module combines text and image features for improved alignment; finally, the Negative Pair Discriminator identifies generated images from negatives.

4.2 Multimodality Fusion Module

The multimodality fusion module is designed to merge visual features and relation-enhanced text features.

[Multimodality Self- and Cross-attention] This fusion is achieved by m cross-modality towers and n semantic self-encoders (Figure 7), inspired by [31]. The visual features ($\mathbf{f}_{img}$) are extracted from an image encoder, comprising a residual convolutional network that downsamples the features. Subsequently, $\mathbf{f}_{img}$ and relation-enhanced text features ($\mathbf{e}_{txt}$) are directed into corresponding Linear Layers (g_{txt} and g_{img}).

$$\begin{aligned} \mathbf{Z}_0^{txt} &= g_{txt}(\mathbf{e}_{txt}), \\ \mathbf{Z}_0^{img} &= g_{img}(\mathbf{f}_{img}) \end{aligned} \tag{3}$$

The semantic feature ($\mathbf{Z}_0^{txt}$) is then processed through n semantic self-encoder ($Encoder^{sem}$) for further self-attention learning, before being integrated with the visual stream in the cross-modality tower ($\mathcal{CMT}$).

$$\begin{aligned} \mathbf{Z}_l^{txt} &= Encoder_{l-1}^{sem}(\mathbf{Z}_{l-1}^{txt}),\ l = 1, .., n, \\ \tilde{\mathbf{Z}}_0^{txt} &= \mathbf{Z}_n^{txt}, \\ \tilde{\mathbf{Z}}_k^{txt}, \mathbf{Z}_k^{img} &= \mathcal{CMT}_{k-1}(\tilde{\mathbf{Z}}_{k-1}^{txt}, \mathbf{Z}_{k-1}^{img}),\ k = 1, .., m \end{aligned} \tag{4}$$

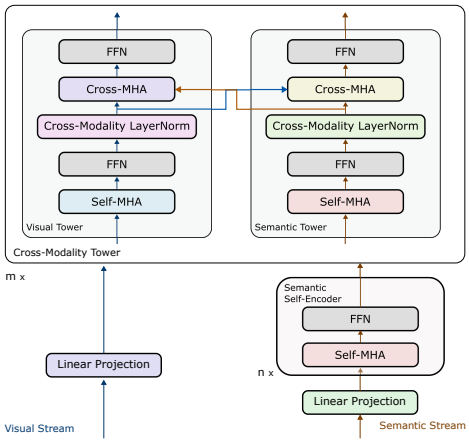


Fig. 7: Structure of attention model. Image and text features are linearly projected, with text features refined by n self-encoders. m cross-modality towers merge streams, each including visual and semantic towers. Each stream's tower has self- and cross-multi-head attention, feedforward, and normalization layers.

The output from the last visual tower, termed visual-fused features ($\mathbf{v}_{fuse} = \tilde{\mathbf{Z}}_m^{img}$), contains self-attention and cross-attention of the two modality features. **[Multimodality Fusion]** We construct a model using the Text Image Residual Gating-based network [27] to fuse visual ($\mathbf{v}_{fuse}$) and text-enhanced ($\mathbf{e}_{txt}$) features, emphasizing text-modified image features over new feature space creation. For improved alignment, it employs semantic-attended visual features ($\mathbf{v}_{fuse}$) rather than solely extracted visual features.

The compose features ($\mathbf{h}$) are calculated in Equation 5. The gating feature design aims to retain the image feature when text prompts are less informative.

$$\begin{aligned} \mathbf{h} &= \mathbf{W}_{gate}\mathbf{f}_{gate} + \mathbf{W}_{res}(1 - \mathbf{f}_{gate}) \odot \mathbf{f}_{res}, \\ \mathbf{f}_{gate} &= ConvNet_2(ConvNet_1([\mathbf{v}_{fuse}; \mathbf{s}_z])) \odot \mathbf{f}_{img}, \\ \mathbf{f}_{res} &= Conv(ConvNet_3([\mathbf{v}_{fuse}; \mathbf{s}_z])) \end{aligned} \tag{5}$$

W_{gate} and W_{res} are weights for gating ($\mathbf{f}gate$) and residual ($\mathbf{f}res$) features. $\odot$ denotes element-wise multiplication, and $[\cdot, \cdot]$ indicates concatenation. $ConvNet$ includes a convolution layer, activation function ($ReLU$ for $ConvNet_1$ and $ConvNet_3$, σ for $ConvNet_2$), and batch normalization, while $Conv$ specifies a convolutional layer alone. The features $\mathbf{h}$ enter an image generation decoder with upsampling residual layers to produce the final image $\tilde{\mathbf{x}}_t$.

4.3 Negative Pair Discriminator

In conventional GANs, the discriminator separates real from generated data. We extend this role by having the discriminator also identify specific data characteristics. Training our discriminator has three objectives: 1) distinguish real from

fake images, 2) discern paired from unpaired image-text sets to ensure relevance, and 3) promote image variation when input content changes. Negative samples are generated accordingly based on the discriminator type.

[U-Net Based Local Discriminator] It focuses on local data features to discern real and fake data per pixel, segmenting images into real and fake regions, akin to [24]. It is defined as:

$$\begin{aligned} \mathbf{d}_{uNet}^{real} &= D_{dec}(D_{enc}(\mathbf{x}_t)) \\ \mathbf{d}_{uNet}^{fake} &= D_{dec}(D_{enc}(\tilde{\mathbf{x}}_t)) \end{aligned} \tag{6}$$

where D_{enc} and D_{dec} are U-Net encoder and decoder, respectively, and $\mathbf{x}_t$ is the ground truth real image.

[Text-conditioned Global Discriminator] It globally evaluates the text-image relationship, enabling effective discrimination between paired and unpaired sets. This generates a scalar indicating if the image relates to the text input rather than an unpaired or randomly sampled negative text prompt:

$$d_{txt-G}^{fake} = D_G(D_{enc}(\tilde{\mathbf{x}}_t) - D_{enc}(\mathbf{x}_{bg}, \mathbf{s})) \tag{7}$$

where D_G is the text-conditioned global discriminator, taking as inputs the sentence features $\mathbf{s}$ and the difference between the encoded canvas image ($\mathbf{x}_{bg}$) and the generated image. We use $\mathbf{x}_t$ to replace $\tilde{\mathbf{x}}_t$ for d_{txt-G}^{real}, and for d_{txt-G}^{unpair}, $\tilde{\mathbf{x}}_t$ is replaced with $\mathbf{x}_t$, and $\mathbf{s}$ is replaced with $\mathbf{s}^* \sim \mathcal{D}$, where $\mathbf{s}^*$ is randomly sampled from dataset $\mathcal{D}$.

[Information-sensitive Global Discriminator] It is designed to be sensitive to the information content, this discriminator plays a critical role in detecting variations resulting from disturbances in the original text inputs.

$$d_{info-G} = similarity(\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_t^{\mathbf{s}*}) \tag{8}$$

Here, we use normalized cosine similarity as loss to encourage the dissimilarity of two generated images, while $\tilde{\mathbf{x}}_t^{\mathbf{s}*}$ is the one generated from the prompt where the informative tokens have been disturbed. We define these informative tokens as attributes (e.g., adjectives) and relations (e.g., adpositions, verbs) based on tokens in the sentences from the Part-of-Speech Tagset (POS tag) [26].

5 Experiments

5.1 Dataset

To address T2I generation challenges, especially complex subject-object relations, we introduced the TV-Logic dataset, featuring structurally complex images. This synthetic dataset comprises three shapes and eight colors, with prompts detailing object positioning relative to existing entities. We've segmented it into six categories for diverse reasoning task assessment.

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**

Table 4: 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).

5.2 Quantitative Results

[Overall Results] In our study, we evaluate the TV-Logic dataset using metrics such as AP, AR, F1, RSIM, and standard evaluations like FID, LIPIPS, SSIM, and PSNR. Our approach is compared with six related works: DALL-E [21] for diverse image generation; U-ViT [2], which combines ViT and U-Net with diffusion models; GeNeVA-GAN [5] and LatteGAN [16] for image manipulation task; "GPT+Blender", utilizing ChatGPT4 [20] and Blender [3] to convert text prompts into 3D scenes; and "GPT+Comp", focusing on scene generation using diffusion models for image components, but they require a Large Language Model to transform logic-rich text prompts into pairwise relations for execution. Quantitative results, displayed in Table 4, highlight our model's standout performance in relation measurement compared to baselines known for object presence scores. On the TV-Logic dataset, while our model exhibits strong object presence scores, it still struggles to accurately measure relationships in complex prompts.
[Category-Specific Results] Table 5 displays the quantitative results on the TV-Logic dataset, breaking down performance across different subcategories. In cases that incorporate comprehensive relational details, our model demonstrates improved results compared to other related works. These findings suggest that emphasizing the three principal factors in this domain can lead to better generation of entity attributes and more accurate interpretation of relationships. However, in scenarios containing multiple entities possessing the same attributes within a scene, or in cases requiring the model to identify the target entity, a decrease in performance is observed for all models. Furthermore, the task of aligning numerical representations in natural language with their corresponding visual outputs presents a significant challenge within the realm of multi-modality, constituting a primary obstacle in this domain.
[Ablation Studies] Our experiments, shown in Table 4, evaluated our architecture's variations to determine each component's impact. In Exp.7 (Focal Informative), the Negative Pair Discriminator's loss function was removed, affect-

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	-

Table 5: Quantitative analysis of the TV-Logic dataset across different subcategories.

ing focus on prompt sections. Exp.8 (Focal Relation) involved substituting the learned attention layer with a uniform matrix, thus eliminating relation-guided feature integration. Exp.9 (Cross Modality Reference) used the direct concatenation of multi-modality features, bypassing the Cross-Modality Tower, which impacted cross-modality alignment and fusion. These modifications, not significantly altering the model's scale, offered a basis for performance comparison, demonstrating the crucial role of each component in our proposed architecture.

5.3 Qualitative Results

Figure 8 displays qualitative outcomes from the TV-Logic dataset across various categories, showcasing ground truth images alongside our baseline outputs. The performance of DALL-E and U-ViT is depicted in the third and fourth columns, respectively. These transformer-based models often misinterpret complex relations and struggle with lengthy text inputs, resulting in inaccuracies in entity depiction. Results for U-ViT and LatteGAN are presented in subsequent columns. They are designed for conditional image generation from textual prompts and visual context, where initially perform well but their performance diminishes with more complex sentences. GPT-Blender and GPT-Comp, employing a compositional strategy, utilize a large language model to convert text prompt to 3D objects and pairwise text prompts for Blender and Composable Diffusion Models, respectively. While effective with basic relational inputs, Composable Diffusion Models face challenges with complex relations and deeper reasoning, relying on ChatGPT for accurate interpretation.

6 Limitation Discussion and Conclusion

Our research introduces the 'logic-rich text-to-image generation' task, underscoring the significance of structural information in this field. We developed the

Fig. 8: 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.

TV-Logic dataset to assess diverse model performances, marking it as the first comprehensive reasoning dataset in this domain. Our proposed baseline model includes a Negative Pair Discriminator, Relation Understanding Module, and Multimodality Fusion Module, aimed at improving reasoning from text to images. Experiments on the TV-Logic dataset demonstrate our model's effectiveness, setting a new benchmark in the field.

The study of logic-rich text-to-image generation, still in its early stages, primarily relies on synthetic and clip-art datasets for evaluation, particularly in relation-focused tasks. However, advances in scene graph generation and computer vision could overcome these challenges. Additionally, there is still room to explore how to handle the ambiguity inherent in text prompts and how to appropriately evaluate the result. Given the field's novelty and inherent complexity, we expect a surge in research contributions in the near future. The failure cases and limitations of the model will be presented in the Appendix.

References

1. Ak, K.E., Lim, J.H., Tham, J.Y., Kassim, A.: Semantically consistent hierarchical text to fashion image synthesis with an enhanced-attentional generative adversarial network. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 3121–3124. IEEE (2019) 3
2. Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., Zhu, J.: All are worth words: A vit backbone for diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22669–22679 (2023) 12
3. Blender Foundation: Blender - a free and open source 3d creation suite. https://www.blender.org/ (2023), accessed: 2024-02-29 12
4. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A.: Generative adversarial networks: An overview. IEEE signal processing magazine **35**(1), 53–65 (2018) 3
5. El-Nouby, A., Sharma, S., Schulz, H., Hjelm, D., Asri, L.E., Kahou, S.E., Bengio, Y., Taylor, G.W.: Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10304–10312 (2019) 12
6. El-Nouby, A., Sharma, S., Schulz, H., Hjelm, D., El Asri, L., Ebrahimi Kahou, S., Bengio, Y., Taylor, G.W.: Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2019) 7
7. Gao, L., Wang, B., Wang, W.: Image captioning with scene-graph based semantic concepts. In: Proceedings of the 2018 10th international conference on machine learning and computing. pp. 225–229 (2018) 3
8. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10696–10706 (2022) 3
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017) 2, 3
10. Hong, S., Yang, D., Choi, J., Lee, H.: Inferring semantic layout for hierarchical text-to-image synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7986–7994 (2018) 3
11. Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1219–1228 (2018) 3
12. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 3
13. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**, 32–73 (2017) 1
14. Liang, W., Jiang, Y., Liu, Z.: Graghvqa: language-guided graph neural networks for graph-based visual question answering. arXiv preprint arXiv:2104.10283 (2021) 3
15. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 1

16. Matsumori, S., Abe, Y., Shingyouchi, K., Sugiura, K., Imai, M.: Lattegan: Visually guided language attention for multi-turn text-conditioned image manipulation. IEEE Access **9**, 160521–160532 (2021) 12
17. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable fidelity and diversity metrics for generative models. ArXiv **abs/2002.09797** (2020) 3
18. Nguyen, K., Tripathi, S., Du, B., Guha, T., Nguyen, T.Q.: In defense of scene graphs for image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1407–1416 (2021) 3
19. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021) 3
20. OpenAI: Gpt-4: Openai's generative pre-trained transformer 4. https://openai.com/ (2023), accessed: 2024-02-29 12
21. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021) 3, 12
22. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 3
23. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29** (2016) 3
24. Schonfeld, E., Schiele, B., Khoreva, A.: A u-net based discriminator for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8207–8216 (2020) 11
25. Teney, D., Liu, L., van Den Hengel, A.: Graph-structured representations for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2017) 3
26. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 human language technology conference of the north american chapter of the association for computational linguistics. pp. 252–259 (2003) 11
27. Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.J., Fei-Fei, L., Hays, J.: Composing text and image for image retrieval-an empirical odyssey. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6439–6448 (2019) 10
28. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (Apr 2004). https://doi.org/10.1109/TIP.2003.819861 2
29. Xiong, P., Zhan, H., Wang, X., Sinha, B., Wu, Y.: Visual query answering by entity-attribute graph matching and reasoning. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 8357–8366 (2019) 3
30. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1316–1324 (2018) 3
31. Xu, X., Wu, C., Rosenman, S., Lal, V., Che, W., Duan, N.: Bridgetower: Building bridges between encoders in vision-language representation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 10637–10647 (2023) 9

32. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10685–10694 (2019) 3
33. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring visual relationship for image captioning. In: Proceedings of the European conference on computer vision (ECCV). pp. 684–699 (2018) 3
34. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 5907–5915 (2017) 3
35. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 2
36. Zhao, B., Yin, W., Meng, L., Sigal, L.: Layout2image: Image generation from layout. International Journal of Computer Vision **128**, 2418–2435 (2020) 3