

EVALUATING SEMANTIC VARIATION IN TEXT-TO-IMAGE SYNTHESIS: A CAUSAL PERSPECTIVE

Anonymous authors

Paper under double-blind review

ABSTRACT

Accurate interpretation and visualization of human instructions are crucial for text-to-image (T2I) synthesis. However, current models struggle to capture semantic variations from word order changes, and existing evaluations, relying on indirect metrics like text-image similarity, fail to reliably assess these challenges. This often obscures poor performance on complex or uncommon linguistic patterns by the focus on frequent word combinations. To address these deficiencies, we propose a novel metric called SemVarEffect and a benchmark named SemVarBench, designed to evaluate the causality between semantic variations in inputs and outputs in T2I synthesis. Semantic variations are achieved through two types of linguistic permutations, while avoiding easily predictable literal variations. Experiments reveal that the CogView-3-Plus and Ideogram 2 performed the best, achieving a score of 0.2/1. Semantic variations in object relations are less understood than attributes, scoring 0.07/1 compared to 0.17-0.19/1. We found that cross-modal alignment in UNet or Transformers plays a crucial role in handling semantic variations, a factor previously overlooked by a focus on textual encoders. Our work establishes an effective evaluation framework that advances the T2I synthesis community’s exploration of human instruction understanding.

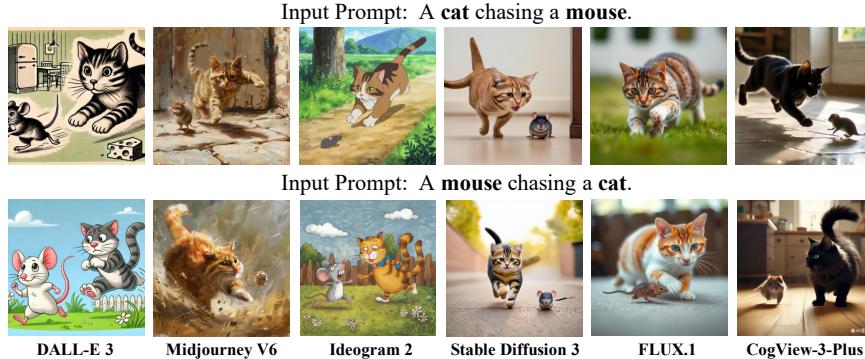


Figure 1: Failed state-of-the-art (SOTA) T2I model examples: different permutations of the same words, different textual semantics, yet similar visual semantics.

1 INTRODUCTION

Accurately interpreting and visually depicting human instructions is essential for text-to-image (T2I) synthesis Cao et al. (2024). Despite advancements in alignment Lee et al. (2023a); Wu et al. (2023); Kirstain et al. (2023), composition Liu et al. (2022); Wang et al. (2024); Li et al. (2024); Feng et al. (2024), and long instructions Yang et al. (2024); Gani et al. (2023), these models still treat text prompts as bags of words, failing to depict the semantic variations in human instructions Yu et al. (2024); Mo et al. (2024). As shown in Fig. 1, existing T2I models generate images with identical semantics, even when the inputs differ semantically (e.g., “a mouse chasing a cat” vs. “a cat chasing a mouse”). This indicates that existing T2I models struggle to accurately capture the semantic variations caused by word orders changes.

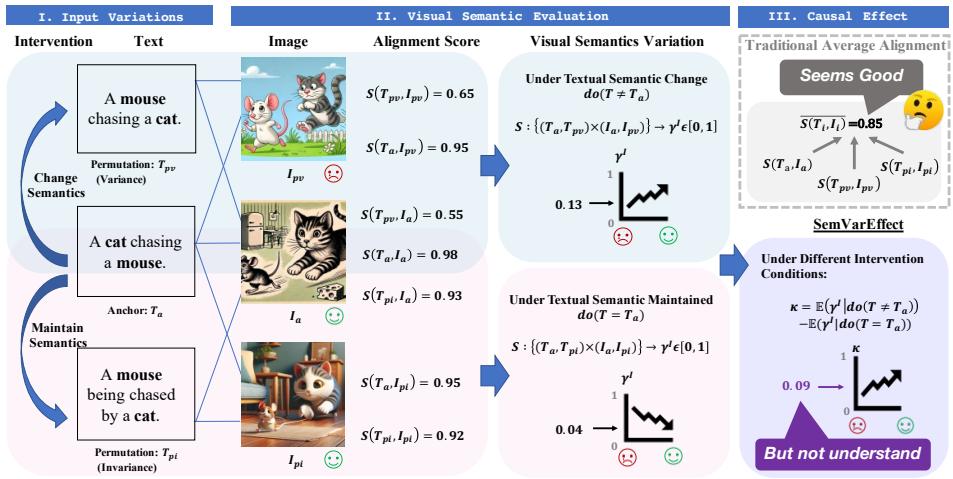


Figure 2: Framework for measuring semantic variation causality in T2I models. Our evaluation consists of three components: (I) **Input Variations** with semantic change/maintenance interventions, (II) **Visual Semantic Evaluation** under both interventions (blue for semantic change, pink for semantic maintenance), and (III) **Causal Effect Calculation** where SemVarEffect (purple) quantifies the difference between intervention outcomes. For Comparison, traditional alignment scores (gray) only measure surface similarity, as shown in the cat-mouse example where high alignment coexists with poor semantic consistency. See Section 2 for mathematical details.

There is a lack of direct metric to evaluate a T2I model’s ability to understand semantic variations caused by word order changes. Existing NLP research typically evaluates semantic variation indirectly through downstream tasks. For example, in language generation Gordon et al. (2020), the input sequences with different word orders are used as the actions in a navigation game and the model is evaluated based on the game’s accuracy. Similarly, in visual-language understanding Thrush et al. (2022); Diwan et al. (2022); Yüksekgönül et al. (2023); Wang et al. (2023); Burapacheep et al. (2024), models are evaluated via cross-modal retrieval and image-text matching, focusing on text-image similarity. In T2I synthesis, the text-image alignment score offers an indirect performance measure but may not fully capture a model’s sensitivity and robustness to word order. For example, as shown in the upper-right of Fig. 2, an average alignment score of 85, as evaluated by GPT-4, might seem satisfying, but it may conceal the model’s proficiency with common word combinations while masking its inadequacy with less frequent or more complex linguistic patterns.

We propose a novel metric, called SemVarEffect, to evaluate the causality of semantic variations between inputs and outputs of T2I models. Our approach uses inputs’ semantics as the only intervention to evaluate the average causal effect (ACE) of this intervention on outputs’ semantic variations, that is, the contribution of inputs to outputs. A significant ACE would indicate that the T2I model can effectively capture and reflect input semantic variations. On the contrary, a small ACE, such as the 0.09 shown in Fig. 2, exposes a considerable weakness in the T2I model’s ability to understand and respond to sentence semantics.

To facilitate the evaluation, we present a new benchmark, called SemVarBench. To avoid overt literal differences, semantic variations are achieved through two types of linguistic permutations Gerner (2012): permutation-variance, where different word orders result in different meanings, and permutation-invariance, where the meaning remains unchanged regardless of word orders. Utilizing pre-defined templates and rules as the guidance in the generation stage, followed by a large amount of annotation and hard sample selection in the validation stage, we constructed a benchmark comprising 11,454 samples, where 10,806 are in the training set and 648 are in the test set. We experimented with a variety of T2I models using our proposed benchmark and metric. The results show that even SOTA models like CogView-3-Plus and Ideogram 2 struggle, achieving scores far from the ideal, which highlights the need for further advancements in handling semantic variations.

Our key contributions are: (1) A first systematic study of semantic variations in T2I synthesis, investigating the causal relationship between input text variations and output images. (2) Sem-

108 VarEffect: A metric quantifying how semantic variations in input text affect T2I output quality.
 109 (3) SemVarBench: An expert-annotated benchmark evaluating semantic variations in T2I synthesis
 110 through permutation-variance and permutation-invariance tests. (4) Comprehensive evaluation
 111 of SOTA T2I models, revealing significant limitations in handling semantic variations and distinct
 112 challenges posed by different variation types, while identifying specific areas for improvement.
 113

114 2 SEMANTIC VARIATION EVALUATION FOR TEXT-TO-IMAGE SYNTHESIS

116 2.1 PRELIMINARY

118 The T2I model f generates images I for each input sentence T , represented as $I = f(T)$. $S(T, I)$ is
 119 the text-image alignment score, measuring text-image similarity. $S(\cdot)$ represents the scoring method.

120 **Linguistic Permutation.** Linguistic permutation refers to changes in word order. Given an anchor
 121 sentence T_a , T_{pv} and T_{pi} are two permutations of T_a . T_{pv} exemplifies permutation-variance, which
 122 shows a change in meaning, while T_{pi} exemplifies permutation-invariance, where the meaning re-
 123 mains unchanged. The expected I_{pv} is a permutation of objects or relations from I_a , while I_{pi} is
 124 semantically equivalent to I_a , preserving the same visual objects and relations after transformation.
 125

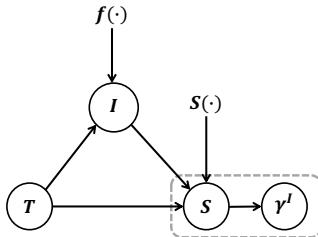
126 2.2 DEFINITION OF VISUAL SEMANTIC VARIATIONS

128 First, we define the visual semantic variations observed from a single sentence T . We decompose
 129 complex semantic variation into minimal discrete steps, called localized changes. For each image
 130 I , the visual semantic variation of a single minimal discrete step $I + \Delta I$, denoted as $\mu_I(T, I)$, is
 131 the difference in alignment scores: $\mu_I(T, I) = S(T, I + \Delta I) - S(T, I)$. When the anchor image I_a
 132 transforms to a permutation image I_p through these minimal discrete steps, the integrated visual
 133 semantic variation is: $\sum_{I_a}^{I_p} \mu_I(T, I) = S(T, I_p) - S(T, I_a)$.

134 Second, we sum the visual semantic variations from multiple text-image pairs to comprehensively
 135 measure these variations. For the sentence T_a , the visual semantic variations $\sum_{I_a}^{I_p} \mu(T_a, I)$ demon-
 136 strate a shift from a matched to a mismatched image-text pair, indicating a negative change. For the
 137 sentence T_{p*} , the visual semantic variations $\sum_{I_a}^{I_p} \mu(T_{p*}, I)$ demonstrate a shift from a mismatched
 138 to a matched image-text pair, indicating a positive change. To measure the total magnitude of these
 139 variations regardless of direction, we use the absolute values. Therefore, the summation of visual
 140 semantic variations is defined as γ^I :

$$141 \quad \gamma^I = \sum_{T \in \{T_a, T_{p*}\}} \left| \sum_{I_a}^{I_p} \mu(T, I) \right| = |S(T_a, I_p) - S(T_a, I_a)| + |S(T_{p*}, I_p) - S(T_{p*}, I_a)|. \quad (1)$$

144 2.3 THE CAUSALITY BETWEEN TEXTUAL AND VISUAL SEMANTIC VARIATIONS



154 Figure 3: Causal relationship
 155 between the input and the out-
 156 put semantic variations.

145 Fig. 3 illustrates the causal relationship between input and output
 146 semantic variations. T is the text input, serving as the input variable,
 147 while I is the generated image, acting as a mediator. S is
 148 the text-image alignment score, influenced by both T and I , and
 149 serves as an intermediate result variable. γ^I denotes visual seman-
 150 tic variation and is the final comparison result variable. $f(\cdot)$ is an
 151 exogenous variable representing a T2I model that maps T to I . $S(\cdot)$
 152 is an exogenous variable representing a scoring function that maps
 153 T and I to S . The dashed line between S and γ^I indicates their de-
 154 rived relationship: γ^I is the difference between two S values under
 155 different output conditions, representing the comparative result of
 156 the alignment scores when the image changes.

157 According to the causal inference theory, we define the average causal effect (ACE) of textual se-
 158 mantic variations on visual semantic variations as the SemVarEffect score. As shown in Fig. 3,
 159 the sentence T serves as an independent variable that influences the generated image I . The visual
 160 semantics variations is jointly influenced by T , I and $S(\cdot)$. Let $\text{do}(T \neq T_a)$ and $\text{do}(T = T_a)$ represent
 161 two types of interventions. $\text{do}(T \neq T_a)$ represents an intervention where T differs in meaning from
 the anchor sentence T_a . The visual semantic variation caused by this intervention is denoted as:

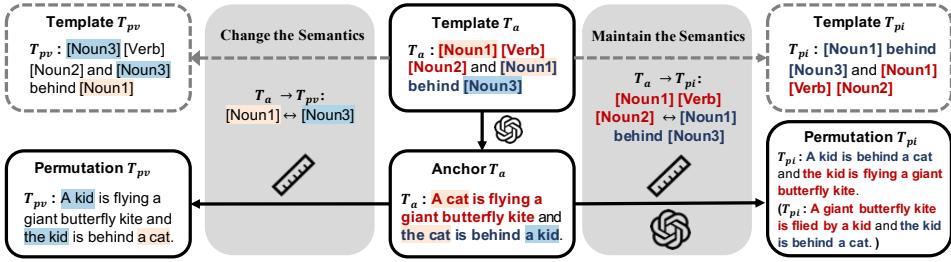


Figure 4: The data collection process of SemVarBench. Top: Templates. Bottom: Generated Sentences. The templates are extracted from the seed pair “a dog is using a wheelchair and the dog is next to a person”/“a person is using a wheelchair and the person is next to a dog”.

$$\gamma_{w/}^I = \mathbb{E}[\gamma^I | \text{do}(T \neq T_a)] = \mathbb{E}[\gamma^I | T = T_{pv}] = |S(T_a, I_{pv}) - S(T_a, I_a)| + |S(T_{pv}, I_{pv}) - S(T_{pv}, I_a)|. \quad (2)$$

$\text{do}(T = T_a)$ represents an intervention where T match the meaning of the anchor sentence T_a . The visual semantic variation caused by this intervention is denoted as:

$$\gamma_{w/o}^I = \mathbb{E}[\gamma^I | \text{do}(T = T_a)] = \mathbb{E}[\gamma^I | T = T_{pi}] = |S(T_a, I_{pi}) - S(T_a, I_a)| + |S(T_{pi}, I_{pi}) - S(T_{pi}, I_a)|. \quad (3)$$

By comparing visual variations under semantic change interventions and semantic maintenance interventions, we determine the ACE of textual semantic variations:

$$\begin{aligned} \kappa &= \mathbb{E}[\gamma^I | \text{do}(T \neq T_a)] - \mathbb{E}[\gamma^I | \text{do}(T = T_a)] = \gamma_{w/}^I - \gamma_{w/o}^I \\ &= |S(T_a, I_{pv}) - S(T_a, I_a)| + |S(T_{pv}, I_{pv}) - S(T_{pv}, I_a)| \\ &\quad - |S(T_a, I_{pi}) - S(T_a, I_a)| - |S(T_{pi}, I_{pi}) - S(T_{pi}, I_a)|, \end{aligned} \quad (4)$$

The SemVarEffect score κ quantifies the influence of input semantic variations on output semantic variations. The alignment score consists of object and relation (triple) components, each contributing up to 0.5 to the total score. Under ideal conditions, where $f(\cdot)$ accurately represents the text through images and $S(\cdot)$ faithfully measures text-image alignment, κ ranges from 0 to 1. κ is maximized when $\gamma_{w/}^I$ reaches its upper bound of 1, which occurs in extreme cases where no relation between objects are identical, and $\gamma_{w/o}^I$ reaches its optimal value of 0. More detailed analysis of the SemVarEffect score can be found in Appendix B.3– B.5.

3 SEMANTIC VARIATION DATASET FOR TEXT-TO-IMAGE SYNTHESIS

We create a semantic variation dataset for T2I synthesis through two types of linguistic permutations. In this Section, we first describe the data characteristics and then introduce collection pipeline.

3.1 CHARACTERISTICS OF DATA

Each sample (T_a, T_{pv}, T_{pi}) consists of three sentences: an anchor sentence T_a and two permutations T_{pv} and T_{pi} . They should adhere to the following characteristics:

Literal Similarity: T_a , T_{pv} and T_{pi} are literally similar, differing only in word order.

Distinct Semantics: T_a and T_{pv} have distinct semantics. T_a and T_{pi} share the same semantics.

Reasonability: T_a , T_{pv} and T_{pi} are semantically reasonable in either the real or fictional world.

Visualizability: T_a , T_{pv} and T_{pi} describe something humans can visualize.

Discrimination: The images evoked by T_a and T_{pv} present distinguishable differences. The images evoked by T_a and T_{pi} appear similar.

Recognizability: The image evoked by T_a , T_{pv} and T_{pi} maintain key elements necessary for recognizing typical scenes and characters.

216 3.2 DATA COLLECTION
217

218 We use LLMs (GPT-3.5) to generate anchor sentences and their permutations, guided by templates.
219 However, LLMs tend to produce patterns common in their training data, which leads to the neglect
220 of less common combinations specified by templates and rules. To address this issue, we employ a
221 different process for generating T_a , T_{pv} and T_{pi} .

222 **Template Acquisition.** We choose all 171 sentence pairs suitable for T2I synthesis from
223 Winoground Thrush et al. (2022); Diwan et al. (2022) as seed pairs. These pairs are used to ex-
224 tract templates and rules for T_a and T_{pv} , while those for T_{pi} are extended manually. To increase
225 diversity, we change the word orders according to the part of speech, including number, adjective,
226 adjective phrase, noun, noun with adjective, noun with clause, noun with verb, noun with pre-
227 positional phrase, verb, verb with adverb, adverb, prepositional and prepositional phrase. In Fig. 4, the
228 top left shows an example of templates for T_a and T_{pv} derived from extraction, while the top right
229 shows the corresponding templates for T_a and T_{pi} derived from manual completion.

230 **Template-guided Generation for T_a .** We use LLMs to generate anchor sentences by filling tem-
231 plate slots based on prior knowledge and maximum likelihood estimation. In Fig. 4, the bottom
232 middle sentence T_a is generated using the template for T_a as a guide.

233 **Rule-guided Permutation for T_{pv} .** T_{pv} is generated by swapping or rearranging words in T_a based
234 on predefined rules, ensuring that T_{pv} introduces semantic variation. This method avoids a ran-
235 dom generation or a semantically equivalent passive structure to T_a , which a common pitfall in
236 autonomous generation by LLMs. By following these rules, T_{pv} includes many rare combinations
237 not commonly found in existing NLP corpora. In Fig. 4, T_{pv} is generated by swapping [Noun1] and
238 [Noun3] in T_a (shown in the top left).

239 **Paraphrasing-guided Permutation for T_{pi} .** T_{pi} can be generated by following rules, such as ex-
240 changing phrases connected by coordinating conjunctions. However, not all sentences contain coor-
241 dinating conjunctions, so we also allow other synonymous transformations, including passive voice
242 and slight rephrasing. Both T_{pi} examples in Fig. 4 are acceptable.

244 3.3 DATA ANNOTATION AND STATISTICS
245

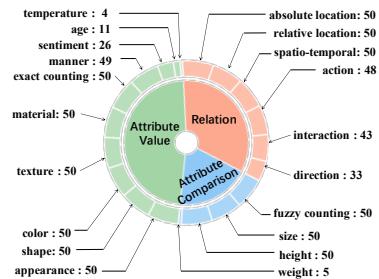
246 **LLM and Human Annotation.** We establish 14 specific
247 criteria to define what constitutes a “valid” input sample.
248 LLMs check each sample against these criteria, labeling
249 them as “yes” or “no” with confidence scores. Samples la-
250 beled “no” with confidence scores above 0.8 are removed.
251 Then, 15 annotators and 3 experts manually verify the re-
252 maining samples. Each sample is independently reviewed
253 by two annotators, with an expert resolving any disagree-
254 ments. This process produced 11,454 valid samples, from
255 which 684 challenging cases are selected for testing based
256 on thresholds and voting. Details are in Appendix C.2.

257 **Scale and Split.** SemVarBench comprises 11,454 samples
258 of (T_a, T_{pv}, T_{pi}) , divided into a training set and a test set.
259 The training set contains 10,806 samples, while the test set
260 consists of 648 samples. All evaluations are on the test set.

261 **Category.** In SemVarBench, samples are divided into 20 categories based on their types of semantic
262 variation. These categories are further classified into three aspects: Relation, Attribute Comparison,
263 and Attribute Values. Fig. 5 shows the distribution of the test set in SemVarBench.

264 4 EXPERIMENTS
265266 4.1 EXPERIMENTAL SETUP
267

268 **T2I Synthesis Models.** We evaluate 13 mainstream T2I models as shown in Tab. 1. For each sen-
269 tence, we generate one image, resulting in a total of $684 \times 3 \times 13$ images. Each input prompt is the
sentence itself, without any negative prompts or additional details expanded by prompt generators.



268 Figure 5: Distribution of semantic
269 variations by category in the semVar-
Bench test set.

Model	Abbr.	Type	#DIM	Text Encoder	#TEP	Image Generator	#IGP	Image Decoder	#IDP	#ToP
Open-source Models										
Stable Diffusion v1.5 Rombach et al. (2022)	SD 1.5	Diffusion	768	CLIP ViT-L	123.06M	UNet	859.52M	VAE	83.65M	1.07B
Stable Diffusion v2.1 Rombach et al. (2022)	SD 2.1	Diffusion	1024	OpenCLIP ViT-H	340.39M	UNet	865.91M	VAE	83.65M	1.29B
Stable Diffusion XL v1.0 Podell et al. (2023)	SD XL 1.0	Diffusion	2048	CLIP ViT-L & OpenCLIP ViT-bigG	123.06M 694.66M	UNet	4.83B	VAE	83.65M	6.51B
Stable Cascade Pernias et al. (2023)	SD CA	Diffusion	1280	CLIP ViT-G	694.66M	UNet	5.15B	VQGAN	18.41M	6.86B
DeepFloyd IF XL Saharia et al. (2022)	DeepFloyd	Diffusion	4096	T5-XXL	4.76B	UNet	6.02B	VAE	55.33B	11.18B
PixArt-alpha XL Chen et al. (2023)	PixArt	Diffusion	4096	Flan-T5-XXL	4.76B	Transformer	611.35M	VAE	83.65M	5.46B
Kolors Team (2024)	Kolors	Diffusion	4096	ChatGLM3	6.24B	UNet	2.58B	VAE	83.65M	8.91B
Stable Diffusion 3[medium] Esser et al. (2024)	SD 3	Diffusion	2048	CLIP ViT-L & OpenCLIP ViT-bigG & T5-XXL	117.92M 662.48M 4.76B	Transformer	2.03B	VAE	83.82M	7.69B
FLUX.I[dev]	FLUX.1	Diffusion	768	CLIP ViT-L & T5-XXL	123.06M 4.76B	Transformer	11.90B	VAE	83.82M	16.87B
API-based Models										
Midjourney V6 DALL-E 3 Betker et al. (2023)	MidJ V6 DALL-E 3	Diffusion	-	-	-	-	-	-	-	-
CogView-3-Plus Ideogram 2	CogV3-Plus Ideogram 2	Diffusion	-	T5-XXL	4.76B	UNet	-	VAE	-	-
		Diffusion	-	T5-XXL ¹	4.76B ¹	Transformer	-	VAE ¹	-	-

¹ The T5-XXL mentioned here is the text encoder of Cogview-3, which is the previous version of Cogview-3-Plus. We have not been able to find specific information about the text encoder and image decoder in the exact materials provided.

Table 1: T2I Models to be evaluated. #DIM represents the pooled dimension of text encoders’ outputs. #TEP, #IGP, #IDP, #ToP represent the parameters of text encoders, image generators, image decoders and whole models.

Evaluators. We use 4 advanced MLLMs as the automatic evaluators to calculate text-image alignment scores: Gemini 1.5 Pro, Claude 3.5 Sonnet, GPT-4o and GPT-4 Turbo. The latter two have demonstrated near-human performance in evaluating the text-image alignment in T2I synthesis Zhang et al. (2023); Chen et al. (2024). We format a sentence and an image in a prompt and feed it into the evaluator, asking it to assign two scores: object accuracy (0-50 points) and relation accuracy (0-50 points). The sum of these two scores is treated as the total score, which is then normalized to [0, 1]. **To validate MLLM effectiveness, we conducted human evaluation with three raters on 80 samples (20 each from Midjourney v6, DALL-E 3, CogView3Plus, and Ideogram2) using the same scoring protocol as our automatic evaluation. We measure MLLMs’ correlation with human preferences using Pearson’s ρ , Spearman’s ϕ , and Cohen’s κ_{cohen} coefficients.**

Metrics. We use 4 metrics: text-image alignment score (\bar{S}_{ii}), our proposed SemVar-Effect (κ), visual semantic variation under semantic change (γ_w^I) and maintenance ($\gamma_{w/o}^I$). For each sample, $\bar{S}_{ii} = \frac{1}{|K|} \sum_{i \in K} S(T_i, I_i)$, where $K = \{a, pv, pi\}$. High \bar{S}_{ii} , γ_w^I and κ , coupled with low $\gamma_{w/o}^I$, indicate a model’s strong causality of input to output. For brevity, we denote \bar{S}_{ii} , γ_w^I , and $\gamma_{w/o}^I$ as \bar{S} , γ_w , and γ_{wo} .

Evaluation Dataset. We evaluate T2I models on the test set in a zero-shot manner. To demonstrate the improvements from fine-tuning, we collected sentences and their generated images from the training set, selecting only those with high quality, high discrimination, and consistent variations as the training data. Details about the selection of the training data are provided in Appendix D.3.

4.2 RESULTS

The results of the influence of inputs semantic variations on outputs semantic variations in T2I synthesis are shown in Tab. 2. The scores for \bar{S} range between 0.6 and 0.8. Despite the alignment score \bar{S} reaching up to 0.8, this does not imply a strong grasp of semantics. The following three metrics provide a more comprehensive view of the model’s ability to handle semantic variations.

Visual Semantic Variation with Changed Textual Semantics. As shown in Tab. 2, the values of γ_w are all below 0.52 for all evaluators, significantly lower than the optimal value of 1. This indicates that none of the T2I models perform at an acceptable level. These models are highly insensitive to semantic variations. This finding aligns with the widely accepted notion that T2I models tend to treat input text as a collection of isolated words, leading them to interpret sentences with minor changes in word order as having the same meaning.

324	Models	Gemini 1.5 Pro				Claude 3.5 Sonnet				GPT-4o				GPT-4 Turbo			
325		$\bar{S}(\uparrow)$	$\gamma_w(\uparrow)$	$\gamma_{wo}(\downarrow)$	$\kappa(\uparrow)$												
326	Open-source Models																
327	SD 1.5	0.55	0.43	0.46	-0.03	0.64	0.19	0.20	-0.01	0.63	0.34	0.33	0.01	0.65	0.32	0.32	0.00
328	SD 2.1	0.58	0.45	0.46	-0.01	0.66	0.21	0.20	0.00	0.65	0.33	0.31	0.02	0.68	0.35	0.34	0.01
329	SD XL 1.0	0.62	0.39	0.39	-0.00	0.69	0.19	0.18	0.00	0.71	0.31	0.28	0.03	0.72	0.32	0.28	0.03
330	SD CA	0.59	0.42	0.41	0.01	0.69	0.19	0.18	0.01	0.67	0.31	0.31	-0.00	0.69	0.32	0.31	0.01
331	DeepFloyd	0.64	0.44	0.44	0.00	0.71	0.20	0.19	0.01	0.69	0.33	0.30	0.03	0.74	0.33	0.28	0.05
332	PixArt	0.60	0.35	<u>0.32</u>	0.02	0.69	0.17	0.15	0.02	0.70	0.29	<u>0.26</u>	0.03	0.71	0.29	0.27	0.02
333	Kolors	0.60	0.41	0.42	-0.01	0.69	0.22	0.22	-0.01	0.69	0.31	0.30	0.01	0.69	0.33	0.30	0.02
334	SD 3	0.67	0.45	0.40	0.05	0.76	0.23	0.19	0.04	0.75	0.36	0.29	0.07	0.76	0.33	0.28	0.05
335	FLUX.1	0.72	0.43	0.35	0.08	0.75	0.23	<u>0.17</u>	0.06	0.72	0.42	0.33	0.10	0.75	0.40	0.30	0.10
336	API-based Models																
337	MidJ V6	0.68	0.46	0.39	0.07	0.73	0.24	0.21	0.03	0.72	0.40	0.33	0.07	0.73	0.38	0.32	0.06
338	DALL-E 3	0.75	0.46	0.33	0.14	0.80	0.25	0.18	0.06	0.82	0.36	0.22	<u>0.13</u>	0.83	0.35	0.30	0.10
339	CogV3-Plus	<u>0.79</u>	0.52	0.35	<u>0.17</u>	0.80	0.28	0.18	0.10	<u>0.81</u>	0.49	0.28	0.20	<u>0.82</u>	0.43	<u>0.26</u>	0.17
340	Ideogram 2	0.80	<u>0.47</u>	0.29	0.18	<u>0.79</u>	<u>0.26</u>	<u>0.17</u>	0.09	<u>0.81</u>	<u>0.46</u>	0.27	0.20	<u>0.81</u>	<u>0.40</u>	0.24	<u>0.15</u>

Table 2: Evaluations on T2I models (left column) using semantic variations under multiple MLLM scores (top row). $\bar{S}_{ii}(\uparrow)$: text-image alignment score. $\kappa(\uparrow)$: SemVar-Effect, measuring input-to-output variation contribution. $\gamma_w^I(\uparrow)$ and $\gamma_{wo}^I(\downarrow)$: visual semantic variation under semantic change and maintenance. **Bold** and underline indicate 1st and 2nd optimal cases. Blue and green indicates average SemVarEffect scores between 0.05 and 0.10, and above 0.10, respectively.

Visual Semantic Variation with Unchanged Textual Semantics. The values of γ_{wo} in Tab. 2 are unexpectedly much higher than the optimal value of 0. Only the models highlighted in blue and green demonstrate slightly better performance, with γ_{wo} scores consistently lower than γ_w . These T2I models illustrate potential semantic variations caused by word order through images, yet struggle to differentiate between meaning-variant and meaning-invariant inputs. These models primarily understand language based on word order rather than the underlying semantics.

Influence of Textual Semantics on Visual Semantic Variations. In Tab. 2, the κ values for all evaluators are below 0.20, indicating considerable room for improvement in T2I models’ understanding of semantic variations. Models with higher alignment scores are more sensitive to semantic variations caused by word orders. However, models highlighted in blue overreact to permutations maintaining the meanings, resulting in higher γ_{wo} values and subsequently lower κ values. These models excel at capturing common alignments but struggles to handle semantic variations.

Human Evaluation. We observe consistent performance trends between human raters and the four MLLMs across all evaluated models. Correlation analysis on CogView3-Plus reveals moderate (up to 0.54) correlation coefficients between machine and human scores, suggesting our selected MLLMs can serve as a reliable proxy for human evaluation. Details are shown in Appendix E.2.

4.3 ANALYSIS

Is a superior text encoder the exclusive solution for T2I models to grasp semantic variations? We explore the relationship between the text encoder’s ability to discriminate semantic variations and the ability of two metrics—alignment scores \bar{S} and visual semantic variation scores γ —to do the same, as illustrated in Fig. 6. We use text similarity¹ to measure the text encoder’s discriminative capability for semantic variations. T2I models like PixArt and Kolors, which utilize T5 and ChatGLM as text encoders, fail to transfer the results of distinguishing semantic variations to image generators, as shown by permutation-variance (indicated by squares). However, T2I models like FLUX.1, which utilize weaker CLIP-T5 hybrid models as text encoders, achieve higher alignment scores and greater differentiation in visual semantic variation scores, despite showing minimal changes in text similarity. These results indicate that a model’s ability to distinguish semantic variations is not only dependent on the text encoder, and that further efforts are needed in cross-modal alignment to effectively transfer these differences to the image generators.

¹Sentences for changed textual semantics unexpectedly show higher text similarity than those for unchanged textual semantics, likely due to the edit distance between our sentences. For further analysis, see Appendix F.

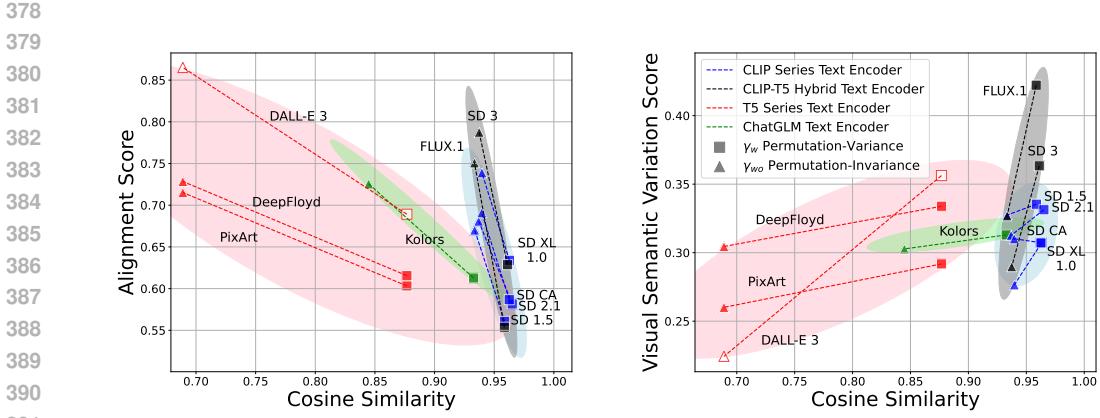


Figure 6: Illustration of the text embedding similarity between the anchor text and the permuted text. Squares represent permutation-variance results (with changed textual semantics), while triangles represent permutation-invariance results (with unchanged textual semantics). The evaluator is GPT-4o. (a) The alignment score between the anchor image I_a and a permutation T_{p^*} decreases as the text similarity between T_a and T_{p^*} increases. (b) The semantic variation score γ increases as the text similarity between T_a and T_{p^*} increases. The cosine similarity for DALL-E 3, an API-driven model, is deduced using T5-XXL, indicated by hollow shapes.

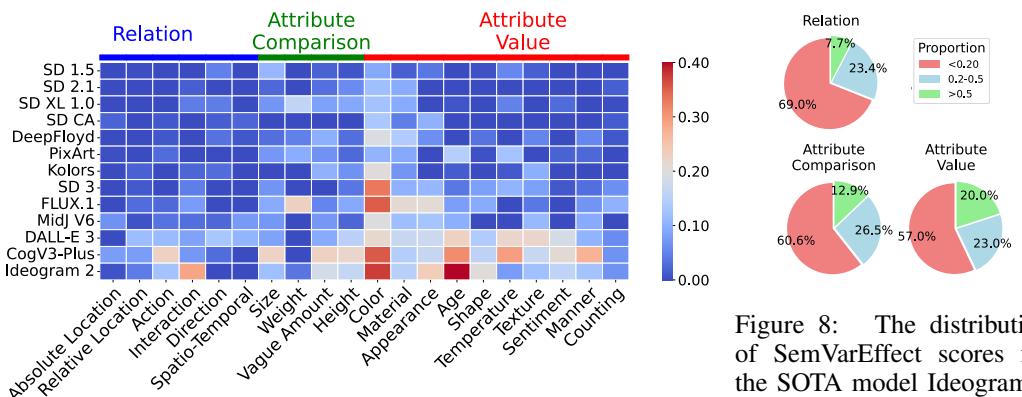


Figure 7: The distribution of categories across different T2I models based on SemVarEffect scores κ . The evaluator is GPT-4 Turbo.

Figure 8: The distribution of SemVarEffect scores for the SOTA model Ideogram 2 across different aspects of the samples. The evaluator is GPT-4 Turbo.

Does the influence of input semantic variations on output semantic variations vary by category? As shown in Fig. 7, the semantic variations in Color have significantly influenced the output of T2I models, with the SemVarEffect score consistently exceeding 0.4 in many models. In contrast, the SemVarEffect scores in other categories are mostly below 0.1. This suggests that T2I models understand semantic variations well only in the case of Color. We found that the SemVarEffect scores of Ideogram 2 in Relation, Attribute Comparison, and Attribute Value are 0.07, 0.13, and 0.19. To compare the distribution of SemVarEffect scores across different aspects, we set 0.2 and 0.5 as thresholds. As shown in Fig. 8, the proportion of high scores in Attribute Values is significantly higher than those in Relation and Attribute Comparison. T2I models lack the capability to discriminate semantic variations, particularly in aspects emphasizing relations and comparisons. Fig. 9 shows failed examples in Relation and Comparison. Although T2I models can generate correct images for common relations, they tend to rigidly adhere to these common relations even when semantic variations occur, leading to incorrect images. More examples are provided in Appendix G.

Does fine-tuning improve T2I model performance on semantic variations? We examine improvements from fine-tuning text encoders and image generators. We use samples in the training set to generated images and select text-images pairs with high alignment scores and high discriminability as training data, details shown in Appendix D.3. As shown in Tab. 3, for categories with sufficient high-quality data, such as Color, supervised fine-tuning (SFT) enhanced the performance

of the T2I model. However, in categories with insufficient high-quality data, such as Direction, SFT led to a decline in performance. Additionally, direct preference optimization (DPO) resulted in performance drops due to failures in permutation-invariance, as evidenced by the increased r_{wo} .

It is crucial to strike a balance between sensitivity and robustness to semantic changes, as this determines whether performance can be enhanced. However, fine-tuning tends to improve sensitivity at the expense of robustness. While T2I models become more sensitive to permutations with different meanings, this discrimination is quickly disrupted by over-sensitivity to permutations with similar meanings, leading to a decline in the model’s overall ability to discern differences. This phenomenon may be attributed to two potential limitations in Diffusion-based T2I Models. First, its cross-attention mechanism only maps tokens to spatial regions without capturing inter-token relationships. Second, the training process lacks semantic-level supervision. While fine-tuning improves token-region correspondence, it cannot enhance the understanding of semantic relationships between tokens. Thus, permutation-variance samples, which differ only in word order but contain identical tokens, challenge the model’s semantic understanding capabilities. This confuses the models and leads to performance declines, especially during DPO. More validation is in Appendix F.4.

Category	Models	GPT-4o			
		$\bar{S}(\uparrow)$	$\gamma_w(\uparrow)$	$\gamma_{wo}(\downarrow)$	$\kappa(\uparrow)$
Color	SD XL	0.73	0.33	0.25	0.08
	+ sft-unet	0.78(↑)	0.38(↑)	0.20(↓)	0.18(↑)
	+ sft-text	0.73(−)	0.40(↑)	0.27(↑)	0.13(↑)
	+ dpo-unet	0.69(↓)	0.43(↑)	0.27(↑)	0.17(↑)
	+ dpo-text	0.68(↓)	0.47(↑)	0.29(↑)	0.18(↑)
Absolute Location	SD XL	0.64	0.29	0.34	-0.05
	+ sft-unet	0.65(↑)	0.34(↑)	0.32(↓)	0.02(↑)
	+ sft-text	0.64(−)	0.31(↑)	0.36(↑)	-0.05(−)
	+ dpo-unet	0.60(↓)	0.29(↑)	0.31(↓)	-0.02(↑)
	+ dpo-text	0.57(↓)	0.33(↑)	0.39(↑)	-0.07(↓)
Height	SD XL	0.77	0.34	0.23	0.10
	+ sft-unet	0.77(−)	0.33(↓)	0.24(↑)	0.09(↓)
	+ sft-text	0.73(↓)	0.39(↑)	0.34(↑)	0.05(↓)
	+ dpo-unet	0.71(↓)	0.34(−)	0.33(↑)	0.02(↓)
	+ dpo-text	0.66(↓)	0.40(↑)	0.53(↑)	-0.13(↓)
Direction	SD XL	0.79	0.20	0.15	0.05
	+ sft-unet	0.77(↓)	0.24(↑)	0.23(↑)	0.01(↓)
	+ sft-text	0.77(↓)	0.23(↑)	0.21(↑)	0.02(↓)
	+ dpo-unet	0.65(↓)	0.23(↑)	0.26(↑)	-0.03(↓)
	+ dpo-text	0.70(↓)	0.29(↑)	0.27(↑)	0.01(↓)

Table 3: Fine-tuned SD XL Results: Performance varies based on the quantity of high-quality samples, which are determined by category and sample size. Left Table: Color and Absolute Location with 4.4k and 1.7k candidates for training. Right Table: Height and Direction with 0.2k and 0.3k candidates for training.

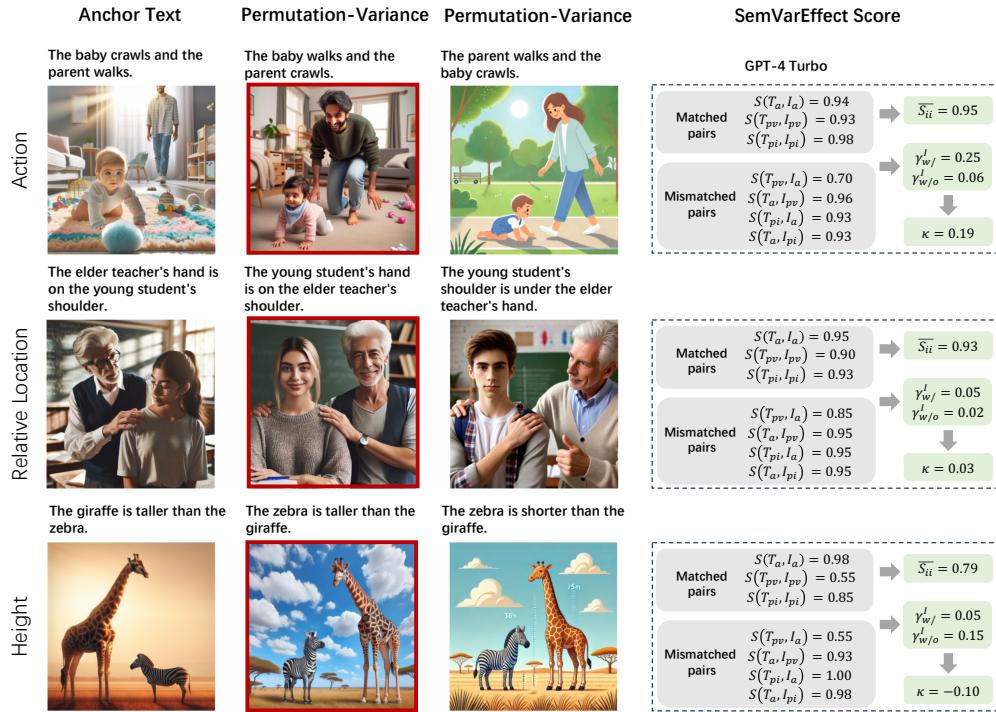


Figure 9: Failed examples of DALL-E 3 on Relation and Attribute Comparison.

486 **Do T2I models’ struggles with semantic relationships stem from training data imbalance?** We
 487 conducted experiments testing performance of fine-tuned SDXL trained with balanced training data.
 488 We used a human-crafted balanced dataset of cat↔dog chasing interactions (80 images per direc-
 489 tion) for training, and tested on two unseen prompt pairs involving the same “chasing” relationship
 490 between common objects. The model showed consistently poor generalization: accuracy remained
 491 low for both anti-commonsense scenarios (mouse↔cat, 3-3/30) and plausible scenarios (bull↔man,
 492 4-5/30). Failure analysis revealed three main categories: (1) Relationship Understanding Failures,
 493 where objects appear either without interaction or with incorrect interactions, indicating the model’s
 494 inability to comprehend the “chasing” concept; (2) Reversed Roles, where the model fails to prop-
 495 erly assign who chases whom; and (3) Missing Objects, where the model fails to generate all re-
 496 quired objects. Even among generated images, correct relationships occurred less frequently than
 497 these failure cases, suggesting random performance rather than true understanding. Thus, the mod-
 498 els’ struggles persist even with perfectly balanced training data, suggesting the core issue lies in
 499 relationship understanding rather than data imbalance. Details are in Appendix F.3.4.

500 5 RELATED WORK

503 **Evaluation of T2I synthesis.** Benchmarks of T2I synthesis primarily focus on general align-
 504 ment Saharia et al. (2022); Yu et al. (2022); Cho et al. (2022), composition Park et al. (2021);
 505 Feng et al. (2023); Park et al. (2021); Hu et al. (2023); Cho et al. (2023b); Li et al. (2024), bias
 506 and fairness Lee et al. (2023b); Luo et al. (2024b;a), common sense Fu et al. (2024) and creativ-
 507 ity Lee et al. (2023b). In these evaluations, the quality of images is measured by detection-based
 508 or alignment-based metrics. Recent research on T2I synthesis has explored samples involving se-
 509 mantic variations caused by word orders, typically using them to evaluate reasoning abilities with
 510 alignment-based metrics Marcus et al. (2022); Lee et al. (2023b); Li et al. (2024). However, a sig-
 511 nificant gap in this research is the underexplored area of whether the generated images consistently
 512 represent fundamental semantic variations within the input text.

513 **Semantic Variation Evaluation in VLMs.** In VLMs, semantic variations caused by word order has
 514 been evaluated by benchmarks like Winoground Thrush et al. (2022) and its expansion in specific
 515 domain Burapacheep et al. (2024). Winoground is designed to challenge models with visio-linguistic
 516 compositional reasoning. It requires models to accurately match two images to their respective
 517 captions, where the two captions are different permutations of the same set of words, resulting in
 518 different meanings. To enhance performance on Winoground, studies have focused on expanding
 519 training datasets with negative samples and optimizing training strategies to handle the resulting
 520 semantic variations Yüksekgönül et al. (2023); Hsieh et al. (2024); Burapacheep et al. (2024).

521 The application of Winoground to T2I synthesis faces several limitations due to the variety and
 522 quantity of its permutations. First, the dataset, with 400 sentence pairs, provides only 171 suitable
 523 for text-image composition analysis Diwan et al. (2022), where samples are classified into three cat-
 524 egories: object, relation, and both. This limited variety is insufficient for a comprehensive evaluation
 525 of T2I models. Second, the suitability of certain samples for T2I model evaluation is problematic.
 526 Winoground primarily focuses on semantic distinctiveness for cross-modal retrieval Yüksekgönül
 527 et al. (2023); Ma et al. (2023); Cascante-Bonilla et al. (2023). It overlooks the criteria essential for
 528 T2I synthesis, such as sentence completeness, clarity of expression, unambiguity, and specificity
 529 in referencing image elements. All of these factors have been carefully considered in the quality
 530 control of our benchmark annotations.

531 6 CONCLUSION

533 We comprehensively study the challenge of semantic variations in T2I synthesis, specifically fo-
 534 cusing on causality between semantic variations of inputs and outputs. We propose a new metric,
 535 SemVarEffect, to quantify the influence of input semantic variations on model outputs, and a novel
 536 benchmark, SemVarBench, designed to examine T2I models’ understanding of semantic variations.
 537 Our experiments reveal that SOTA T2I models struggle with semantic variations, scoring below 0.2
 538 on our benchmark. Fine-tuning shows limited improvement, improving sensitivity but at the cost of
 539 robustness. These findings underscore the need for better cross-modal understanding of relation in
 540 semantics, particularly for capturing inter-token dependencies in T2I synthesis.

540 REFERENCES
541

- 542 Anthropic. Claude 3.5 sonnet, 2024. URL <https://www.anthropic.com/news/claude-3-5-sonnet>.
- 543
- 544 Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and
545 Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image
546 models. *CoRR*, abs/2304.05390, 2023. doi: 10.48550/ARXIV.2304.05390. URL <https://doi.org/10.48550/arXiv.2304.05390>.
- 547
- 548
- 549 James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang
550 Zhuang, Joyce Lee, Yafei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao,
551 and Aditya Ramesh. Improving image generation with better captions. 2023. URL <https://cdn.openai.com/papers/dall-e-3.pdf>.
- 552
- 553
- 554 Jirayu Burapacheep, Ishan Gaur, Agam Bhatia, and Tristan Thrush. Colorsnap: A color and word
555 order dataset for multimodal evaluation. In Lun-Wei Ku, Andre Martins, and Vivek Srikanth
556 (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand*
557 and virtual meeting, August 11-16, 2024, pp. 1716–1726. Association for Computational Linguis-
558 tics, 2024. URL <https://aclanthology.org/2024.findings-acl.99>.
- 559
- 560 Pu Cao, Feng Zhou, Qing Song, and Lu Yang. Controllable generation with text-to-image diffusion
561 models: A survey. *arXiv preprint arXiv:2403.04279*, 2024.
- 562
- 563 Paola Cascante-Bonilla, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim,
564 Rameswar Panda, Gülden Varol, Aude Oliva, Vicente Ordonez, Rogério Feris, and Leonid Karlinsky.
565 Going beyond nouns with vision & language models using synthetic data. *CoRR*, abs/2303.17590,
566 2023. doi: 10.48550/ARXIV.2303.17590. URL <https://doi.org/10.48550/arXiv.2303.17590>.
- 567
- 568 Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James
569 Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- α : Fast training of diffusion transformer
570 for photorealistic text-to-image synthesis, 2023.
- 571
- 572 Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin
573 Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, et al. Mj-bench: Is your multimodal reward
574 model really a good judge for text-to-image generation? *arXiv preprint arXiv:2407.04842*, 2024.
- 575
- 576 Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social
577 biases of text-to-image generative transformers. *CoRR*, abs/2202.04053, 2022.
- 578
- 579 Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit
580 Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-
581 grained evaluation for text-to-image generation. *CoRR*, abs/2310.18235, 2023a. doi: 10.48550/
582 ARXIV.2310.18235. URL <https://doi.org/10.48550/arXiv.2310.18235>.
- 583
- 584 Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for text-to-image generation and
585 evaluation. *CoRR*, abs/2305.15328, 2023b. doi: 10.48550/ARXIV.2305.15328. URL <https://doi.org/10.48550/arXiv.2305.15328>.
- 586
- 587 Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. Why is winoground
588 hard? investigating failures in visuolinguistic compositionality. In *Proceedings of the 2022*
589 *Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi,*
590 *United Arab Emirates, December 7-11, 2022*, pp. 2236–2250. Association for Computational
591 Linguistics, 2022.
- 592
- 593 Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam
Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for
high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*,
2024.

- 594 Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun R. Akula, Pradyumna Narayana, Sugato
 595 Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for
 596 compositional text-to-image synthesis. In *The Eleventh International Conference on Learning
 597 Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.
- 598 Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu,
 599 Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and gen-
 600 eration with large language models. *Advances in Neural Information Processing Systems*, 36,
 601 2024.
- 602 Xingyu Fu, Muyu He, Yujie Lu, William Yang Wang, and Dan Roth. Commonsense-t2i chal-
 603 lenge: Can text-to-image generation models understand commonsense? *arXiv preprint
 604 arXiv:2406.07546*, 2024.
- 605 Hanan Gani, Shariq Farooq Bhat, Muzammal Naseer, Salman Khan, and Peter Wonka. Llm
 606 blueprint: Enabling text-to-image generation with complex and detailed prompts. *arXiv preprint
 607 arXiv:2310.10640*, 2023.
- 608 Matthias Gerner. Predicate-induced permutation groups. *J. Semant.*, 29(1):109–144, 2012. doi:
 609 10.1093/JOS/FFR007. URL <https://doi.org/10.1093/jos/ffr007>.
- 610 Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta
 611 Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv
 612 preprint arXiv:2212.10015*, 2022.
- 613 Jonathan Gordon, David Lopez-Paz, Marco Baroni, and Diane Bouchacourt. Permutation equiv-
 614 ariant models for compositional generalization in language. In *8th International Conference
 615 on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenRe-
 616 view.net, 2020. URL <https://openreview.net/forum?id=SylVNerFvr>.
- 617 Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrape:
 618 Fixing hackable benchmarks for vision-language compositionality. *Advances in neural informa-
 619 tion processing systems*, 36, 2024.
- 620 Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A.
 621 Smith. TIFA: accurate and interpretable text-to-image faithfulness evaluation with question an-
 622 swering. *CoRR*, abs/2303.11897, 2023. doi: 10.48550/ARXIV.2303.11897. URL <https://doi.org/10.48550/arXiv.2303.11897>.
- 623 Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehen-
 624 sive benchmark for open-world compositional text-to-image generation. *CoRR*, abs/2307.06350,
 625 2023.
- 626 Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Ad-
 627 vances in neural information processing systems*, 34:21696–21707, 2021.
- 628 Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-
 629 a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural
 630 Information Processing Systems*, 36:36652–36663, 2023.
- 631 Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel,
 632 Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human
 633 feedback. *arXiv preprint arXiv:2302.12192*, 2023a.
- 634 Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yun-
 635 zhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, Minguk Kang, Taesung
 636 Park, Jure Leskovec, Jun-Yan Zhu, Li Fei-Fei, Jiajun Wu, Stefano Ermon, and Percy Liang. Holis-
 637 tic evaluation of text-to-image models. *CoRR*, abs/2311.04287, 2023b. doi: 10.48550/ARXIV.
 638 2311.04287. URL <https://doi.org/10.48550/arXiv.2311.04287>.
- 639 Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Xide Xia, Pengchuan Zhang,
 640 Graham Neubig, and Deva Ramanan. Evaluating and improving compositional text-to-visual
 641 generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog-
 642 nition*, pp. 5290–5301, 2024.

- 648 Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual
 649 generation with composable diffusion models. In *European Conference on Computer Vision*, pp.
 650 423–439. Springer, 2022.
- 651 Hanjun Luo, Ziye Deng, Ruizhe Chen, and Zuozhu Liu. Faintbench: A holistic and precise bench-
 652 mark for bias evaluation in text-to-image models. *arXiv preprint arXiv:2405.17814*, 2024a.
- 653 Hanjun Luo, Haoyu Huang, Ziye Deng, Xuecheng Liu, Ruizhe Chen, and Zuozhu Liu. Bigbench: A
 654 unified benchmark for social bias in text-to-image generative models based on multi-modal llm.
 655 *arXiv preprint arXiv:2407.15240*, 2024b.
- 656 Hanjun Luo, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. @
 657 CREPE: can vision-language foundation models reason compositionally? In *IEEE/CVF Con-
 658 ference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada,
 659 June 17-24, 2023*, pp. 10910–10921. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01050. URL
 660 <https://doi.org/10.1109/CVPR52729.2023.01050>.
- 661 Gary Marcus, Ernest Davis, and Scott Aaronson. A very preliminary analysis of DALL-E 2. *CoRR*,
 662 abs/2204.13807, 2022. doi: 10.48550/ARXIV.2204.13807. URL <https://doi.org/10.48550/arXiv.2204.13807>.
- 663 Wenyi Mo, Tianyu Zhang, Yalong Bai, Bing Su, Ji-Rong Wen, and Qing Yang. Dynamic prompt op-
 664 timizing for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer
 665 Vision and Pattern Recognition*, pp. 26627–26636, 2024.
- 666 Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for
 667 compositional text-to-image synthesis. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Pro-
 668 ceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1,
 669 NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.
- 670 Pablo Pernias, Dominic Rampas, Mats L. Richter, Christopher J. Pal, and Marc Aubreville. Wuer-
 671 stchen: An efficient architecture for large-scale text-to-image diffusion models, 2023.
- 672 Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe
 673 Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image
 674 synthesis. *CoRR*, abs/2307.01952, 2023. doi: 10.48550/ARXIV.2307.01952. URL <https://doi.org/10.48550/arXiv.2307.01952>.
- 675 Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-
 676 resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer
 677 Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp.
 678 10674–10685. IEEE, 2022.
- 679 Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Den-
 680 ton, Seyed Kamyar Seyed Ghasempour, Raphael Gontijo Lopes, Burcu Karagol Ayan,
 681 Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic
 682 text-to-image diffusion models with deep language understanding. In *NeurIPS*,
 683 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html.
- 684 Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of
 685 score-based diffusion models. *Advances in neural information processing systems*, 34:1415–
 686 1428, 2021.
- 687 Kolors Team. Kolors: Effective training of diffusion model for photorealistic text-to-image synthe-
 688 sis. *arXiv preprint*, 2024.
- 689 Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and
 690 Candace Ross. Winoground: Probing vision and language models for visio-linguistic compo-
 691 sitionality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022,
 692 New Orleans, LA, USA, June 18-24, 2022*, pp. 5228–5238. IEEE, 2022.

702 Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam,
 703 Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using
 704 direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision
 705 and Pattern Recognition*, pp. 8228–8238, 2024.

706 Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu,
 707 and Lijuan Wang. Equivariant similarity for vision-language foundation models. In *IEEE/CVF
 708 International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp.
 709 11964–11974. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01102. URL [https://doi.org/
 710 10.1109/ICCV51070.2023.01102](https://doi.org/10.1109/ICCV51070.2023.01102).

711 Zhenyu Wang, Enze Xie, Aoxue Li, Zhongdao Wang, Xihui Liu, and Zhenguo Li. Divide and
 712 conquer: Language models can plan and self-correct for compositional text-to-image generation.
 713 *arXiv preprint arXiv:2401.15688*, 2024.

714 Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score:
 715 Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF
 716 International Conference on Computer Vision*, pp. 2096–2105, 2023.

717 Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and CUI Bin. Mastering text-
 718 to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Forty-first
 719 International Conference on Machine Learning*, 2024.

720 Chang Yu, Junran Peng, Xiangyu Zhu, Zhaoxiang Zhang, Qi Tian, and Zhen Lei. Seek for incan-
 721 tations: Towards accurate text-to-image diffusion synthesis through prompt engineering. *arXiv
 722 preprint arXiv:2401.06345*, 2024.

723 Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Va-
 724 sudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana
 725 Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive mod-
 726 els for content-rich text-to-image generation. *Trans. Mach. Learn. Res.*, 2022, 2022. URL
 727 <https://openreview.net/forum?id=AFDcYJKhND>.

728 Mert Yüksekgönül, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When
 729 and why vision-language models behave like bags-of-words, and what to do about it? In *The
 730 Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda,
 731 May 1-5, 2023*. OpenReview.net, 2023. URL <https://openreview.net/forum?id=KRLUvxh8uaX>.

732 Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan,
 733 William Yang Wang, and Linda Ruth Petzold. Gpt-4v (ision) as a generalist evaluator for vision-
 734 language tasks. *arXiv preprint arXiv:2311.01361*, 2023.

735
 736
 737
 738
 739
 740
 741
 742
 743
 744
 745
 746
 747
 748
 749
 750
 751
 752
 753
 754
 755

756 The Appendix is organized as follows:
 757

- 758 • Section A provides detailed illustrations of four types of semantic variation results in T2I
 759 synthesis.
- 760 • Section B describes the data requirements, as well as the properties of the alignment func-
 761 tion and the SemVarEffect score.
- 762 • Section C details the construction process of the benchmark.
- 763 • Section D presents the implementation details of the evaluation.
- 764 • Section E presents more experimental results.
- 765 • Section F provides further analysis of the results.
- 766 • Section G visualizes more successful and failed examples in the evaluation.
- 767 • Section H discusses the limitations of our evaluation and benchmark.

770 A FOUR TYPES OF SEMANTIC VARIATIONS RESULTS IN T2I SYNTHESIS

772 The results of semantic variations in T2I synthesis, both in text and images, can be divided into four
 773 types, as shown in Fig. 10.

- 775 • **Image Changing Semantics with Text Changing Semantics:** The semantic consistency
 776 between the input and output in the first quadrant suggests that the model tends to under-
 777 stand the different semantics introduced by linguistic permutations. In this case, the value
 778 of $\gamma_{w/o}$ will tend to 1.
- 779 • **Image Maintaining Semantics with Text Changing Semantics:** The semantic inconsis-
 780 tency between the input and output in the fourth quadrant suggests that the model does not
 781 understand the different semantics introduced by linguistic permutations. In this case, the
 782 value of $\gamma_{w/o}$ will tend to 0.
- 783 • **Image Changing Semantics with Text Maintaining Semantics:** The semantic consis-
 784 tency between the input and output in the third quadrant suggests that the model tends to
 785 understand the similar semantics introduced by linguistic permutations. In this case, the
 786 value of $\gamma_{w/o}$ will tend to 0.
- 787 • **Image Maintaining Semantics with Text Maintaining Semantics:** The semantic inconsis-
 788 tency between the input and output in the second quadrant suggests that the model does
 789 not understand the similar semantics introduced by linguistic permutation. In this case, the
 790 value of $\gamma_{w/o}$ will tend to 1.

792 B PROPERTIES OF SEMVAREFFECT

794 B.1 PRELIMINARY

796 A T2I generation model f consists of one or more text encoders, image generators, and an image
 797 decoder. The T2I generation model f generates images $I \in \mathcal{I}$ for each input textual prompt $T \in \mathcal{T}$.
 798 \mathcal{T} represents the textual space, and \mathcal{I} represents the visual space. $S(T, I)$ is the alignment score
 799 between T and I .

800 Let T_a be an anchor textual prompt. Let T_{p*} represent a permutation of T_a , where T_{pv} is a permuta-
 801 tion with a different meaning than T_a , and T_{pi} is a permutation with the same meaning as T_a . Let I_a ,
 802 I_{pv} , and I_{pi} be the resulting images generated by a T2I model from T_a , T_{pv} , and T_{pi} , respectively.
 803 We expect that I_{p*} should be a rearrangement of the objects or relations found within I_a .

805 B.2 TEXTUAL VS. VISUAL SEMANTIC VARIATIONS

807 The measurement of semantic variations in the transition from (T_a, I_a) to (T_{p*}, I_{p*}) can be defined
 808 from two perspectives: (1) textual semantic variations r^T : The semantic changes in the text, ob-
 809 served through differences between images I_a and I_{p*} , and (2) visual semantic variations r^I : The
 semantic changes in the images, observed through differences between texts T_a and T_{p*} .

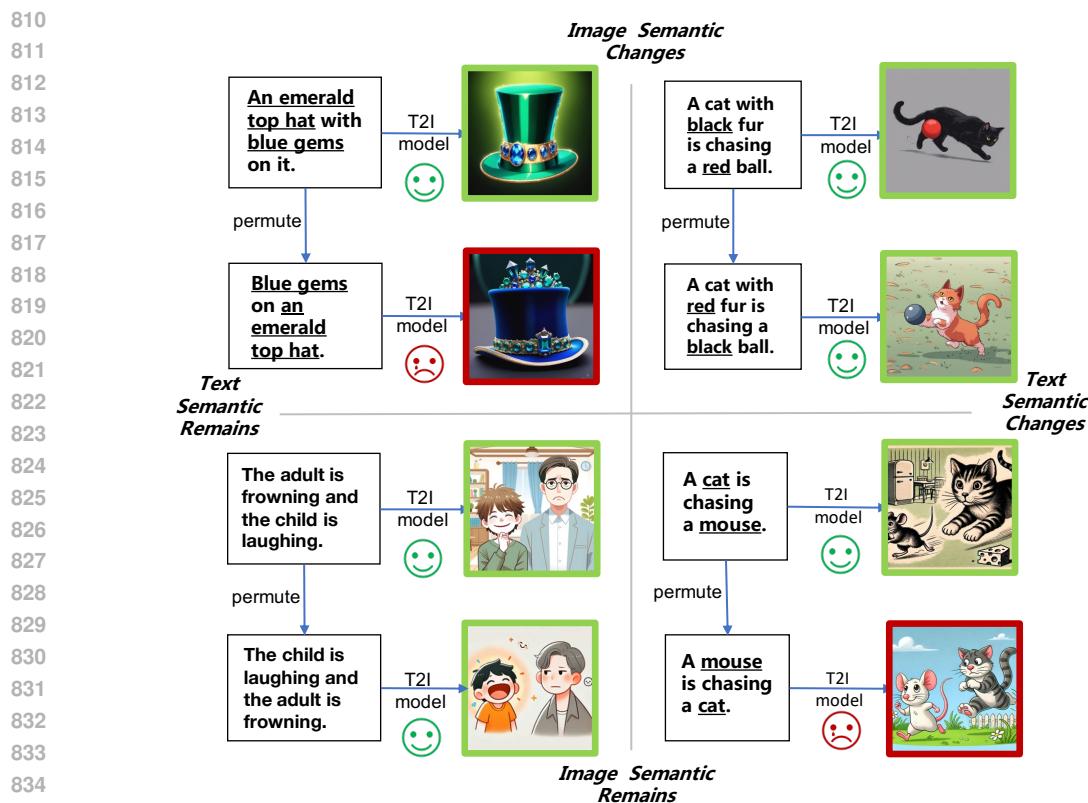


Figure 10: Four types of semantic variation results in T2I synthesis. The images with a red border represent an incorrect output. The images with a green border represent a correct output.

Specifically, we define the textual semantic variations observed from a single image I . The initial anchor sentence T_a is transformed into a permutation T_{p*} after a series of localized linguistic permutations. For each change from T to $T + \Delta T$ in the text space, the textual semantic variation at position T , denoted as $\mu_T(T, I)$, is defined as the difference in alignment scores between the modified text and the original text with the image: $\mu_T(T, I) = S(T + \Delta T, I) - S(T, I)$. The textual semantic variation from T_a to T_{p*} is defined as the sum of the semantic variations produced by each localized permutation: $\sum_{T_a}^{T_{p*}} \mu_T(T, I) = S(T_{p*}, I) - S(T_a, I)$. Therefore, the integrated textual semantic variation is defined as $\gamma^T = \sum_{T \in \{T_a, T_{p*}\}} \left| \sum_{T_a}^{T_{p*}} \mu(T, I) \right|$.

Similarly, we define the visual semantic variation observed from a single textual prompt T . Let $\mu_I(T, I)$ be the visual semantic variation observed from a given text T in the image space. The visual semantic variation from I_a to I_{p*} is defined as the sum of the visual semantic variations produced by each localized modification: $\sum_{I_a}^{I_{p*}} \mu_I(T, I) = S(I_{p*}, T) - S(I_a, T)$. Therefore, the integrated visual semantic variation is defined as $\gamma^I = \sum_{T \in \{T_a, T_{p*}\}} \left| \sum_{I_a}^{I_{p*}} \mu(T, I) \right|$.

We have not evaluated the influence by measuring the synchronicity of semantic changes between images and text, which has been applied in VLM Wang et al. (2023). This is because semantic variations introduce a unique challenge in the evaluation of T2I synthesis: images are not independent; they are influenced by both the input textual prompt and the inherent characteristics of the model, complicating the independent assessment of semantic changes.

Therefore, we conduct the evaluation by measuring the influence of external intervention on semantic variations within the corresponding images of T2I synthesis to avoid directly imposing interventions on images.

864 **B.3 PROPERTIES OF ALIGNMENT SCORES S**
 865

866 **Definition of Text-Image Alignment Score.** To facilitate semantic analysis, we structured these
 867 permutations by objects and triples. Changing the word order affects the arrangement of objects
 868 or relations and leads to changes in syntactic dependencies and semantics. Let T_{p*} represent any
 869 permutation of T_a . T_a and T_{p*} share the same set of objects set V and set of relations R . The triple
 870 set E in T_a is a subset of $V \times R \times V$. Some triples in T_{p*} may differ from those in T_a , but they have
 871 the same number of triples. For example, the initial triple set of T_a contains (apple, on, box),
 872 (girl, touch, apple) and (girl, NULL, box). After swapping box and apple, the
 873 triple set of T_{p*} contains (box, on, apple), (girl, touch, box) and (girl, NULL,
 874 apple).
 875 To calculate the fine-grained alignment scores for objects and triples, we define the alignment score
 876 S between T and I as the sum of the object and triple alignment scores:
 877

$$878 \quad S(T, I) = \sum_{i=1}^{|V|} S_{obj_i}(T, I) + \sum_{j=1}^{|E|} S_{tri_j}(T, I), \quad (5)$$

880 where $|V|$ is the number of objects mentioned in T and $|E|$ is the number of triples mentioned in T .
 881 The components of the alignment score are defined as piecewise functions:
 882

$$883 \quad S_{obj_i}(T, I) = \begin{cases} w_{v_i} & \text{if the } i\text{-th object matches,} \\ 0 & \text{if the } i\text{-th object does not match,} \end{cases} \quad (6)$$

$$884 \quad S_{tri_j}(T, I) = \begin{cases} w_{e_j} & \text{if the } j\text{-th triple matches,} \\ 0 & \text{if the } j\text{-th triple does not match,} \end{cases}$$

885 where w_{v_i} and w_{e_j} are the weighted matching score for the i -th object and j -th triple. We obtain the
 886 alignment function S that satisfies the constraints in Eq. 10. Consequently, the alignment score of a
 887 matched text-image pair is calculated as:
 888

$$889 \quad S(T_{p*}, I_{p*}) = S(T_a, I_a) = \sum_{v_i \in V_{MA}} w_{v_i} + \sum_{e_j \in E_{MA}} w_{e_j}, \quad (7)$$

890 where V_{MA} and E_{MA} represent the exactly matched objects and triples between a text prompt and
 891 its generated image, with $|V_{MA}| = |V|$ and $|E_{MA}| = |E|$. The alignment score for a mismatched
 892 text-image pair is calculated as:
 893

$$894 \quad S(T_{p*}, I_a) = S(T_a, I_{p*}) = \sum_{v_i \in V_{MI}} w_{v_i} + \sum_{e_j \in E_{MI}} w_{e_j}, \quad (8)$$

895 where V_{MI} and E_{MI} represent the partially matched objects and triples between a text prompt and
 896 a mismatched image, with $|V_{MI}| = |V|$ and $0 \leq |E_{MI}| \leq |E|$.
 897 **Range of S .** If f accurately depicts the text through images and S faithfully measures the semantic
 898 changes between text space and image space, any alignment score $S(T, I)$ is bounded by:
 899

$$900 \quad \sum_{v_i \in V} w_{v_i} \leq S(T, I) \leq \sum_{v_i \in V} w_{v_i} + \sum_{e_j \in E} w_{e_j}. \quad (9)$$

901 In our implementation, we set the value of $S(T, I)$ as an integer between 0 and 100, where the
 902 object accuracy is between 0 and 50 and the triple accuracy is in between 0 and 50. Then we
 903 normalize it into a real number within the range [0, 1]. Based on the assumption of f mentioned
 904 above, $0.5 \leq S(T, I) \leq 1$.
 905 However, limitations in the capabilities of the model $f(\cdot)$ and the alignment function $S(\cdot)$, often
 906 prevent the alignment score values from achieving the property in Eq. 10. For example, if a model
 907 f generated a low-quality image I , it may fail to accurately depict all target objects (V), leading
 908 to $|V_{MA}| < |V|$ and $|V_{MI}| < |V|$. This results in an object accuracy below 0.5 (as illustrated
 909 in the bottom case of Fig. 24) and inconsistent relation accuracy (see cases in Fig. 19 and Fig.
 910 20). Furthermore, inaccuracies in the scoring approach $S(\cdot)$ may incorrectly evaluate the similarity
 911 between text prompts and generated images, causing unpredictable fluctuations in semantic variation
 912 measurements (as illustrated in Fig. 26).

918 **Identity Relation for S .** In the ideal scenario where f accurately transforms all semantic variations
 919 from text space to image space, the alignment scores would satisfy the constraints:
 920

$$S(T_a, I_a) \equiv S(T_{p*}, I_{p*}) \text{ and } S(T_{p*}, I_a) \equiv S(T_a, I_{p*}). \quad (10)$$

922 Eq. 10 is also demonstrated under the assumption that the alignment function is an equivariant
 923 map in the continuous textual feature space \mathcal{T} and visual feature space \mathcal{I} , as detailed in Wang
 924 et al. (2023). This assumption ensures that the alignment scores vary consistently with the semantic
 925 changes from images or text. The property is crucial for determining the characteristics of the data
 926 in SemVarBench and for designing the alignment functions.

927 However, the low-quality image I , resulting from a poorly performing model f , and the inaccuracies
 928 in the approach to scoring S , often prevent the alignment scores from meeting the criteria of Eq.
 929 10. These limitations make the direct comparison of textual and visual semantic variation scores
 930 unreliable in T2I synthesis. This approach was previously explored in Wang et al. (2023).

931 B.4 PROPERTY OF VISUAL SEMANTIC VARIATIONS γ^I

933 We analyze the theoretical relationship between visual semantic variations and model performance.
 934 According to Eqs. 1, 7 and 8, we derive the visual semantic variations as follows:

$$936 \gamma^I = 2 \left| \sum_{v_i \in \{V_{MA} - V_{MI}\}} w_{v_i} + \sum_{e_j \in \{E_{MA} - E_{MI}\}} w_{e_j} \right|. \quad (11)$$

939 For both permutation-variance and permutation-invariance, the value of $\sum_{v_i \in \{V_{MA} - V_{MI}\}} w_{v_i}$ re-
 940 mains constant, which we denote as C_1 . As a result, the visual semantic variation can be simplified
 941 to $\gamma^I = 2 \left| C_1 + \sum_{e_j \in \{E_{MA} - E_{MI}\}} w_{e_j} \right|$, primarily depending on the size of the set $E_{MA} - E_{MI}$.
 942 However, the size of the set varies dramatically between the two settings:

- 944 • In permutation-variance settings (T_a, T_{pv}), the optimal value for the set $E_{MA} - E_{MI}$ is
 945 its maximum set E , resulting in an observed positive correlation between visual semantic
 946 variation γ_w^I and model performance.
- 948 • In permutation-invariance settings (T_a, T_{pi}), the optimal value for the set $E_{MA} - E_{MI}$ is
 949 its minimum set \emptyset , resulting in an observed negative correlation between visual semantic
 950 variation $\gamma_{w/o}^I$ and model performance.

951 Therefore, we conclude that the visual semantic variations in permutation-variance and permutation-
 952 invariance differ significantly.

- 954 • A higher γ_w^I value indicates that the model effectively captures and reflects the intended
 955 semantic transformation in the input text.
- 956 • A lower $\gamma_{w/o}^I$ value indicates that the model maintains semantic consistency in the images
 957 despite variations in the input text.

959 B.5 PROPERTY OF SEMVAREFFECT SCORE κ

961 The SemVarEffect score on visual semantic variations, κ , is defined as the difference between γ_w^I /
 962 $\gamma_{w/o}^I$. It quantifies the model's ability to discriminate between significant and negligible seman-
 963 tic changes in the text.

- 965 • If κ is large, it suggests that the model is sensitive to semantic changes, recognizing vari-
 966 tions in meaning. However, this does not necessarily imply strong alignment. The model
 967 might detect changes in semantics but still struggle to fully capture all objects and relation-
 968 ships described in the text, indicating a gap between sensitivity and complete alignment.
- 969 • If κ is small or close to zero, it suggests that the model either fails to reflect meaningful
 970 semantic changes or overreacts to minor text variations. Regardless of the overall alignment
 971 score, the model may generate similar images even in the presence of significant semantic
 972 differences in the input text.

972 **C CONSTRUCTION DETAILS**
973
974

975 **C.1 DATA COLLECTION**
976
977

978

Seed Sentence Pairs from Winoground	Templates & Rule
<i>caption_0</i> : a bird eats a snake <i>caption_1</i> : a snake eats a bird	T_a : [Noun1] [Verb (vt)] [Noun2] T_{pv} : [Noun2] [Verb (vt)] [Noun1] $T_a \rightarrow T_{pv}$: [Noun1] \leftrightarrow [Noun2]
<i>caption_0</i> : a person is in a helicopter which is in a car <i>caption_1</i> : a person is in a car which is in a helicopter	T_a : [Noun1] [Verb (vi)] [Prepositional Phrase1 (location)] which is in [Prepositional Phrase2 (location)] T_{pv} : [Noun1] [Verb (vi)] [Prepositional Phrase2 (location)] which is in [Prepositional Phrase1 (location)] $T_a \rightarrow T_{pv}$: [Prepositional Phrase1 (location)] \leftrightarrow [Prepositional Phrase2 (location)]
<i>caption_0</i> : there are some pineapples in boxes, and far more pineapples than boxes <i>caption_1</i> : there are some boxes containing pineapples, and far more boxes than pineapples	T_a : ([Prepositional Phrase1 (location)],)(There be)[Noun1] [locate in] [Noun2], and far more [Noun1] than [Noun2] T_{pv} : ([Prepositional Phrase1 (location)],)(There be)[Noun2] [contain] [Noun1], and far more [Noun2] than [Noun1] $T_a \rightarrow T_{pv}$: [Noun1] \leftrightarrow [Noun2]
<i>caption_0</i> : the person sitting down is supporting the person standing up <i>caption_1</i> : the person standing up is supporting the person sitting down	T_a : [Noun1] (which) [Verb1 (vi)] [Verb (vt)] [Noun2] (which) [Verb2 (vi)] T_{pv} : [Noun1] (which) [Verb2 (vi)] [Verb (vt)] [Noun2] (which) [Verb1 (vi)] $T_a \rightarrow T_{pv}$: [Verb1 (vi)] \leftrightarrow [Verb2 (vi)]
<i>caption_0</i> : the person with green legs is running quite slowly and the red legged one runs faster <i>caption_1</i> : the person with green legs is running faster and the red legged one runs quite slowly	T_a : [Noun1] [Prepositional Phrase1/Relative Clause1 (appearance)] [Verb1 (vi)] slowly and [Noun2] [Prepositional Phrase2/Relative Clause2 (appearance)] [Verb2 (vi)] faster T_{pv} : [Noun1] [Prepositional Phrase1/Relative Clause1 (appearance)] [Verb1 (vi)] faster and [Noun2] [Prepositional Phrase2/Relative Clause2 (appearance)] [Verb2 (vi)] slowly $T_a \rightarrow T_{pv}$: slowly \leftrightarrow faster

1000 Table 4: Examples of extracted templates and transformation rules between templates of (T_a, T_{pv}) .
1001
1002

1003 **Template Acquisition** We designate 171 compositional cases in Winoground Thrush et al. (2022),
1004 which are labeled as “no-tag” in subsequent research Diwan et al. (2022), as SEED₀ and SEED₁.
1005 The template of T_{pv} , the permutation with semantic changes from, is extracted from each pair of
1006 seeds by human annotators. Then, we define the rule of T_{pi} , which is the permutation without
1007 semantic changes, as the original template of T_{pi} . An examples is illustrated as follows.
1008

1009

T_a : [Noun1] [Verb] [Noun2] and [Noun1] behind [Noun3]
T_{pv} : [Noun3] [Verb] [Noun2] and [Noun3] behind [Noun1]
T_{pi} : [Noun1] behind [Noun3] and [Noun1] [Verb] [Noun2]
$T_a \rightarrow T_{pv}$: [Noun1] \leftrightarrow [Noun3]
$T_a \rightarrow T_{pi}$: [Noun1] [Verb] [Noun2] \leftrightarrow [Noun1] behind [Noun3]

1014 If there is no coordinating conjunction such as *and* and *while* for the template of T_{pi} , the template
1015 can be set *NULL*. In this case, the permutation T_{pi} will be generated depends on the LLM according
1016 to other solutions.
1017

1018

T_a : [Noun1] [Verb1 (vi)] [Verb (vt)] [Noun2] [Verb2 (vi)]
T_{pv} : [Noun1] [Verb2 (vi)] [Verb (vt)] [Noun2] [Verb1 (vi)]
T_{pi} : NULL
$T_a \rightarrow T_{pv}$: [Verb1 (vi)] \leftrightarrow [Verb2 (vi)]
$T_a \rightarrow T_{pi}$: NULL

1024 **Template-guided Generation for T_a .** The prompt for generating the T_a , which is guided by the
1025 templates and seed pairs, is:
1026

1026 Assuming you are a linguist, you have the ability to create a similar sentence following the structure of given
 1027 sentences.

1028 The given two sentences are $\{\text{SEED}_0\}$ and $\{\text{SEED}_1\}$. The structure of them are both “ $\{\text{Template } T_a\}$ ”. Please
 1029 create a similar “ $\{\text{Template } T_a\}$ ” sentence as “ TEXT0 ”, and diversify your sentence as much as possible by
 1030 using different themes, scenes, objects, predicate, verbs, and modifiers.

1031
 1032 Output a list containing $\{\text{NUM}\}$ json objects that contain the following keys: TEXT0 . Use double quotes
 1033 instead of single quotes for key and value. Now, let’s start. The output json object list:

1034
 1035 **Rule-guided Permutation for T_{pv} .** The prompt for generating T_{pv} based on T_a and its correspond-
 1036 ing rule is:

1037 Assuming you are a linguist, you have the ability to judge the structure of existing sentences and imitate more
 1038 new sentences with similar structure but varied content.

1039
 1040 Step 1: Input some sentences structured by $\{\text{Template } T_a\}$ and $\{\text{Template } T_{pv}\}$. We call each sentence as
 1041 “ TEXT0 ”.

1042 Step 2: For each “ TEXT0 ”, perform the change which is “ $\{\text{RULE of } T_a \rightarrow T_{pv}\}$ ” and keep the other words
 1043 unchanged as “ TEXT1 ”.

1044 For example, $\text{TEXT0}=\{\text{TEXT0}\}$. Only swap/move $\{\text{RULE of } T_a \rightarrow T_{pv}\}$ and keep the other words
 1045 unchanged to generate $\text{TEXT1}=\{\text{TEXT1}\}$.

1046 Output a list containing $\{\text{NUM}\}$ json objects that contain the following keys: TEXT0 , TEXT1 . Use double
 1047 quotes instead of single quotes for key and value. Now, let’s start. The input is: $\{\text{TEXT0}\}$. The output json
 1048 object list:

1049
 1050 **Paraphrasing-guided Permutation for T_{pi} .** The prompt for generating T_{pi} based on T_a and its
 1051 corresponding rule is:

1052 [Instruction]
 1053 Please generate a sentence that has a similar length and meaning in the following six ways:
 1054 1. Change the word order: For example, “a red and yellow dog” can be changed to “a yellow and red dog.” In
 1055 some languages, adjusting the order of words in a sentence can create a new sentence form without changing
 1056 the meaning. For instance, “I like you” can be adjusted to “You are the person I like”.
 1057 2. Passive voice: For example, “a kid is flying a yellow kite” can be changed to “a yellow kite is being flown by
 1058 a kid.”
 1059 3. Change the description: For example, “a boy is playing with a girl” can be changed by paraphrasing and
 1060 altering the sentence structure to “a boy is playing. He is near a girl.”
 1061 4. Use synonyms: Replace words in the sentence with their synonyms. For example, “happy” can be replaced
 1062 with “joyful”.
 1063 5. Use infinitive or gerund forms: For example, “He likes to run” can be changed to “He enjoys running”.
 1064 6. Simplify or expand: You can either simplify the sentence structure or add additional information
 1065 to create a new sentence. For example, “The quick, brown fox jumps over the lazy dog” can be simplified
 1066 to “The fox jumps over the dog”, or expanded to “The fox, which is quick and brown, jumps over the lazy dog”.

1067 Now, please generate a similar sentence for input prompt given at the end. Provide one sentence for each of the
 1068 six methods. If a sentence cannot be generated using a particular method, please output “None”.

1069 Add the results as a list of JSON objects, containing 6 JSON objects. Each object should include the keys:
 1070 number, modification method, and sentence.

1071 [Prompt]
 1072 “ $\{\text{TEXT0}\}$ ”

1073 C.2 DATA ANNOTATION

1074
 1075 **Criteria for Valid Samples.** The primary challenge in annotation lies in defining the criteria for
 1076 what qualifies as “valid”. For T2I synthesis models, we define “valid” input text based on 14 specific
 1077 criteria. First, we illustrate these criteria through examples of T_a and T_{pv} . Second, for T_{pi} , we
 1078 require it to apply one of the six synonymous transformations defined in the prompt for generating
 1079 T_{pi} . From a semantic perspective, T_{pi} must be strictly consistent with T_a , ensuring the consistency
 and accuracy of the entire dataset. We list these criteria and examples in the following.

Type	Valid Criteria	Example	✓/✗
Basic	Complete Expression	T_a : Swinging on the swing and off the metal chains. T_{pv} : Swinging off the swing and on the metal chains. T_{pi} : Swinging off the metal chains and on the swing.	✗ ✗ ✗
	Clear and Concrete Objects	T_a : A brighter sun is shining on a dimmer object. T_{pv} : A dimmer sun is shining on a brighter object. T_{pi} : A dimmer object is shined on by a brighter sun.	✗ ✗ ✗
	Reasonable Semantics	T_a : An engineer builds a bridge. T_{pv} : A bridge builds an engineer. T_{pi} : A bridge is built by an engineer.	✓ ✗ ✓
Visualizable	Visually Depicted Elements	T_a : There are more salads than burgers on the menu. T_{pv} : There are more burgers than salads on the menu. T_{pi} : There are less burgers than salads on the menu.	✗ ✗ ✗
	Static Scene or Multiple Exposure Scene	T_a : The wave is moving faster and the fish is swimming slowly. T_{pv} : The fish is swimming faster and the wave is moving slowly. T_{pi} : The fish is swimming slowly and the wave is moving faster.	✗ ✗ ✗
	Moderate Details	T_a : In the library, there are a stack of books and some more magazines. T_{pv} : In the library, there are a stack of magazine and some more books. T_{pi} : In the library, there are some more magazines and a stack of books.	✗ ✗ ✗
Discriminative	Quantifiable Comparison	T_a : There are more ants than bees in the garden. T_{pv} : There are more bees than ants in the garden. T_{pi} : There are less bees than ants in the garden.	✗ ✗ ✗
	Modification Rules	T_a : A sharp knife is on a dull cutting board. T_{pv} : A dull cutting board is under a sharp knife. T_{pi} : A dull cutting board is under a sharp knife.	✗ ✗ ✗
	Distinct Textual Semantics	T_a : The boat is on the dock and the fisherman is on the pier. T_{pv} : The boat is on the pier and the fisherman is on the dock. T_{pi} : The fisherman is on the pier and the boat is on the dock.	✗ ✗ ✗
Recognizable	Visually Distinguishable	T_a : There's a delicious chocolate cake with a bitter coffee frosting. T_{pv} : There's a bitter chocolate cake with a delicious coffee frosting. T_{pi} : There's a bitter coffee frosting with a delicious chocolate cake.	✗ ✗ ✗
	Item-Specific Scene	T_a : There are more books than shelves in this library. T_{pv} : There are more shelves than books in this library. T_{pi} : There are less shelves than books in this library.	✓ ✗ ✓
	Item-Specific Character	T_a : A photographer wearing a camera strap with his lens in the air and a videographer wearing a tripod. T_{pv} : A photographer wearing a tripod with his lens in the air and a videographer wearing a camera strap. T_{pi} : A videographer wearing a tripod and a photographer wearing a camera strap with his lens in the air.	✗ ✗ ✗
1117	Attire-based Character	T_a : The soldier in the barracks is cleaning equipment and the officer in the office is reviewing reports. T_{pv} : The soldier in the barracks is reviewing reports and the officer in the office is cleaning equipment. T_{pi} : The officer in the office is reviewing reports and the soldier in the barracks is cleaning equipment.	✗ ✗ ✗
	Action-based Character	T_a : The businessman is wearing navy suit and red tie. T_{pv} : The businessman is wearing red suit and navy tie. T_{pi} : The businessman is wearing red tie and navy suit.	✗ ✗ ✗

Table 5: Error Examples of LLM-generated permutation-based sentences (T_a , T_{pv} , T_{pi}) and the criteria they violate.

- Basic
 - Complete Expression: Both sentences should be complete and free from obvious linguistic errors.
 - Clear and Concrete Objects: Both sentences must be clear and unambiguous, contextually or inherently, and specifically describe tangible objects, steering clear of abstract concepts.
 - Meaningful Sentence: Both sentences must maintain logical coherence in their respective contexts. The reasonable definition includes real-world plausibility or scenarios typically seen as implausible in virtual or imaginative settings (like children’s literature, animations, or science fiction), such as flying pigs or dinosaurs piloting planes. For example, “a shorter person can reach a higher shelf while a taller one cannot” is not reasonable in any world.
- Visualizable
 - Visually Depicted Element: Both sentences must convey visual elements, including objects, scenes, actions, and attributes, ensuring that the text prompts are visually depictable and the image content is identifiable during evaluation.

- 1134 – Static Scene or Multiple Exposure Scene: Both sentences should be visually repre-
 1135 sentable through images alone, negating the need for video, audio, or other sensory
 1136 inputs like touch and smell. Temporal aspects, procedures, and comparisons in test
 1137 cases must be conveyable within a single image’s scope.
- 1138 – A Moderate Level of Details: Sentences should maintain a moderate level of detail
 1139 with similar scales for objects and scenes. Excessive or mismatched scales can result
 1140 in sentences that are challenging to depict. For example, comparing the quantity of
 1141 books and magazines “in a library” is less suitable than “on a table”.
- 1142 – Quantifiable Comparison: Comparisons in both sentences should be quantifiable, us-
 1143 ing measures like counts, areas, or volumes. For example, “There are more students in
 1144 the classroom than words on the blackboard” are difficult to compare quantitatively.
- 1145 • Discriminative
- 1146 – Following Permutation Rules: Generated samples T_{pv} must strictly follow the desig-
 1147 nated manual template, including word swapping and moving.
- 1148 – Distinct Textual Semantics: Two sentences must have distinct textual semantics. Oth-
 1149 erwise, the pairs are considered invalid.
- 1150 – Visually Distinguishable: Two sentences should be visually distinct, with clear differ-
 1151 entiation regarding the visual characteristics of the objects or scenes described. Subtle
 1152 differences requiring very close observation are not considered distinct visual differ-
 1153 ences.
- 1154 • Recognizable
- 1155 – Item-Specific Scenes: Scenes in sentences should be identifiable, maintaining key
 1156 elements for recognition. Otherwise, identification may be challenging. For instance,
 1157 a sentence describing a library where “bookshelves outnumber books” might be
 1158 unrecognizable, as we typically expect a library to contain many books.
- 1159 – Item-Specific Characters: When a sentence depicts a character through associations
 1160 with specific items, these items or behaviors should remain consistent for easy iden-
 1161 tification. If not, the character may be hard to recognize. For instance, chefs are
 1162 usually associated with “chef’s attire, cooking utensils, and kitchens”.
- 1163 – Attire-Based Characters: When a sentence presents characters identifiable by their at-
 1164 tire, such as firefighters, police officers, soldiers, doctors,
 1165 and nurses, their clothing should remain consistent for clear recognition. Changes
 1166 in attire could obscure their identities.
- 1167 – Action-Based Characters: When a sentence features characters defined by specific
 1168 actions or interactions, such as bartenders (mixing drinks), businessmen (ne-
 1169 gotiating), journalists (interviewing), divers (deep-sea diving), their typical
 1170 activities should be consistent. Altering distinctive features or placing characters in
 1171 unusual scenarios may obscure their identities.

1172 **Automatic Annotation.** We employ machine-human hybrid verification to filter out invalid samples
 1173 that violate any characteristic. We use LLMs to judge whether each sample violates any of the
 1174 specific criteria, labeling them “yes” or “no” and providing confidence scores. The samples whose
 1175 confidence exceeds a threshold of 0.8 are removed from the dataset. We initially collected 48K
 1176 samples, each including 3 sentences. The automatic filtering helped eliminate over 42% of them,
 1177 resulting in a final corpus of 27K samples.

1178 **Human Annotation.** We use 15 annotators and 3 experienced experts to manually verify the sam-
 1179 ples. All annotators have linguistic knowledge and are provided with detailed annotation guidelines.
 1180 Each sample is independently annotated by two annotators. Then an experienced expert reviews the
 1181 controversial annotations and makes the final decision. After annotation, we randomly sampled 100
 1182 samples from valid samples to assess annotation accuracy. Two experts evaluated that 99% of the
 1183 samples were valid. Finally, we got 11,479 valid, non-duplicated samples.

1184 **Hard Samples Selection.** To effectively evaluate T2I models, it is crucial to select challenging
 1185 samples rather than simple ones. Initially, we generate images using SOTA models like DALL-E3,
 1186 and flagging those with alignment scores below 0.7. Then we aggregate the votes from these models
 1187 to determine the most representative candidates, and select those with the highest votes for further

Category	Train	Test	Total
Relation			
Absolute Location	1,716	50	1,766
Relative Location	1,111	50	1,161
Action	216	48	264
Interaction	153	43	196
Direction	342	33	375
Spatio-temporal	234	50	284
Attribute Comparison			
Vague amount	1,839	50	1,889
Size	2,168	50	3,118
Height	253	50	303
Weight	5	5	10
Attribute Value			
Color	4,451	50	4,501
Appearance	1,972	50	2,022
Texture	542	50	592
Shape	190	50	240
Size	516	50	566
Material	227	50	277
Manner	194	49	243
Sentiment	88	26	114
Age	22	11	33
Temperature	14	4	18
Counting	614	50	664
Total	15,518	819	14,699
Total(deduplication)	11,454	684	10,770

Table 6: Statistics of SemVarBench.

filtering. To ensure diversity, we categorize these samples based on permutation types, as shown in Fig. 5, with a maximum limit of 50 samples per category. Finally, 684 samples were included in our benchmark.

C.3 DATA STATISTICS

Category. The samples in SemVarBench are divided into 20 categories based on their permutation types. Furthermore, these categories are further classified into three aspects based on triple types, as illustrated in Tab. 23. These aspects are *Relation*, *Attribute Comparison* and *Attribute Value*. Specifically, *Relation* aspect includes six categories: *Action*, *Interaction*, *Absolute Location*, *Relative Location*, *Spatial-Temporal*, *Direction*. *Attribute Contrast* includes four categories: *Size*, *Height*, *Weight*, *Vague Amount*. *Attribute Value* includes ten categories: *Color*, *Counting*, *Texture*, *Material*, *Shape*, *Age*, *Sentiment*, *Temperature*, *Manner*, and *Appearance*.

Scale and Split. SemVarBench comprises 11,454 valid samples of (T_a, T_{pv}, T_{pi}) , totaling 34,362 sentences. We divide it into a training set and a test set. The training set contains 10,806 samples, while the test set consists of 648 challenging samples for effective evaluation, as shown in Tab. 6. All our evaluations are conducted on the test set.

Distribution. Since some permutations contain multiple words, they may fall into more than one category. In the training set, 51.06% of the permutation involves only one category, 35.12% for two categories, and 7.35% for three categories and 0.5% for more than four categories. In the test set, 82.75% of permutations involve only one category, 14.77% involve two categories, and 2.49% involve three categories. As a result, the total count of categorized samples exceeds the actual number of unique samples.

SemVarBench vs. Other benchmarks. Compared with existing benchmarks, SemVarBench focuses on the understanding of semantic variations for text-to-image synthesis, which includes two types of permutation: permutation-variance and permutation-invariance. Other comparisons, such as source, scale, annotation and split, are illustrated in Tab. 7.

Benchmark	Concentration	Data Source	#Prompts	Annotation	Split
DrawBench Saharia et al. (2022)	General	Human	200	Human	Test
PartiPromps Yu et al. (2022)	General	Human	1600	Human	Test
PaintSkills Cho et al. (2022)	General	Template	73.3K	–	Train/Test
HRS-Bench Bakr et al. (2023)	General	Template & LLM	45.0K	Human	Test
SR _{2D} Gokhale et al. (2022)	Compositional	Dataset	25.3K	–	Test
ABC-6K Feng et al. (2023)	Compositional	Dataset	6.4K	–	Test
CC-500 Feng et al. (2023)	Compositional	Template	500	–	Test
TIFA v1.0 Hu et al. (2023)	Compositional	Dataset	4.1K	–	Test
VPEval-skill Cho et al. (2023b)	Compositional	Dataset	3.8K	–	Test
DSG-1K Cho et al. (2023a)	Compositional	Dataset	1.1K	–	Test
T2I-CompBench Huang et al. (2023)	Compositional	Template & LLM	6.0K	–	Train/Test
Winoground Thrush et al. (2022)	Permutation-Variance	Human	800	Human	Test
SemVarBench(ours)	Permutation-Variance Permutation-Invariance	Template & LLM	34K	LLM & Human	Train/Test

Table 7: Comparison between SemVarBench and other T2I synthesis benchmarks.

D DETAILS OF EXPERIMENT SETTING

D.1 T2I SYNTHESIS MODELS

We generate one image using the mainstream T2I diffusion models in Fig. 1: Stable Diffusion v1.5² (denoted as SD 1.5), Stable Diffusion v2.1³ (denoted as SD 2.1), Stable Diffusion XL v1.0⁴ (denoted as SD XL 1.0), Stable Cascade⁵ (denoted as SD CA), DeepFloyd IF XL⁶ (denoted as DeepFloyd), PixArt-alpha XL⁷ (denoted as PixArt), Kolors, Stable Diffusion 3 [medium]⁸ (denoted as SD 3), FLUX.1 [dev]⁹ (denoted as FLUX.1), Midjourney V6¹⁰ (denoted as MidJ V6), DALL-E 3¹¹, CogView3-Plus¹² (denoted as CogV3-Plus), Ideogram 2¹³. The schedulers for SD 1.5 and SD 2.1 are set to DPM-Solver++, while all other settings are left as default.

D.2 EVALUATOR

We use four advanced MLLMs as the evaluators to demonstrate the general applicability of our proposed evaluation metrics: Gemini 1.5 Pro, Claude 3.5 Sonnet, GPT-4o and GPT-4 Turbo. GPT-4o and GPT-4 Turbo have been shown to achieve near-human performance in evaluating alignment in T2I synthesis models Zhang et al. (2023); Chen et al. (2024). Claude 3.5 Sonnet outperforms GPT-4o and Gemini 1.5 Pro Anthropic (2024). The versions of these MLLMs used are as follows: Gemini 1.5 Pro (gemini-1.5-pro-001), Claude 3.5 Sonnet (claude-3-5-sonnet-20240620), GPT-4o (gpt-4o-2024-05-13), and GPT-4 Turbo (gpt-4-turbo-2024-04-09). The alignment score components follow the division outlined in Zhang et al. (2023), with the exception of the aesthetic score component, which has been omitted. The complete prompt is as follows.

²The model used is ruwnayml/stable-diffusion-v1-5, which is now deprecated. A mirror is available at: <https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5>

³<https://huggingface.co/stabilityai/stable-diffusion-2-1>

⁴<https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0>; <https://huggingface.co/stabilityai/stable-diffusion-xl-refiner-1.0>

⁵<https://huggingface.co/stabilityai/stable-cascade-prior>; <https://huggingface.co/stabilityai/stable-cascade>

⁶<https://huggingface.co/DeepFloyd/IF-I-XL-v1.0>; <https://huggingface.co/DeepFloyd/IF-II-L-v1.0>; <https://huggingface.co/stabilityai/stable-diffusion-x4-upscaler>

⁷<https://huggingface.co/PixArt-alpha/PixArt-XL-2-1024-MS>

⁸<https://huggingface.co/stabilityai/stable-diffusion-3-medium>

⁹<https://huggingface.co/black-forest-labs/FLUX.1-dev>

¹⁰<https://www.midjourney.com/home>

¹¹<https://openai.com/index/dall-e-3/>

¹²<https://www.bigmodel.cn/dev/api/image-model/cogview>

¹³<https://about.ideogram.ai/2.0>

1296 Does the generated image align with the given prompt?
 1297
 1298 [Instruction] Carefully assess the generated image in terms of relevance to the prompt and object accuracy.
 1299 Notice that the image is digitally created or artificially generated, and I hope you help feedback on the quality
 1300 of a generated image rather than discussing the content of a real photograph.
 1301 Use the following criteria to guide your evaluation: with Relevance (0-50 points), Object Accuracy
 1302 (0-50 points). After providing your explanation, you must rate the generated image by strictly following
 1303 this format: “[rating]”, for example: “Relevance (0-50 points): [[35]], Object Accuracy (0-50 points): [[30]]”.
 1304 [Prompt]
 1305 {prompt}
 1306

1307 After receiving outputs from LLMs, we utilize regular expressions to extract scores. In our experi-
 1308 ments, the outputs from four evaluators mentioned above consistently followed the specified format
 1309 as defined in the prompt. We also tested Qwen-VL-Chat, Qwen-VL-Plus, Qwen-VL-Max, and
 1310 LLAVA-1.6, which exhibited poor adherence to the specified format and required a more complex
 1311 extraction process. To simplify the evaluation process, we decided to adopt results exclusively from
 1312 Gemini 1.5 Pro, Claude 3.5 Sonnet, GPT-4o, and GPT-4-Turbo.
 1313

1314 D.3 TRAINING SETTING

1315
 1316 **Training Data Selection.** The training set of SemVarBench comprises 10,806 samples. We inves-
 1317 tigate the improvement from fine-tuning the T2I model Stable Diffusion XL v1.0. We select the
 1318 generated images whose alignment scores meet the requirements. These constraints are as follows.

1319 First, the generated image should be approximately aligned with its corresponding text prompt.
 1320

$$\begin{cases} S(T_a, I_a) > C_2, \\ S(T_{pv}, I_{pv}) > C_2, \end{cases} \quad (12)$$

1323 where C_2 is a threshold.
 1324

1325 Second, the alignment scores between matched text-image pairs should be higher than those between
 1326 mismatched text-image pairs.

$$\begin{cases} S(T_a, I_a) > S(T_a, I_{pv}), \\ S(T_a, I_a) > S(T_{pv}, I_a), \\ S(T_{pv}, I_{pv}) > S(T_a, I_{pv}), \\ S(T_{pv}, I_{pv}) > S(T_{pv}, I_a), \end{cases} \quad (13)$$

1331 Third, the visual semantic variations observed from different text prompts should be the same when
 1332 the initial image and the final image are the same.
 1333

$$S(T_a, I_a) - S(T_a, I_{pv}) \approx S(T_{pv}, I_{pv}) - S(T_{pv}, I_a), \quad (14)$$

1335 Similarly, the textual semantic variations observed from different images should be the same when
 1336 the initial text prompt and the final text prompt are the same.
 1337

$$S(T_a, I_a) - S(T_{pv}, I_a) \approx S(T_{pv}, I_{pv}) - S(T_a, I_{pv}), \quad (15)$$

1339 Utilizing this approximate equality relationship in Eq. 14 and Eq. 15, we constrain the alignment
 1340 score using the following inequality:

$$\begin{cases} |(S(T_a, I_a) - S(T_a, I_{pv})) - (S(T_{pv}, I_{pv}) - S(T_{pv}, I_a))| < C_3, \\ |(S(T_a, I_a) - S(T_{pv}, I_a)) - (S(T_{pv}, I_{pv}) - S(T_a, I_{pv}))| < C_3, \end{cases} \quad (16)$$

1344 In our experiments, we utilized Stable Diffusion XL v1.0 to generate an image for each text prompt
 1345 within the training set. To select the training data, we designated $C_2 = 0.8$ and $C_3 = 0.1$. Ulti-
 1346 mately, we selected 327 samples, resulting in 981 sentences.
 1347

1348 **Supervised Fine-Tuning (SFT).** Each text-image pair (T_i, I_i) is incorporated into the training set.
 1349 For every sample (T_a, T_{pv}, T_{pi}) , which results in three text-image pairs: (T_a, I_a) , (T_{pv}, I_{pv}) and
 (T_{pi}, I_{pi}) , leading to a total of 981 diverse pairs. The selected set of samples is denoted as D_s . The

loss function for SFT remains unchanged Kingma et al. (2021); Song et al. (2021), which is defined as

$$\mathcal{L}(\theta) = \mathbb{E}_{(x,y) \in \mathcal{D}_s} \left[\|\epsilon - \epsilon_\theta(z_t, t, y)\|_2^2 \right], \quad (17)$$

where x, y, t, z_t are the representations of the image I_i , text prompt T_i , timestamp t , and the latent representation of the image at timestamp t , respectively. We conducted two separate fine-tuning processes using the diffusers library¹⁴: only fine-tuned the LoRA model on either the UNet or on the text encoder for 5000 steps, with a training batch size of 1. Our computational resources included a NVIDIA GeForce RTX 4090 with 25.2 GB of VRAM and a 16-core AMD EPYC 9354 processor, with 60.1 GB of system memory available. We train the LoRA model with a rank of 4 on UNet or text encoders, and the training process takes approximately 0.5 hours.

Direct Policy Optimization (DPO). In our experiments, we added text-image tuples of the form (T_i, I_i, I_j) to the training set, where the semantic content of T_i does not match T_j . For each input T_i , I_i represents the chosen image and I_j the rejected one. For every sample (T_a, T_{pv}, T_{pi}) , this results in four text-image tuples: (T_a, I_a, I_{pv}) , (T_{pv}, I_{pv}, I_a) , (T_{pv}, I_{pv}, I_{pi}) , and (T_{pi}, I_{pi}, I_{pv}) , resulting in a total of 1,308 tuples. The loss function for DPO remains unchanged Wallace et al. (2024), which is defined as

$$\begin{aligned} \mathcal{L}(\theta) = & -\mathbb{E}_{(x^w, x^l, y) \sim \mathcal{D}_s, z_t^w \sim q(z_t^w | x^w), z_t^l \sim q(z_t^l | x^l)} \log \sigma(\\ & -\beta (\|\epsilon^w - \epsilon_\theta(z_t^w, t, y)\|_2^2 - \|\epsilon^w - \epsilon_{ref}(z_t^w, t, y)\|_2^2 - \\ & (\|\epsilon^l - \epsilon_\theta(z_t^l, t, y)\|_2^2 - \|\epsilon^l - \epsilon_{ref}(z_t^l, t, y)\|_2^2))), \end{aligned} \quad (18)$$

where $x^w, x^l, y, t, z_t^w, z_t^l, \sigma$ are the representations of the chosen image I_i , the rejected image I_j , text prompt T_i , timestamp t , the latent representation of the chosen image at timestamp t , the latent representation of the rejected image at timestamp t and the sigmoid function, respectively. We executed two separate fine-tuning processes using the DiffusionDPO¹⁵: only fine-tuned the LoRA model on either the UNet or on the text encoder for 5000 steps, with a training batch size of 1. Our computational resources included an Tesla V100-SXM2 with 32GB of VRAM and a 11-core Intel(R) Xeon(R) Platinum 8163 processor, with 88.0 GB of system memory available. We train the LoRA model with a rank of 4 on UNet or text encoders, and the training process takes approximately 4.5 hours.

E MORE EXPERIMENT RESULTS

E.1 RESULTS ON CATEGORIES

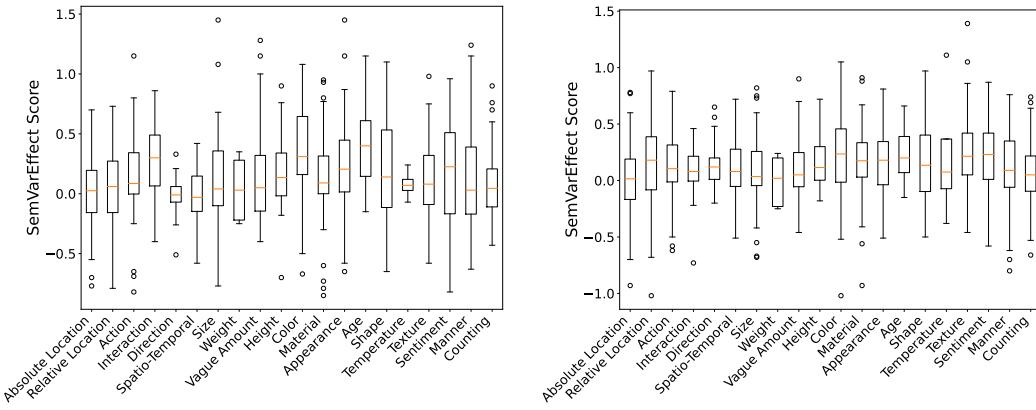


Figure 11: The distribution of SemVarEffect scores across various categories for the Ideogram 2 model and DALL E-3 model, as evaluated by GPT-4 Turbo. Left: Ideogram 2. Right: DALL E-3.

¹⁴The diffusers library support fine-tuning Unet and Unet + text encoder. We made minor modifications to support fine-tuning only the text encoder. The url of scripts provided by diffusers is: https://github.com/huggingface/diffusers/tree/main/examples/text_to_image/

¹⁵We used the code provided by the diffusers library, which supports fine-tuning of Unet. We made minor modifications to support fine-tuning only the text encoder. The url of scripts provided by diffusers is: https://github.com/huggingface/diffusers/tree/main/examples/research_projects/diffusion_dpo/

1404	1405	1406	Models	Relation					Attribute Comparison			
				Absolute Location	Relative Location	Action	Interaction	Direction	Spatial -Temporal	Size	Weight	Vague Amount
Open Source Models												
1407	Stable Diffusion v1.5	-0.01	0.00	-0.06	-0.11	<u>0.05</u>	-0.01	0.11	-0.01	-0.07	0.01	
1408	Stable Diffusion v2.1	-0.01	-0.06	-0.08	0.02	<u>-0.02</u>	-0.00	0.03	-0.10	0.02	0.06	
1409	Stable Diffusion XL v1.0	-0.02	-0.08	-0.01	0.05	<u>0.05</u>	-0.05	0.07	<u>0.16</u>	0.03	0.09	
1410	Stable Cascade	0.02	-0.03	0.01	-0.03	0.02	0.02	-0.02	-0.09	0.08	-0.01	
1411	DeepFloyd IF XL	-0.01	-0.00	0.01	-0.01	0.04	0.01	0.03	0.05	-0.04	0.03	
1412	PixArt-alpha XL	0.00	-0.01	0.03	0.00	<u>-0.04</u>	-0.03	0.07	0.10	0.10	0.03	
1413	Kolors	-0.03	0.02	-0.07	0.02	0.03	-0.02	-0.06	-0.10	0.07	0.07	
1414	Stable Diffusion 3	-0.03	0.01	-0.02	0.05	<u>-0.08</u>	-0.04	0.07	-0.02	0.10	0.04	
1415	FIUX.1	-0.03	0.03	0.03	<u>0.08</u>	-0.04	-0.00	0.09	0.23	0.05	0.09	
API-based Models												
1416	Midjourney V6	0.07	0.01	0.04	0.03	0.03	<u>0.08</u>	0.07	-0.12	0.07	0.02	
1417	DALL-E 3	-0.00	0.12	0.11	<u>0.08</u>	0.13	0.11	0.08	-0.00	0.09	0.15	
1418	CogView3-Plus	0.08	<u>0.08</u>	0.23	0.07	0.03	-0.01	0.23	-0.03	0.23	0.22	
1419	Ideogram 2	0.01	0.04	<u>0.13</u>	0.29	-0.02	-0.02	<u>0.12</u>	0.04	<u>0.17</u>	0.17	

Table 8: The results of SemVarEffect κ on aspects *Relation* and *Attribute Comparison*. The evaluator is GPT-4 Turbo.

1420	1421	1422	Models	Attribute Value									AVG
				Color	Material	Appearance	Age	Shape	Temperature	Texture	Sentiment	Manner	
Open Source Models													
1423	Stable Diffusion v1.5	0.09	0.02	0.04	-0.20	0.01	0.06	0.02	-0.06	-0.05	-0.07	-0.01	
1424	Stable Diffusion v2.1	0.12	0.10	-0.03	-0.09	-0.00	-0.06	-0.03	-0.10	-0.01	0.02	0.00	
1425	Stable Diffusion XL v1.0	0.13	0.09	-0.01	0.01	-0.01	0.05	-0.00	0.03	-0.00	0.00	0.02	
1426	Stable Cascade	0.14	0.05	0.10	-0.03	-0.02	-0.15	-0.03	-0.05	0.06	0.01	0.04	
1427	DeepFloyd IF XL	0.19	0.14	0.06	-0.19	0.04	-0.02	0.05	-0.05	0.06	0.01	0.04	
1428	PixArt-alpha XL	0.11	0.09	0.00	0.15	0.01	0.13	-0.02	0.02	0.03	-0.00	0.02	
1429	Kolors	0.21	0.07	-0.01	0.01	-0.09	0.01	0.10	-0.01	-0.04	0.00	0.01	
1430	Stable Diffusion 3	0.33	0.10	0.11	0.04	0.08	0.11	0.08	0.01	0.03	0.06	0.06	
1431	FIUX.1	<u>0.35</u>	0.21	<u>0.21</u>	0.08	0.09	-0.13	0.04	-0.06	0.07	0.10	0.09	
API-based Models													
1432	Midjourney V6	0.20	0.12	0.13	0.10	-0.03	-0.21	0.11	-0.05	0.09	-0.02	0.05	
1433	DALL-E 3	0.22	<u>0.17</u>	0.17	0.23	0.14	<u>0.22</u>	0.22	<u>0.19</u>	0.11	0.06	0.12	
1434	CogView3-Plus	<u>0.35</u>	0.15	0.17	<u>0.31</u>	<u>0.16</u>	0.30	<u>0.17</u>	0.21	0.27	0.07	0.15	
1435	Ideogram 2	<u>0.37</u>	0.15	0.24	0.42	0.20	0.08	0.12	0.16	<u>0.13</u>	<u>0.07</u>	0.13	

Table 9: The results of SemVarEffect κ on aspects *Attribute Value*. The evaluator is GPT-4 Turbo. AVG represents the average effect score of all samples on aspect *Relation*, *Attribute Comparison* and *Attribute Value*. The evaluator is GPT-4 Turbo.

Effects of Semantic Variations on Different Categories. The impact of semantic variations is not uniform across different semantic classes, as shown in Fig. 7, with exact scores listed in Tabs. 8 and Tab. 9. For *Relation*, most models consistently show low scores, as indicated by the dark blue shading in Fig. 7. This suggests that models handle samples involving *Relations*—such as *Absolute Location*, *Relative Location*, and *Actions*—with limited accuracy. For *Attribute Value*, models such as Ideogram2 perform significantly better at capturing attributes such as *Color*, as shown by the prominent red shading in Fig. 7. These models demonstrate a clear advantage in both generating and recognizing these attributes. In contrast, models such as DALL-E 3 and CogV3-Plus display a more balanced yet average performance across most categories (shaded in light orange and light blue). For *Attribute Comparison* (e.g., *Size*, *Weight*, *Height*), most models score lower, indicating their weaker ability to handle complex attribute comparisons.

Although most T2I models struggle with capturing semantic variations in many categories, some categories, such as *Color* and *Age*, demonstrate slightly better performance, as indicated by higher median values. Fig. 11 illustrates the distribution of SemVarEffect scores across various categories for the Ideogram 2 model, while Fig. 12 shows the scores for different T2I models in the *Color* and *Direction* categories. Most categories have medians (marked by the orange line) close to zero, indicating that T2I models generally struggle to capture the semantic variations introduced by word order changes, particularly in the *Direction* category. However, some categories, such as *Weight* and *Color*, show slightly higher median values, indicating that semantic variation caused by word order changes may have a minor positive effect in these instances. Categories such as *Absolute Location* and *Counting* show greater variability in model responses, while categories such as *Sentiment* and *Texture* show more consistent effects with narrower distributions.

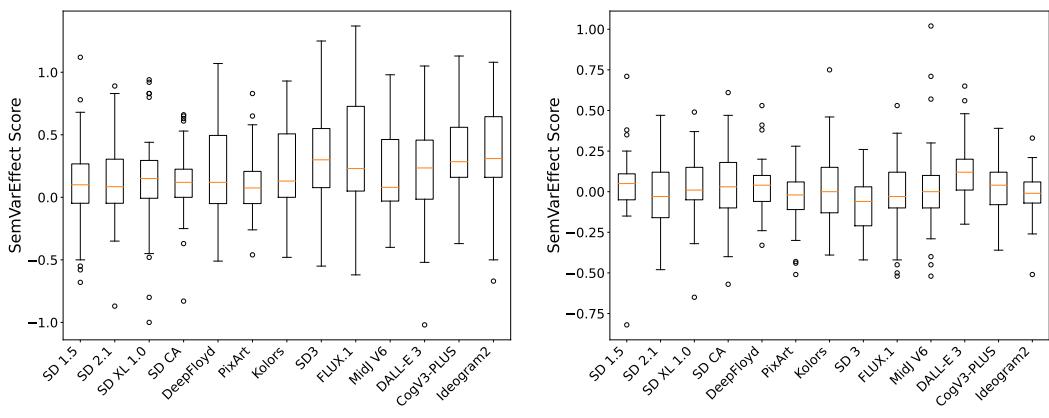


Figure 12: The distribution of SemVarEffect scores across various T2I models within the Color and Direction categories, as evaluated by GPT-4 Turbo. Top: Color. Bottom: Direction.

E.2 RESULTS ON HUMAN EVALUATION

Human Evaluation. To validate the effectiveness of MLLMs we used, we add human evaluation with three raters on 80 samples (20 each from SOTA models: Midjourney v6, DALL-E 3, CogView3-Plus, and Ideogram2) through stratified sampling (one for each category). Following the same scoring protocol as our automatic evaluation, each rater scores the semantic alignment of matched or mismatched image-text pair, and we used their mean scores to calculate SemVarEffect. The results demonstrate the reliability of our MLLM-based evaluation approach. First, we observe consistent performance trends between human raters and the four MLLMs across all evaluated models (Table 10). Second, our correlation analysis on CogView3-Plus reveals moderate Pearson’s ρ , Spearman’s ϕ , and Cohen’s Kappa κ_{cohen} coefficients between machine and human scores (Table 11), suggesting our selected MLLMs can serve as a reliable proxy for human evaluation. This validates our MLLM-based evaluation while confirming T2I models’ current limitations.

Models	$\bar{S}(\uparrow)$	$\gamma_w(\uparrow)$	$\gamma_{wo}(\downarrow)$	$\kappa(\uparrow)$
Midjourney V6	0.59	0.37	0.49	-0.12
DALL-E 3	0.63	0.53	0.50	0.03
CogView3-Plus	0.69	0.52	0.33	0.19
Ideogram 2	0.74	0.50	0.31	0.19

Table 10: Evaluation human rating results of different T2I models in understanding semantic variations.

Models	$\rho(\uparrow)$	$\phi(\uparrow)$	$\kappa_{cohen}(\uparrow)$
GPT-4o	0.54	0.53	0.53
GPT-4v	0.50	0.51	0.54
Claude-3.5-Sonnet	0.42	0.37	0.37
Gemini-Pro-1.5	0.27	0.11	0.23

Table 11: Correlation coefficients between GPT-4o/v, Claude-3.5-Sonnet, Gemini-Pro-1.5 and human evaluations on SemVarEffect of CogView3-Plus.

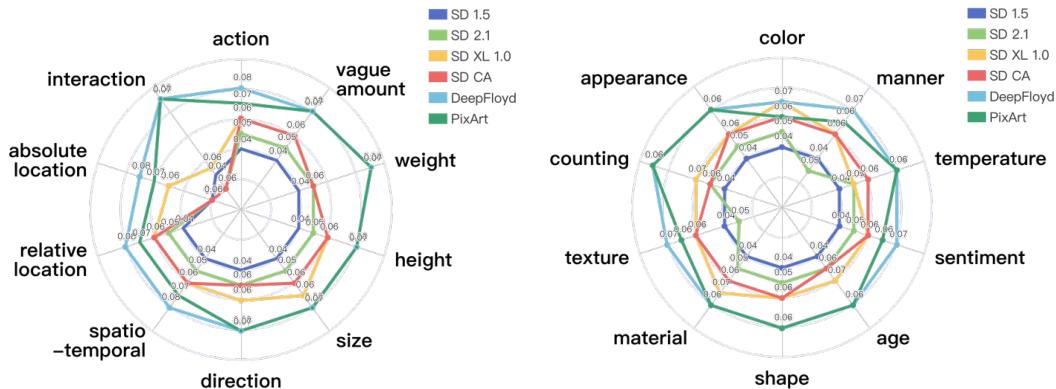
F MORE ANALYSIS

F.1 TEXT ENCODER

Do different text encoders themselves in the text space distinguish semantic variations caused by linguistic permutations? We explore the efficacy of diverse text encoders in discerning such nuances. Fig. 13 compares the text similarity between T_a and T_{pv} across models utilizing different text encoder models. SD 1.5, SD 2.1, SD XL v1.0, and SC utilize CLIP series models as text encoders, while DeepFloyd, PixArt, and DALL-E 3 utilize T5 series models. The similarity metric is depicted as $1 - \text{cosine}(T_a, T_{pv})$, with higher values indicating a stronger ability of the text encoder to differentiate between the semantics of two sentences. This indicates that the choice of text encoder significantly influences the model’s semantic discrimination capabilities.

Category	(T_a, T_{pv})	(T_a, T_{pi})	(T_a, T_{random})
Relation			
Absolute Location	11.94	27.24	52.40
Relative Location	12.26	28.62	46.98
Action	13.94	31.85	48.35
Interaction	13.56	29.58	44.26
Direction	12.03	27.18	50.21
Spatio-temporal	17.40	42.74	59.94
Attribute Comparison			
Vague amount	19.38	36.56	50.38
Size	11.38	26.00	46.82
Height	13.04	23.00	31.22
Weight	10.00	26.20	22.20
Attribute Value			
Color	11.80	30.86	46.90
Material	12.40	28.42	41.92
Appearance	13.86	44.14	59.14
Age	14.73	34.73	46.64
Shape	13.34	33.48	40.98
Temperature	11.00	27.50	38.50
Texture	11.74	31.20	54.22
Sentiment	11.96	33.15	48.65
Manner	13.37	33.71	52.90
Counting	8.44	29.06	44.18
Average	13.12	31.65	54.30

Table 12: The average edit distance between sentences in different categories.

Figure 13: The semantic discrimination capabilities of different text encoders measured by $1 - \cosine(T_a, T_{pv})$.

Why do permutations without semantic changes exhibit higher text similarity scores compared to those with semantic changes? This phenomenon is closely related to our dataset’s construction methodology, where T_{pi} is generated by swapping two long phrases located on either side of a coordinating conjunction or a predicate, such as the *and* in Fig. 4. We observed that permutations with semantic changes in our benchmark have significantly smaller edit distances from the anchor sentence compared to synonymous sentences, as shown in Tab. 12. The average edit distances between (T_a, T_{pv}) , (T_a, T_{pi}) and (T_a, T_{random}) are 13, 32 and 53. As our analysis does not rely on the similarity scores of synonymous sentences, this does not affect our previous findings.

F.2 EVALUATION METRICS

F.2.1 ALIAGNMENT SCORE & SEMVAREFFECT SCORE

Detailed Analysis on Alignment Scores vs. SemVarEffect Score Fig. 14 illustrates that although the distributions of the SemVarEffect score and the alignment score are similar, the SemVarEffect score demonstrates a higher degree of differentiation, especially when it comes to distinguishing between FLUX.1 and SD 3. Based on the alignment score, it could be concluded that FLUX.1, SD 3, and SD XL 1.0 have comparable performance levels and they may be grouped into the same

cluster. However, based on the SemVarEffect score, it becomes evident that FLUX.1 and SD 3 differ distinctly from SD XL 1.0. SD XL 1.0 responds more similarly to semantic variations caused by word order changes in a manner similar to SD 1.5, SD 2.1, and SD CA. Correspondingly, we observe that when using the T5-XXL series model as the text encoder, the difference between DALL-E 3 and other models, such as PixArt and DeepFloyd, becomes more pronounced when assessed by the SemVarEffect score.

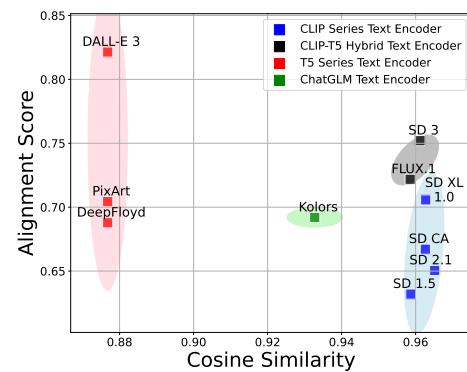


Figure 14: A comparison of alignment scores and the SemVarEffect score under the same conditions of text similarity. The squares are results of permutations of permutation-variance. The evaluator is GPT-4 Turbo.

It should be noted that SemVarEffect is not intended to challenge alignment evaluation, but to serve as a complementary metric to alignment score \bar{S} . This metric focuses only on the consistency of variations rather than the absolute quality of generations. Thus, a model would still score well on SemVarEffect even if it consistently generates incorrect attributes (such as believing oranges are blue and bananas are green), as long as it maintains consistent semantic variations. SemVarEffect helps validate the reliability of high alignment scores \bar{S} : high \bar{S} with low κ indicates limited semantic understanding, while high \bar{S} with high κ suggests effective semantic understanding.

F.3 TRAINING DATA

F.3.1 DATA FILTER CHOICES

How does the filtering standards of training data using MMLM potentially affect the evaluation results, particularly for categories with limited samples? We examine filtering standards across data-rich and data-limited categories to study the effects of data scale and data quality. The filtering criteria contains:

- Alignment Score Thresholds: Values of 0.8 and 0.6 for C_2 (defined in Eq. 12)
- Existence of Strict Filtering on Semantic Variations: Clear distinction between matched and unmatched text-image pairs through strict filtering criteria (defined in Eq. 13-16).

– High filtering. High filtering standards can reduce available data in categories with limited data, leading to decreased alignment scores and SemVarEffect scores. Given the same initial data size, Absolute_location has lower alignment scores (> 0.6) than Height/Direction (> 0.7) due to stricter filtering effects, as shown in Table 3. However, due to its large initial data pool (1.7k candidates), Absolute_location retains sufficient high-quality samples for effective fine-tuning under the “0.8+strict” setting, despite a low retention rate, as shown in Table 13. This comparison between categories with different data scales demonstrates that initial data volume is crucial for achieving improvements.

– Relaxing filtering. Results show that relaxing filtering criteria for more training data leads to worse performance compared to the strictest standards. Removing “0.8+strict filtering” leads to decreased performance across all categories, even those with sufficient data like Color and Absolute_location, as shown in Table 13. For categories with limited data, “0.8+strict filtering” minimizes the decrease in both alignment scores and semantic variation effects., as shown in Table 14. This demonstrates that even with limited data, lowering filtering standards is not a viable solution.

F.3.2 COMMONSENSE BIASES

Is the model’s insensitivity to word order variations caused by commonsense biases in training data? To investigate this question, we conducted controlled experiments with balanced training

Filter Constrains	$\bar{S}(\uparrow)$	$\gamma_w(\uparrow)$	$\gamma_{wo}(\downarrow)$	$\kappa(\uparrow)$
zeroshot	0.73	0.33	0.25	0.08
0.8 + strict filtering	0.78	0.44	0.24	0.20
0.8 + no filtering	0.70	0.42	0.29	0.14
0.6 + strict filtering	0.73	0.44	0.23	0.22
0.6 + no filtering	0.72	0.44	0.27	0.19

Filter Constrains	$\bar{S}(\uparrow)$	$\gamma_w(\uparrow)$	$\gamma_{wo}(\downarrow)$	$\kappa(\uparrow)$
zeroshot	0.64	0.29	0.34	-0.05
0.8 + strict filtering	0.65	0.42	0.35	0.07
0.8 + no filtering	0.54	0.25	0.31	-0.05
0.6 + strict filtering	0.56	0.36	0.38	-0.02
0.8 + no filtering	0.57	0.31	0.40	-0.09

Table 13: Fine-tuning Results on Color (left) and Absolute Location (right) with different data filtering standard.

Filter Constrains	$\bar{S}(\uparrow)$	$\gamma_w(\uparrow)$	$\gamma_{wo}(\downarrow)$	$\kappa(\uparrow)$
zeroshot	0.77	0.34	0.23	0.10
0.8 + strict filtering	0.77	0.27	0.09	0.07
0.8 + no filtering	0.74	0.32	0.28	0.04
0.6 + strict filtering	0.72	0.29	0.38	0.05
0.6 + no filtering	0.73	0.33	0.29	0.04

Filter Constrains	$\bar{S}(\uparrow)$	$\gamma_w(\uparrow)$	$\gamma_{wo}(\downarrow)$	$\kappa(\uparrow)$
zeroshot	0.79	0.20	0.15	0.05
0.8 + strict filtering	0.77	0.25	0.24	0.02
0.8 + no filtering	0.75	0.24	0.24	0.00
0.6 + strict filtering	0.75	0.19	0.28	-0.09
0.8 + no filtering	0.74	0.20	0.27	-0.07

Table 14: Fine-tuning Results on Height (left) and Direction (right) with different data filtering standard.

data with commonsense and anti-commonsense and evaluated the model’s generalization to novel object pairs. For training, we created a balanced dataset consisting of 40 plausible images for “cat chasing mouse” (human-selected from DALL-E 3 generations) and 40 anti-commonsense images for “mouse chasing cat”. The anti-commonsense images were manually created, where 31 images were generated via DALL-E 3 with professional designer editing, and 9 images were created via vector graphics compositing, as shown in Figure 15. To evaluate the model’s performance, we categorized generation errors into five types:

- Missing Objects: generated images lacking one or more required objects.
- No Interaction: all objects present but without any interaction.
- Wrong Interaction: objects interacting but not performing the required chasing action.
- Wrong Direction: objects running but not in a chasing formation (e.g., running in opposite directions).
- Reversed Roles: correct chasing action but with reversed subject-object roles (e.g., mouse chasing cat when cat chasing mouse was required).

Our training performance results showed that while generation quality improved (Right cases increased from 0-2 to 10-14), Reversed Role errors also increased, suggesting the model learned to generate chase scenes but struggled with directional semantics, as shown in Table 15. When testing on novel object pairs with the same “chasing” relationship (both plausible pairs like “hippo↔elephant” and “bull↔man”), we observed consistently poor performance across all new pairs, with similar increasing trends of No Interaction errors and Reversed Roles, as shown in Table 16 and Table 17. Notably, there was no significant difference between plausible and anti-commonsense scenarios. These findings suggest the core challenge isn’t commonsense bias, but rather a fundamental limitation in processing directional relationships. The model struggles equally

Class	Reasons	SD XL		FT SD XL (trained on mouse↔cat)	
		mouse→cat	cat→mouse	mouse→cat	cat→mouse
Wrong	Missing Objects	12	14	4	2
	No Interaction	3	1	7	7
	Wrong Interaction	4	5	3	2
	Wrong Direction	7	8	0	5
	Reversed Role	2	0	6	0
Right	Partial/Full Match	0	2	10	14

Table 15: Training performance of SD XL before and after fine-tuning on balanced mouse↔cat data.

	Class	Reasons	SD XL bull→man	SD XL man→bull	FT (trained on mouse↔cat) bull→man	FT (trained on mouse↔cat) man→bull
Wrong	Missing Objects		0	0	0	0
	No Interaction		6	12	14	15
	Wrong Interaction		5	2	4	3
	Wrong Direction		16	12	8	3
	Reversed Role		2	2	2	6
Right	Partial/Full Match		1	2	2	3

Table 16: Testing performance on novel bull↔man pairs after fine-tuning SD XL on balanced mouse↔cat data.

	Class	Reasons	SD XL hippo→elephant	SD XL elephant→hippo	FT (trained on mouse↔cat) hippo→elephant	FT (trained on mouse↔cat) elephant→hippo
Wrong	Missing Objects		11	16	4	9
	No Interaction		11	10	16	14
	Wrong Interaction		7	3	4	2
	Wrong Direction		0	1	0	0
	Reversed Role		1	0	4	1
Right	Partial/Full Match		0	0	2	4

Table 17: Testing performance on novel hippo↔elephant pairs after fine-tuning SD XL on balanced mouse↔cat data.

with both plausible and anti-commonsense scenarios, indicating an inability to establish proper subject-object relationships regardless of semantic plausibility.

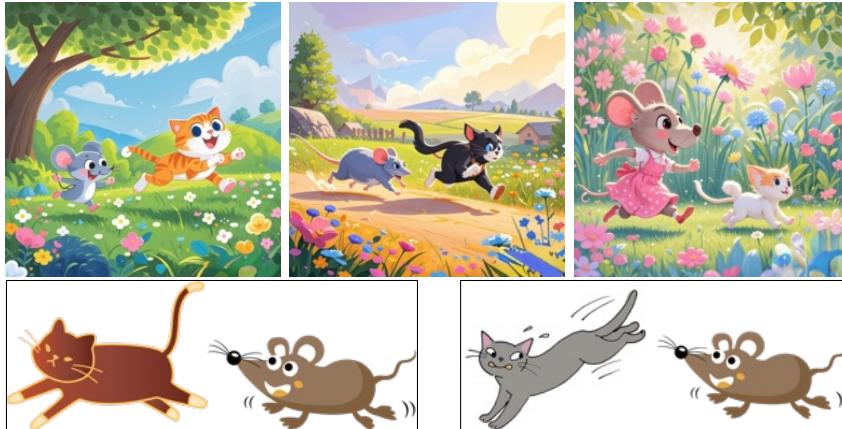


Figure 15: Examples of anti-commonsense scenario “a mouse chasing a cat”. Top: DALL-E 3 generated images with professional designer editing in Photoshop. Bottom row: Vector graphics compositions.

F.3.3 PLAUSIBLE SCENARIOS

Can current T2I models effectively distinguish different semantic relations in plausible scenarios? We investigate advanced commercial T2I models’ ability to handle semantic variations through two plausible scenarios: “A cat chasing a dog” and “A dog chasing a cat”. Both scenarios are possible in real life. Using DALL-E 3 and CogView3-Plus, we generated 30 images for each prompt and evaluated them based on strict criteria: (a) both animals, (b) running, (c) in the same direction, (d) clear spatial relationships (chaser behind chased), and (e) proper chase interaction. Results shows significant performance differences, as shown in Table 18. For “A dog chasing a cat”, DALL-E 3 achieving 70% accuracy (21/30). However, for “A cat chasing a dog”, performance dropped dramatically. DALL-E 3 achieved only 13.3% accuracy while CogView3-Plus failed completely. Failed cases either No Interaction or Reversed Role errors. Even though these scenarios are more plausible

	Class	Reasons	DALL-E 3		Cogview3-Plus	
			cat→dog	dog→cat	cat→dog	dog→cat
Wrong	Miss Objects	0	0	0	0	0
	No Interaction	17	5	12	22	
	Wrong Interaction	0	0	0	0	
	Wrong Direction	0	3	0	0	
	Reversed Role	9	1	18	0	
Right	Partial/Full Match	4	21	0	8	

Table 18: Error analysis of advanced commercial text-to-image models (DALL-E 3 and Cogview3-Plus) on directional relationship generation.

	Class	Reasons	SD XL		FT SD XL (trained on cat↔dog)	
			mouse→cat	cat→mouse	mouse→cat	cat→mouse
Wrong	Missing Objects	12	14	12	21	
	No Interaction	4	5	2	1	
	Wrong Interaction	3	1	4	2	
	Wrong Direction	7	8	4	1	
	Reversed Role	2	0	5	2	
Right	Partial/Full Match	0	2	3	3	

Table 19: Testing performance on novel mouse↔cat pairs after fine-tuning SD XL on balanced cat↔dog data.

than "a mouse chasing a cat", particularly "a cat chasing a dog", current advanced T2I models still struggle with semantic role reversals.

F.3.4 DATA IMBALANCE

Do T2I models' struggles with semantic relationships stem from training data imbalance? We conducted experiments testing performance of fine-tuned SDXL trained with balanced training data. We used a human-filtered balanced dataset of cat↔dog chasing interactions for training, where 80 images for "a dog chasing a cat" and 80 images for "a cat chasing a dog" selected from DALL-E 3 generations. We used two unseen prompt pairs involving the same "chasing" relationship between common objects for testing. The experiment results are shown in Table 19 and Table 20. Despite balanced training, the model showed consistently poor generalization: accuracy remained low for both anti-commonsense scenarios (mouse↔cat, 3-3/30) and plausible scenarios (bull↔man, 4-5/30). Failure analysis revealed three main categories: (1) Missing Objects, where the model fails to generate all required objects; (2) Relationship Understanding Failures, where objects appear either without interaction or with incorrect interactions (e.g. No Interaction, Wrong Interaction and Wrong Direction), indicating the model's inability to comprehend the "chasing" concept; and (3) Reversed Roles, where the model fails to properly assign who chases whom. Even among generated images, correct relationships occurred less frequently than these failure cases, suggesting random performance rather than true understanding. Thus, the models' struggles persist even with perfectly balanced training data, suggesting the core issue lies in relationship understanding rather than data imbalance.

	Class	Reasons	SD XL		FT SD XL (trained on cat↔dog)	
			bull→man	man→bull	bull→man	man→bull
Wrong	Missing Objects	0	0	0	0	
	No Interaction	6	12	14	15	
	Wrong Interaction	5	2	4	3	
	Wrong Direction	16	12	8	3	
	Reversed Role	2	2	2	6	
Right	Partial/Full Match	1	2	2	3	

Table 20: Testing performance on novel bull↔man pairs after fine-tuning SD XL on balanced cat↔dog data.

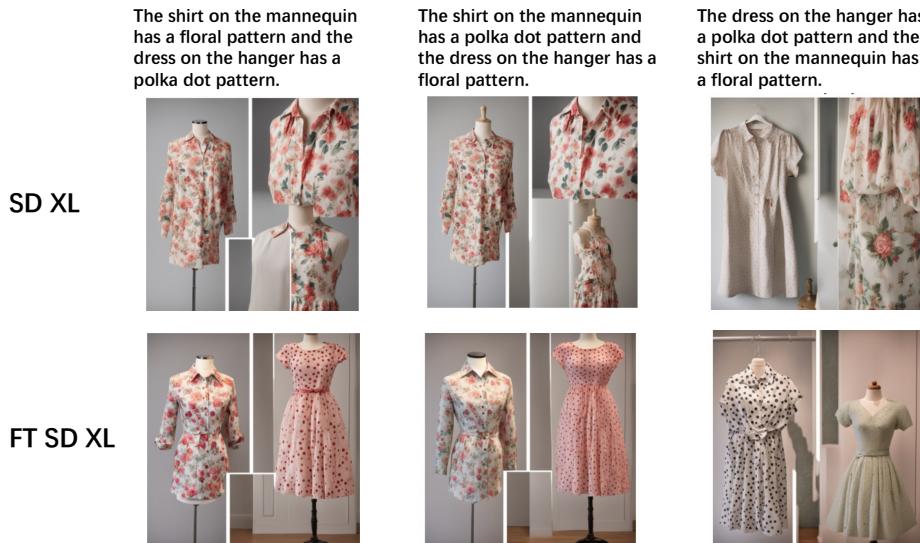
1782 F.4 TRAINING MECHANISM
 1783
 1784 F.4.1 TOKEN-LEVEL IMPROVE OR SEMANTIC-LEVEL IMPROVE

1785 **Does fine-tuning enhance token-level or semantic-level understanding?** The semantic per-
 1786 formance of fine-tuned SD XL are shown in Table 3. To distinguish between improvements in token-
 1787 level or semantic level, we additional conducted experiments at token level. We instructed GPT-4
 1788 to verify whether “words with specific meaning” (including nouns, verbs, adjectives, adverbs, and
 1789 relationship-describing prepositions) from the input text are reflected in the generated image. For ex-
 1790 ample, in “a dog chasing a cat”, GPT-4 would identify three content words (“dog”, “chasing”, “cat”) and
 1791 verify their presence independently. Our analysis of Absolute_location and Height categories
 1792 revealed improved token-level accuracy after fine-tuning, as shown in Table 21. While fine-tuning
 1793 improves token accuracy for the Height category, it actually leads to degradation in semantic un-
 1794 derstanding. This pattern suggests that the model, while better at incorporating individual tokens
 1795 from the prompt after fine-tuning, fails to maintain or improve its understanding of the semantic
 1796 relationships between these elements.

1797 This phenomenon is further illustrated in Figure 16, where the fine-tuned model successfully in-
 1798 cludes most prompted elements (e.g., “shirt”, “mannequin”, “dress”, “hanger”, “floral” and “polka
 1799 dot”) but fails to establish correct semantic relationships between them. For instance, while both
 1800 clothing items and patterns appear in the generated images, their associations are incorrect, demon-
 1801 strating enhanced token-level accuracy but persistent semantic relationship errors. These findings
 1802 indicate that current fine-tuning approaches may prioritize token-level matching over semantic com-
 1803 prehension, suggesting a need for training strategies that better preserve and enhance semantic un-
 1804 derstanding. More examples are illustrated in Figure 17.

Model	Semantic Token Ratio			Model	Semantic Token Ratio		
	T_a	T_{pv}	T_{pi}		T_a	T_{pv}	T_{pi}
SDXL	0.709	0.640	0.716	SDXL	0.881	0.660	0.861
FT SDXL	0.718	0.654	0.716	FT SDXL	0.886	0.662	0.866

1805 Table 21: Semantic Token Ratio comparing base model and fine-tuned model on Absolute_Location
 1806 (left) and Height (right). We first filter out semantic tokens and verify their visual representation in
 1807 the image. The filter and the verifier are both GPT-4o.



1831 Figure 16: Qualitative comparison showing the disconnect between token presence and semantic un-
 1832 derstanding after fine-tuning: while fine-tuning improves token presence (e.g., “shirt”, “mannequin”,
 1833 “dress”, “hanger”, “floral” and “polka dot” all appear in the image), the model fails to capture cor-
 1834 rect semantic relationships between these tokens. This illustrates enhanced token-level accuracy but
 1835 persistent semantic relationship errors.

1836	Model	Testset	$\bar{S}(\uparrow)$	$S_{a,a}(\uparrow)$	$S_{pv,pv}(\uparrow)$	$S_{pi,pi}(\uparrow)$	$(S_{a,a} + S_{pv,pv})/2(\uparrow)$	$(S_{a,a} + S_{pi,pi})/2(\uparrow)$	$\gamma_w(\uparrow)$	$\gamma_{wo}(\downarrow)$	$\kappa(\uparrow)$
1837	SD XL	-	0.786	0.768	0.779	0.809	0.774	0.788	0.201	0.155	0.046
1838	FT SD XL	Imageset 2	0.784	0.802(\uparrow)	0.758	0.793	0.780(\uparrow)	0.797(\uparrow)	0.244(\uparrow)	0.205	0.038
1839		Imageset 3	0.790(\uparrow)	0.806(\uparrow)	0.754	0.809	0.780(\uparrow)	0.807(\uparrow)	0.209(\uparrow)	0.262	-0.053
1840		Imageset 6	0.787(\uparrow)	0.800(\uparrow)	0.768	0.794	0.784(\uparrow)	0.797(\uparrow)	0.244(\uparrow)	0.224	0.02
1841		Imageset 8	0.798(\uparrow)	0.841(\uparrow)	0.756	0.797	0.799(\uparrow)	0.819(\uparrow)	0.230(\uparrow)	0.269	-0.04

Table 22: Analysis of alignment metrics revealing the shortcut phenomenon in fine-tuning on Direction category. While anchor scores ($S_{a,a}$) show consistent improvements (\uparrow) across different Imagesets, permutation scores ($S_{pv,pv}$, $S_{pi,pi}$) remain unchanged or decrease after fine-tuning SD XL (FT SD XL).

F.4.2 SHORTCUT PHENOMENON IN DATA-LIMITED CATEGORIES DURING FINE-TUNING

We observed a “shortcut” phenomenon during fine-tuning models, that is, some improvement of average alignment score is misleading. For data-limited categories like Direction, we generated 8 images per prompt and constructed multiple test samples using various image combinations. The results reveals inconsistent patterns, where average alignment scores \bar{S} (averaged across T_a , T_{pv} , and T_{pi}) increased in 4 sets (Imageset 2,3,6,8) but decreased in 4 sets (Imageset 1,4,5,7) of experiments, as shown in Table 22. For all improved results, they consistently exhibited substantial gains in alignment score of the original text T_a and decreases in alignment score of the permutations T_{pv} and T_{pi} . The magnitude of anchor improvements was so large that it artificially inflated all other average alignment metrics, masking performance issues in other aspects.

This finding raises concerns about evaluation methods in current literature, which often evaluate generation using only the average of alignment scores. Improvements in alignment scores may be misleading. For instance, the fine-tuned model shows higher average scores (0.78 vs. 0.77 on Imageset 2, with Imageset 3,6,8 showing similar improvements) compared to zeroshot. However, our deeper analysis using SemVarEffect κ scores revealed consistent declines, primarily due to γ_{wo} not decreasing as expected, indicating the model fails to understand true semantics.

G MORE CASE STUDIES

In this section, we present examples that demonstrate an understanding of semantic variations and examples that do not. Examples that grasp semantic variations typically have high alignment scores (\bar{S}_{ii}) and high effect scores (κ), as illustrated in Fig. 18. Conversely, examples that lack this understanding often have high alignment scores (\bar{S}_{ii}) but low effect scores (κ), as depicted in Figures 19 and 20. The SemVarEffect scores allow us to distinguish models’ abilities to accurately interpret and visually represent semantic variations. However, in practice, evaluation accuracy can be significantly affected by errors in generated images or biases in evaluators’ ratings. Severe errors can particularly distort the evaluation’s accuracy, as evidenced in Figures 24 and 26. To enhance the accuracy of our evaluations, we will utilize more precise evaluators in future work.

H LIMITATION

We would like to highlight that the size of SemVarBench is constrained by the necessity for manual verification due to the unsatisfactory accuracy of LLM’s validation, which incurs high costs. Furthermore, the scale of evaluation is limited by the high costs of image generation and LLM-based evaluation, both in terms of time and money, thus restricting the extent of such evaluations.

1890		
1891		
1892		
1893		
1894	Aspect	Category
1895		Action
1896	Relation	Interaction
1897		Absolute Location
1898		Relative Location
1899		Spatial-Temporal
1900		Direction
1901		Attribute Comparison
1902		Size
1903		Height
1904		Weight
1905		Vague Amount
1906		Color
1907	Attribute Values	Counting
1908		Texture
1909		Material
1910		Shape
1911		Age
1912		Sentiment
1913		Temperature
1914		Manner
1915		Appearance
1916		
1917		
1918		
1919		
1920		
1921		
1922		
1923		
1924		
1925		
1926		
1927		
1928		
1929		
1930		
1931		
1932		
1933		
1934		
1935		
1936		
1937		
1938		
1939		
1940		
1941		
1942		
1943		

Table 23: Permutation-based valid sentences (T_a, T_{pv}, T_{pi}) in diverse categories.

1944	The mountain in the distance has snowy peak and the hill by the river has green peak.			
1945				
1946				
1947				
1948				
1949				
1950	SD XL			
1951				
1952				
1953				
1954				
1955				
1956				
1957	FT SD XL			
1958				
1959				
1960				
1961				
1962				
1963				
1964				
1965				
1966				
1967	SD XL			
1968				
1969				
1970				
1971				
1972				
1973				
1974				
1975	FT SD XL			
1976				
1977				
1978				
1979				
1980				
1981				
1982				
1983				
1984				
1985	SD XL			
1986				
1987				
1988				
1989				
1990				
1991				
1992				
1993	FT SD XL			
1994				
1995				
1996				
1997				

Figure 17: More examples for qualitative comparison showing the disconnect between token presence and semantic understanding after fine-tuning.

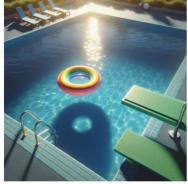
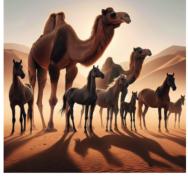
	Anchor Text	Permutation-Variance	Permutation-Variance	SemVarEffect Score	
1998					
1999					
2000					
2001					
2002	There are more smiles than frowns in the photograph.				
2003					
2004					
2005					
2006					
2007					
2008					
2009					
2010	In the pool, there are four floaties and one diving board.				
2011					
2012					
2013					
2014					
2015					
2016					
2017					
2018	The camels are taller than the horses.				
2019					
2020					
2021					
2022					
2023					
2024					
2025	Copper pots with ceramic plates.				
2026					
2027					
2028					
2029					
2030					
2031					
2032					
2033	There's a sleek, modern phone with an old, chunky computer.				
2034					
2035					
2036					
2037					
2038					
2039					
2040	The happy dog is wagging its tail while the cat is sleeping.				
2041					
2042					
2043					
2044					
2045					
2046					
2047					
2048					
2049					
2050					
2051					

Figure 18: The cases which understand semantic variations.

	Anchor Text	Permutation-Variance	Permutation-Variance	SemVarEffect Score
2052				
2053				
2054				
2055				
2056	At the park, few benches and many trees.	At the park, few trees and many benches.	At the park, many trees and few benches.	GPT-4 Turbo
2057				<p>Matched pairs $S(T_a, l_a) = 0.93$ $S(T_{pv}, l_{pv}) = 0.60$ $S(T_{pi}, l_{pi}) = 0.95$</p> <p>Mismatched pairs $S(T_{pv}, l_a) = 0.60$ $S(T_a, l_{pv}) = 0.65$ $S(T_{pi}, l_a) = 0.94$ $S(T_a, l_{pi}) = 0.93$</p>
2058				$\overline{S}_{ll} = 0.83$ $\gamma_{w/}^l = 0.28$ $\gamma_{w/o}^l = 0.01$ $\kappa = 0.27$
2059				
2060				
2061				
2062				
2063	The bag on the hook is heavy and the one on the table is not.	The bag on the table is heavy and the one on the hook is not.	The one on the table is not heavy and the bag on the hook is.	
2064				
2065				
2066				
2067				
2068				
2069				
2070				
2071	Baked potato; first put the butter on the baked potato, and then put the sour cream on top.	Baked potato; first put the sour cream on the baked potato, and then put the butter on top.	Baked potato; first put the butter on the baked potato, then top it with sour cream.	
2072				
2073				
2074				
2075				
2076				
2077				
2078				
2079				
2080	The computer is on the desk and the phone is on the nightstand.	The computer is on the nightstand and the phone is on the desk.	The phone is on the nightstand and the computer is on the desk.	
2081				
2082				
2083				
2084				
2085				
2086				
2087				
2088	A happy family is walking next to a sad ghost.	A sad family is walking next to a happy ghost.	Next to a sad ghost, a happy family is walking.	
2089				
2090				
2091				
2092				
2093				
2094				
2095	The paintings on the wall are realistic and the ones on the floor are abstract.	The paintings on the wall are abstract and the ones on the floor are realistic.	The ones on the floor are abstract and the paintings on the wall are realistic.	
2096				
2097				
2098				
2099				
2100				
2101				
2102				
2103				
2104				
2105				

Figure 19: The cases which don't understand semantic variations.

2106
 2107
 2108
 2109
 2110
 2111
 2112
 2113
 2114
 2115
 2116
 2117
 2118
 2119
 2120
 2121
 2122
 2123
 2124
 2125
 2126
 2127
 2128
 2129
 2130
 2131
 2132
 2133
 2134
 2135
 2136
 2137
 2138
 2139
 2140
 2141
 2142
 2143
 2144
 2145
 2146
 2147
 2148
 2149
 2150
 2151
 2152
 2153
 2154
 2155
 2156
 2157
 2158
 2159

Anchor Text

The baby crawls and the parent walks.



Permutation-Variance

The baby walks and the parent crawls.



Permutation-Variance

The parent walks and the baby crawls.



SemVarEffect Score

GPT-4 Turbo

Matched pairs	$S(T_a, I_a) = 0.94$ $S(T_{pv}, I_{pv}) = 0.93$ $S(T_{pi}, I_{pi}) = 0.98$	$\overline{S}_{ii} = 0.95$
Mismatched pairs	$S(T_{pv}, I_a) = 0.70$ $S(T_a, I_{pv}) = 0.96$ $S(T_{pi}, I_a) = 0.93$ $S(T_a, I_{pi}) = 0.93$	$\gamma_{w/o}^I = 0.06$ $\kappa = 0.19$

A full glass is next to an empty plate.



An empty glass is next to a full plate.



An empty plate is next to a full glass.



The skater wears a denim vest over a graphic t-shirt with a round neck collar.



The skater wears a graphic vest over a denim t-shirt with a round neck collar.



A denim vest is worn by the skater over a graphic t-shirt with a round neck collar.



The elder teacher's hand is on the young student's shoulder.



The young student's hand is on the elder teacher's shoulder.



The young student's shoulder is under the elder teacher's hand.



The mountain in the distance has snowy peak and the hill by the river has green peak.



The mountain in the distance has green peak and the hill by the river has snowy peak.



The hill by the river has a green peak and the mountain in the distance has a snowy peak.



A robot is serving tea to a group of children next to a parent.



A parent is serving tea to a group of children next to a robot.



A robot next to a parent is serving tea to a group of children.



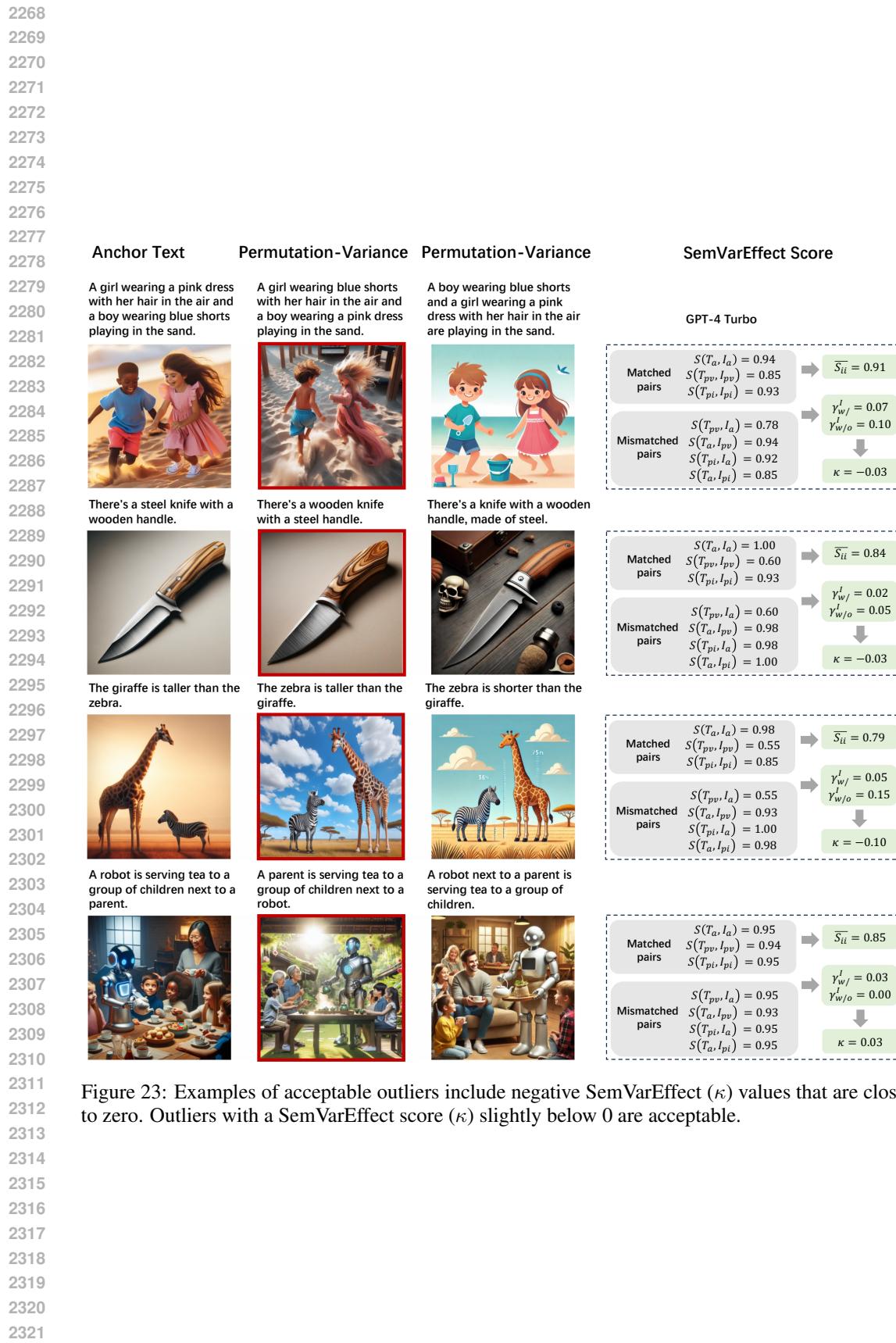
Figure 20: More cases which don't understand semantic variations.

	Anchor Text	Permutation-Variance	Permutation-Variance	SemVarEffect Score
2160				
2161				
2162				
2163				
2164	Four kids riding bikes on the street and one kid skateboarding.	One kid riding bikes on the street and four kids skateboarding.	On the street, four kids are riding bikes and one kid is skateboarding.	GPT-4 Turbo
2165				<p>Matched pairs $S(T_a, I_a) = 0.80$ $S(T_{pv}, I_{pv}) = 0.83$ $S(T_{pi}, I_{pi}) = 0.93$</p> <p>Mismatched pairs $S(T_{pv}, I_a) = 0.30$ $S(T_a, I_{pv}) = 0.55$ $S(T_{pi}, I_a) = 0.90$ $S(T_a, I_{pi}) = 0.93$</p>
2166				$\bar{S}_{li} = 0.85$ $\gamma_{w/o}^I = 0.78$ $\gamma_{w/o}^I = 0.16$ $\kappa = 0.62$
2167				
2168				
2169				
2170				
2171				
2172	The child in the stroller is sleeping and the adult on the bench is reading.	The child in the stroller is reading and the adult on the bench is sleeping.	The adult on the bench is reading and the child in the stroller is sleeping.	
2173				<p>Matched pairs $S(T_a, I_a) = 0.98$ $S(T_{pv}, I_{pv}) = 0.80$ $S(T_{pi}, I_{pi}) = 0.98$</p> <p>Mismatched pairs $S(T_{pv}, I_a) = 0.50$ $S(T_a, I_{pv}) = 0.65$ $S(T_{pi}, I_a) = 0.92$ $S(T_a, I_{pi}) = 0.93$</p>
2174				$\bar{S}_{li} = 0.92$ $\gamma_{w/o}^I = 0.63$ $\gamma_{w/o}^I = 0.11$ $\kappa = 0.52$
2175				
2176				
2177				
2178				
2179				
2180	There's a plastic cup with a ceramic saucer.	There's a ceramic cup with a plastic saucer.	With a ceramic saucer, there's a plastic cup.	
2181				<p>Matched pairs $S(T_a, I_a) = 0.75$ $S(T_{pv}, I_{pv}) = 0.80$ $S(T_{pi}, I_{pi}) = 0.89$</p> <p>Mismatched pairs $S(T_{pv}, I_a) = 0.35$ $S(T_a, I_{pv}) = 0.50$ $S(T_{pi}, I_a) = 0.93$ $S(T_a, I_{pi}) = 0.95$</p>
2182				$\bar{S}_{li} = 0.81$ $\gamma_{w/o}^I = 0.70$ $\gamma_{w/o}^I = 0.24$ $\kappa = 0.45$
2183				
2184				
2185				
2186				
2187	The waiter is covering the eyes of the customer with a menu.	The customer is covering the eyes of the waiter with a menu.	The waiter is covering the customer's eyes with a menu.	
2188				<p>Matched pairs $S(T_a, I_a) = 0.90$ $S(T_{pv}, I_{pv}) = 0.60$ $S(T_{pi}, I_{pi}) = 0.70$</p> <p>Mismatched pairs $S(T_{pv}, I_a) = 0.30$ $S(T_a, I_{pv}) = 0.45$ $S(T_{pi}, I_a) = 0.65$ $S(T_a, I_{pi}) = 0.60$</p>
2189				$\bar{S}_{li} = 0.73$ $\gamma_{w/o}^I = 0.75$ $\gamma_{w/o}^I = 0.35$ $\kappa = 0.40$
2190				
2191				
2192				
2193				
2194				
2195				
2196	The child on the swing is higher than the other children on the seesaw.	The child on the swing is lower than the other children on the seesaw.	The other children on the seesaw are lower than the child on the swing.	
2197				<p>Matched pairs $S(T_a, I_a) = 0.95$ $S(T_{pv}, I_{pv}) = 0.25$ $S(T_{pi}, I_{pi}) = 0.92$</p> <p>Mismatched pairs $S(T_{pv}, I_a) = 0.89$ $S(T_a, I_{pv}) = 0.91$ $S(T_{pi}, I_a) = 0.65$ $S(T_a, I_{pi}) = 0.93$</p>
2198				$\bar{S}_{li} = 0.71$ $\gamma_{w/o}^I = 0.68$ $\gamma_{w/o}^I = 0.29$ $\kappa = 0.39$
2199				
2200				
2201				
2202				
2203				
2204				
2205				
2206				
2207				
2208				
2209				
2210				
2211				
2212				
2213				

Figure 21: Cases with minor errors which understand semantic variations.

	Anchor Text	Permutation-Variance	Permutation-Variance	SemVarEffect Score
2214				
2215				
2216				
2217	A child wearing a superhero cape with their fists in the air and a parent wearing a business suit.	A child wearing a business suit with their fists in the air and a parent wearing a superhero cape.	A parent wearing a business suit and a child wearing a superhero cape with their fists in the air.	GPT-4 Turbo
2218				<p>Matched pairs $S(T_a, I_a) = 0.88$ $S(T_{pv}, I_{pv}) = 0.60$ $S(T_{pi}, I_{pi}) = 0.95$</p> <p>Mismatched pairs $S(T_{pv}, I_a) = 0.15$ $S(T_a, I_{pv}) = 0.98$ $S(T_{pi}, I_a) = 0.82$ $S(T_a, I_{pi}) = 0.95$</p>
2219				$\overline{S}_{ll} = 0.81$ $\gamma_{w/}^l = 0.55$ $\gamma_{w/o}^l = 0.20$ $\kappa = 0.35$
2220				
2221				
2222				
2223				
2224				
2225				
2226	The waiter is wearing a black vest over a white shirt.	The waiter is wearing a white vest over a black shirt.	A black vest is being worn by the waiter over a white shirt.	<p>Matched pairs $S(T_a, I_a) = 1.00$ $S(T_{pv}, I_{pv}) = 0.93$ $S(T_{pi}, I_{pi}) = 0.93$</p> <p>Mismatched pairs $S(T_{pv}, I_a) = 0.65$ $S(T_a, I_{pv}) = 0.93$ $S(T_{pi}, I_a) = 1.00$ $S(T_a, I_{pi}) = 1.00$</p>
2227				$\overline{S}_{ll} = 0.95$ $\gamma_{w/}^l = 0.35$ $\gamma_{w/o}^l = 0.07$ $\kappa = 0.28$
2228				
2229				
2230				
2231				
2232				
2233	The baby's foot is on the mother's chest.	The mother's foot is on the baby's chest.	The mother's chest is under the baby's foot.	<p>Matched pairs $S(T_a, I_a) = 0.75$ $S(T_{pv}, I_{pv}) = 0.53$ $S(T_{pi}, I_{pi}) = 0.55$</p> <p>Mismatched pairs $S(T_{pv}, I_a) = 0.30$ $S(T_a, I_{pv}) = 0.65$ $S(T_{pi}, I_a) = 0.60$ $S(T_a, I_{pi}) = 0.65$</p>
2234				$\overline{S}_{ll} = 0.61$ $\gamma_{w/}^l = 0.33$ $\gamma_{w/o}^l = 0.15$ $\kappa = 0.18$
2235				
2236				
2237				
2238				
2239				
2240				
2241	Two balloons tied to a chair and three balloons floating in the air.	Three balloons tied to a chair and two balloons floating in the air.	Two balloons are tied to a chair, and in the air, three balloons are floating.	<p>Matched pairs $S(T_a, I_a) = 0.65$ $S(T_{pv}, I_{pv}) = 0.70$ $S(T_{pi}, I_{pi}) = 0.65$</p> <p>Mismatched pairs $S(T_{pv}, I_a) = 0.65$ $S(T_a, I_{pv}) = 0.65$ $S(T_{pi}, I_a) = 0.65$ $S(T_a, I_{pi}) = 0.65$</p>
2242				$\overline{S}_{ll} = 0.67$ $\gamma_{w/}^l = 0.05$ $\gamma_{w/o}^l = 0.00$ $\kappa = 0.05$
2243				
2244				
2245				
2246				
2247				
2248				
2249	Chefs in white uniforms with a golden frying pan in their hands.	Chefs in golden uniforms with a white frying pan in their hands.	In white uniforms with a golden frying pan in their hands, chefs.	<p>Matched pairs $S(T_a, I_a) = 0.95$ $S(T_{pv}, I_{pv}) = 0.55$ $S(T_{pi}, I_{pi}) = 0.81$</p> <p>Mismatched pairs $S(T_{pv}, I_a) = 0.55$ $S(T_a, I_{pv}) = 0.78$ $S(T_{pi}, I_a) = 0.98$ $S(T_a, I_{pi}) = 0.83$</p>
2250				$\overline{S}_{ll} = 0.77$ $\gamma_{w/}^l = 0.17$ $\gamma_{w/o}^l = 0.29$ $\kappa = -0.12$
2251				
2252				
2253				
2254				
2255				
2256				
2257	A younger child is hugging the leg of an older parent.	An older parent is hugging the leg of a younger child.	The leg of an older parent is being hugged by a younger child.	<p>Matched pairs $S(T_a, I_a) = 0.70$ $S(T_{pv}, I_{pv}) = 0.65$ $S(T_{pi}, I_{pi}) = 0.95$</p> <p>Mismatched pairs $S(T_{pv}, I_a) = 0.35$ $S(T_a, I_{pv}) = 0.65$ $S(T_{pi}, I_a) = 0.70$ $S(T_a, I_{pi}) = 0.95$</p>
2258				$\overline{S}_{ll} = 0.77$ $\gamma_{w/}^l = 0.35$ $\gamma_{w/o}^l = 0.50$ $\kappa = -0.15$
2259				
2260				
2261				
2262				
2263				
2264				
2265				
2266				
2267				

Figure 22: Cases with minor errors which don't understand semantic variations. Several alignment scores, which are incorrect according to GPT-4V, are labeled in red.



	Anchor Text	Permutation-Variance	Permutation-Variance	SemVarEffect Score
2322				
2323				
2324	The person in the hat is smiling and the person without a hat is frowning.	The person in the hat is frowning and the person without a hat is smiling.	The person without a hat is frowning and the person in the hat is smiling.	GPT-4 Turbo
2325				<p>Matched pairs $S(T_a, I_a) = 0.95$ $S(T_{pv}, I_{pv}) = 0.40$ $S(T_{pi}, I_{pi}) = 0.30$</p> <p>Mismatched pairs $S(T_{pv}, I_a) = 1.00$ $S(T_a, I_{pv}) = 1.00$ $S(T_{pi}, I_a) = 1.00$ $S(T_a, I_{pi}) = 0.60$</p>
2326				$\bar{S}_{li} = 0.55$ $\gamma_{w/o}^l = 0.65$ $\gamma_{w/o}^l = 1.05$ $\kappa = -0.40$
2327				
2328				
2329				
2330				
2331				
2332	All people eat with a fork except for one who eats with chopsticks.	All people eat with chopsticks except for one who eats with a fork.	Except for one who eats with chopsticks, all people eat with a fork.	
2333				
2334				
2335				
2336				
2337				
2338				
2339				
2340	The wooden spoon is in the drawer and the metal spatula is on the counter.	The metal spoon is in the drawer and the wooden spatula is on the counter.	The metal spatula is on the counter and the wooden spoon is in the drawer.	
2341				
2342				
2343				
2344				
2345				
2346				
2347				
2348	The hot coffee is in the mug and the cold tea is in the glass.	The cold tea is in the mug and the hot coffee is in the glass.	The cold tea is in the glass and the hot coffee is in the mug.	
2349				
2350				
2351				
2352				
2353				
2354				
2355				
2356	The ice cream in the cone is melting while the ice cream in the cup is frozen.	The ice cream in the cup is melting while the ice cream in the cone is frozen.	The ice cream in the cup is frozen while the ice cream in the cone is melting.	
2357				
2358				
2359				
2360				
2361				
2362				
2363				
2364	The pockets on the left side of the jacket are big and the ones on the right side are small.	The pockets on the left side of the jacket are small and the ones on the right side are big.	The jacket has big pockets on the left side and small ones on the right side.	
2365				
2366				
2367				
2368				
2369				
2370				
2371				
2372				
2373				
2374				
2375				

Figure 24: Examples of acceptable outliers include negative κ values that are with a SemVarEffect score outside the range [0,1], being considered unacceptable. This discrepancy may be due to incorrect text-image alignment scores provided by evaluators or low quality images.

2376

2377

2378

2379

2380 A shiny ring is next to a dull watch.



2381

2382

2383

2384

2385

2386

2387

A police officer in a black uniform is holding a white flashlight.



2388

2389

2390

2391

2392

2393

2394

2395

2396

A green apple with a brown stem.



2397

2398

2399

2400

2401

2402

2403

2404

The pizza on the tray is round and the sandwich on the plate is square.



2405

2406

2407

2408

2409

2410

The happy child is in the pool and the worried parent is at the edge.



2411

2412

A bird with colorful feathers is flying above a bird without feathers.



2413

2414

2415

2416

2417

2418

A bird without feathers is flying above a bird with colorful feathers.



2419

2420

2421

2422

2423

2424

2425

2426

2427

2428

2429

Anchor Text

Permutation-Variance

Permutation-Variance

SemVarEffect Score

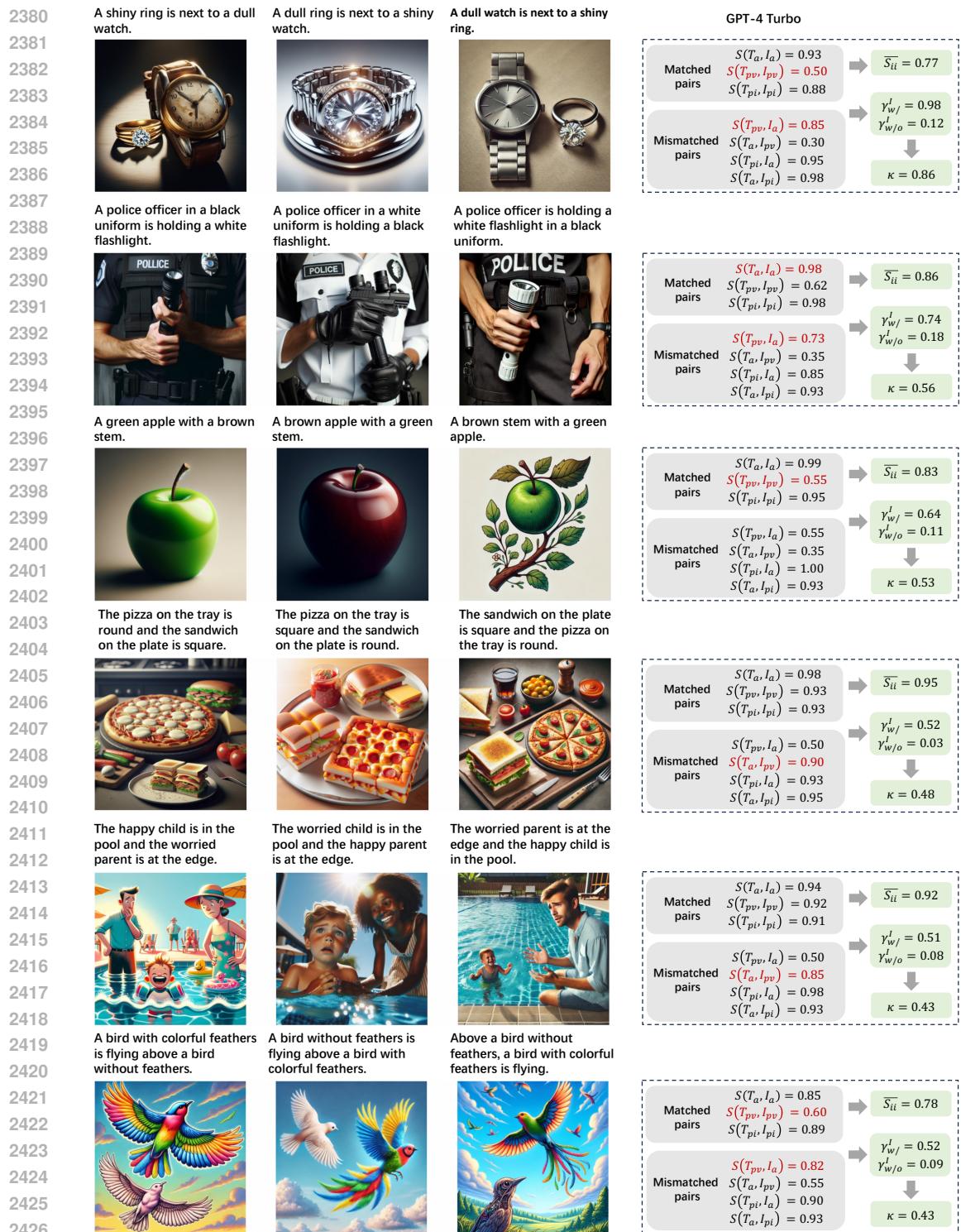


Figure 25: Errors only due to incorrect scoring by GPT-4V, where images are essentially correct.

	Anchor Text	Permutation-Variance	Permutation-Variance	SemVarEffect Score
2430				
2431				
2432				
2433	The happy couple is at the restaurant and the grumpy waiter is at the table.	The grumpy couple is at the restaurant and the happy waiter is at the table.	At the restaurant, the happy couple and the grumpy waiter are at the table.	GPT-4 Turbo
2434				<p>Matched pairs $S(T_a, I_a) = 0.98$ $S(T_{pv}, I_{pv}) = 0.93$ $S(T_{pi}, I_{pi}) = 0.94$</p> <p>Mismatched pairs $S(T_{pv}, I_a) = 0.60$ $\textcolor{red}{S(T_a, I_{pv}) = 0.93}$ $S(T_{pi}, I_a) = 0.97$ $S(T_a, I_{pi}) = 0.95$</p>
2435				$\bar{S}_{ii} = 0.95$
2436				$\gamma_{w/}^I = 0.38$
2437				$\gamma_{w/o}^I = 0.06$
2438				$\kappa = 0.32$
2439				
2440				
2441	A giant squid attacking a ship, and the squid is bigger than the ship.	A ship attacking a giant squid, and the ship is bigger than the squid.	A squid, bigger than the ship, is attacking a ship.	
2442				<p>Matched pairs $S(T_a, I_a) = 0.93$ $S(T_{pv}, I_{pv}) = 0.93$ $S(T_{pi}, I_{pi}) = 0.95$</p> <p>Mismatched pairs $S(T_{pv}, I_a) = 0.75$ $\textcolor{red}{S(T_a, I_{pv}) = 0.94}$ $S(T_{pi}, I_a) = 0.93$ $S(T_a, I_{pi}) = 0.98$</p>
2443				$\bar{S}_{ii} = 0.94$
2444				$\gamma_{w/}^I = 0.19$
2445				$\gamma_{w/o}^I = 0.07$
2446				$\kappa = 0.12$
2447				
2448				
2449	The swimmer in the pool is swimming towards the edge.	The swimmer is at the edge swimming towards the pool.	Towards the edge, the swimmer in the pool is swimming.	
2450				<p>Matched pairs $S(T_a, I_a) = 1.00$ $S(T_{pv}, I_{pv}) = 0.95$ $S(T_{pi}, I_{pi}) = 0.80$</p> <p>Mismatched pairs $S(T_{pv}, I_a) = 0.70$ $\textcolor{red}{S(T_a, I_{pv}) = 0.95}$ $S(T_{pi}, I_a) = 0.90$ $S(T_a, I_{pi}) = 0.85$</p>
2451				$\bar{S}_{ii} = 0.92$
2452				$\gamma_{w/}^I = 0.30$
2453				$\gamma_{w/o}^I = 0.25$
2454				$\kappa = 0.05$
2455				
2456				
2457	There's a silver spoon with a gold handle.	There's a gold spoon with a silver handle.	There's a gold handle with a silver spoon.	
2458				<p>Matched pairs $S(T_a, I_a) = 0.95$ $S(T_{pv}, I_{pv}) = 0.50$ $S(T_{pi}, I_{pi}) = 0.60$</p> <p>Mismatched pairs $S(T_{pv}, I_a) = 0.91$ $\textcolor{red}{S(T_a, I_{pv}) = 1.00}$ $S(T_{pi}, I_a) = 0.97$ $S(T_a, I_{pi}) = 1.00$</p>
2459				$\bar{S}_{ii} = 0.68$
2460				$\gamma_{w/}^I = 0.46$
2461				$\gamma_{w/o}^I = 0.42$
2462				$\kappa = 0.04$
2463				
2464				
2465	The coffee in the mug is black and the tea in the cup is green.	The coffee in the mug is green and the tea in the cup is black.	The tea in the cup is green and the coffee in the mug is black.	
2466				<p>Matched pairs $S(T_a, I_a) = 1.00$ $S(T_{pv}, I_{pv}) = 0.90$ $S(T_{pi}, I_{pi}) = 0.99$</p> <p>Mismatched pairs $S(T_{pv}, I_a) = 0.85$ $\textcolor{red}{S(T_a, I_{pv}) = 0.94}$ $S(T_{pi}, I_a) = 0.83$ $S(T_a, I_{pi}) = 1.00$</p>
2467				$\bar{S}_{ii} = 0.96$
2468				$\gamma_{w/}^I = 0.11$
2469				$\gamma_{w/o}^I = 0.16$
2470				$\kappa = -0.05$
2471				
2472				
2473	The happy baby is in the crib and the unhappy baby is in the stroller.	The unhappy baby is in the crib and the happy baby is in the stroller.	The unhappy baby is in the stroller and the happy baby is in the crib.	
2474				<p>Matched pairs $S(T_a, I_a) = 1.00$ $S(T_{pv}, I_{pv}) = 0.99$ $S(T_{pi}, I_{pi}) = 0.80$</p> <p>Mismatched pairs $S(T_{pv}, I_a) = 1.00$ $\textcolor{red}{S(T_a, I_{pv}) = 1.00}$ $S(T_{pi}, I_a) = 0.95$ $S(T_a, I_{pi}) = 0.98$</p>
2475				$\bar{S}_{ii} = 0.93$
2476				$\gamma_{w/}^I = 0.01$
2477				$\gamma_{w/o}^I = 0.17$
2478				$\kappa = -0.16$
2479				
2480				
2481				
2482				
2483				

Figure 26: Errors only due to incorrect scoring by GPT-4V, where images are essentially correct. The errors heavily influence the SemVarEffect scores.