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.

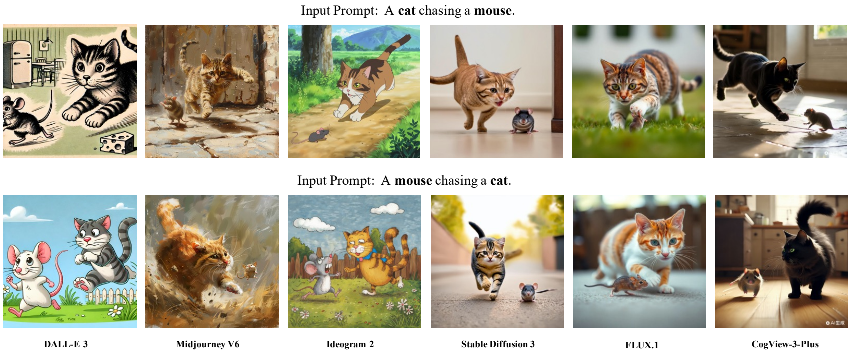

Input Prompt: A **cat** chasing a **mouse**.
Input Prompt: A **mouse** chasing a **cat**.
DALL-E 3
Midjourney V6
Ideogram 2
Stable Diffusion 3
FLUX.1
CogView-3-Plus

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.

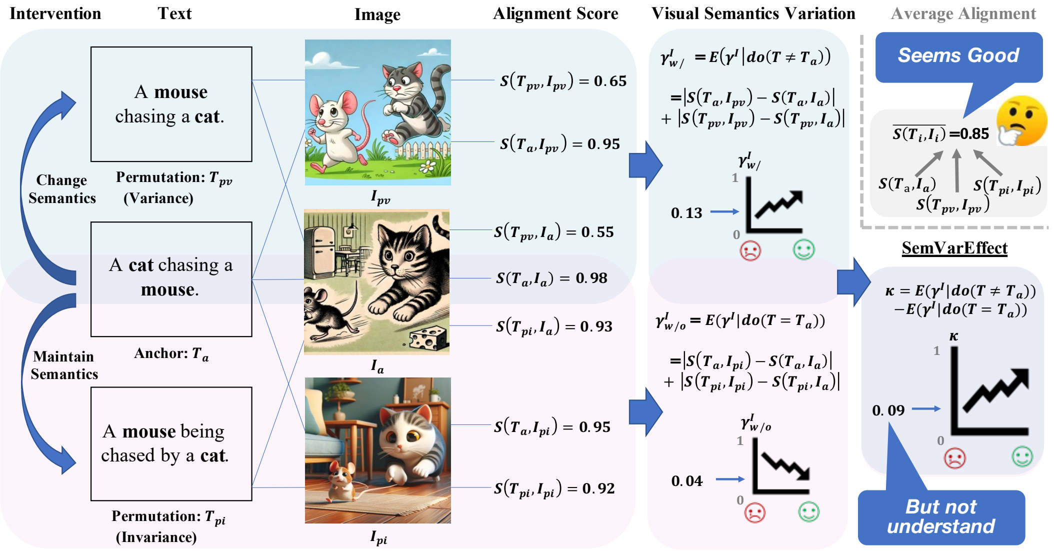

Intervention
Text
Image
Alignment Score
Visual Semantics Variation
Average Alignment
A **mouse** chasing a **cat**.
Permutation: T_{pv} (Variance)
Change Semantics
A **cat** chasing a **mouse**.
Anchor: T_a
Maintain Semantics
A **mouse** being chased by a **cat**.
Permutation: T_{pi} (Invariance)
I_{pv}
I_a
I_{pi}
$S(T_{pv}, I_{pv}) = 0.65$
$S(T_a, I_{pv}) = 0.95$
$S(T_{pv}, I_a) = 0.55$
$S(T_a, I_a) = 0.98$
$S(T_{pi}, I_a) = 0.93$
$S(T_a, I_{pi}) = 0.95$
$S(T_{pi}, I_{pi}) = 0.92$
$\gamma^I_{w/} = E(\gamma^I | do(T \neq T_a))$
$= |S(T_a, I_{pv}) - S(T_a, I_a)| + |S(T_{pv}, I_{pv}) - S(T_{pv}, I_a)|$
$\gamma^I_{w/}$
1
0.13
0
$\gamma^I_{w/o} = E(\gamma^I | do(T = T_a))$
$= |S(T_a, I_{pi}) - S(T_a, I_a)| + |S(T_{pi}, I_{pi}) - S(T_{pi}, I_a)|$
$\gamma^I_{w/o}$
1
0.04
0
Seems Good
$\overline{S(T_i, I_i)} = 0.85$
$S(T_a, I_a)$
$S(T_{pv}, I_{pv})$
$S(T_{pi}, I_{pi})$
SemVarEffect
$\kappa = E(\gamma^I | do(T \neq T_a)) - E(\gamma^I | do(T = T_a))$
κ
1
0.09
0
But not understand

Figure 2: Measuring the causality between semantic variations in inputs and outputs of T2I models. **Blue areas** show alignment scores (by GPT-4) and resulting semantic variations when the semantics are changed. **Pink areas** show alignment scores and resulting semantic variations when the semantics are maintained. **Purple areas on the far right** show the SemVarEffect score, reflecting the causal contribution of input changes to output variations based on the two interventions. **Gray areas** show the average alignment score commonly used in previous T2I studies, focusing on similarity rather than causality. For detailed explanations of the symbols and calculations, see Section 2.

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 contributions are summarized as follows: (1) We are the first to conduct a comprehensive investigation into the problem of semantic variations in T2I models. (2) We propose SemVarEffect, a novel metric specifically designed to measure the causality between semantic variations in inputs and outputs of T2I synthesis. (3) We propose SemVarBench, a high-quality, expert-annotated benchmark for evaluating semantic variations in T2I synthesis, avoiding predictable literal differences and focusing on two key linguistic permutations: permutation-variance and permutation-invariance. This benchmark sets a new standard for evaluating T2I models on semantic understanding. (4) We conduct a comprehensive benchmarking of SOTA T2I models on SemVarBench, revealing significant limitations in their handling of semantic variations. Our findings suggest several areas for improvement, including text encoding and cross-modal semantic alignment techniques, and offer insights into the challenges posed by different types of semantic variations.

2 SEMANTIC VARIATION EVALUATION FOR TEXT-TO-IMAGE SYNTHESIS

2.1 PRELIMINARY

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

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

2.2 DEFINITION OF VISUAL SEMANTIC VARIATIONS

First, we define the visual semantic variations observed from a single sentence T. For each I and its localized variation $I + \Delta I$ in the image space, the visual semantic variation at I, denoted as $\mu_I(T, I)$, is the difference in alignment scores between the two images for the same sentence: $\mu_I(T, I) = S(T, I + \Delta I) - S(T, I)$. If the anchor image I_a is transformed into a permutation image I_p through a series of localized changes, the total visual semantic variation from I_a to I_p is the sum of variations across all localized changes: $\sum_{I_a}^{I_{p*}} \mu_I(T, I) = S(T, I_{p*}) - S(T, I_a)$.

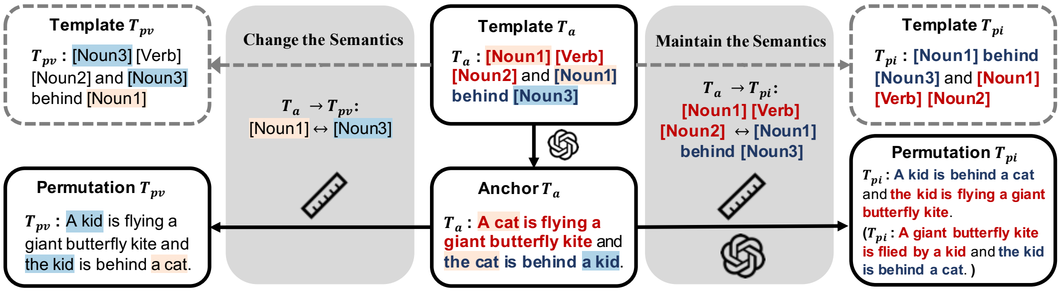
Second, we integrate the visual semantic variations observed across multiple sentences. For the sentence T_a, the visual semantic variations $\sum_{I_a}^{I_{p*}} \mu(T_a, I)$ demonstrate a shift from a matched to a mismatched image-text pair, indicating a negative change. For the sentence T_{p*}, the visual semantic variations $\sum_{I_a}^{I_{p*}} \mu(T_{p*}, I)$ demonstrate a shift from a mismatched to a matched image-text pair, indicating a positive change. To measure the total magnitude of these variations regardless of direction, we use the absolute values. Therefore, the integrated visual semantic variations γ^I is defined as:

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

2.3 THE CAUSALITY BETWEEN TEXTUAL AND VISUAL SEMANTIC VARIATIONS

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

According to the causal inference theory, we define the average causal effect (ACE) of textual semantic variations on visual semantic variations as the SemVarEffect score. As shown in Fig. 3, the sentence T serves as an independent variable that influences the generated image I. The visual 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 two types of interventions. $\text{do}(T \neq T_a)$ represents an intervention where T differs in meaning from



T_{pv} : [Noun3] [Verb] [Noun2] and [Noun3] behind [Noun1]

T_a : [Noun1] [Verb] [Noun2] and [Noun1] behind [Noun3]

T_{pi} : [Noun1] behind [Noun3] and [Noun1] [Verb] [Noun2]

T_{pv} : A kid is flying a giant butterfly kite and the kid is behind a cat.

T_a : A cat is flying a giant butterfly kite and the cat is behind a kid.

T_{pi} : A kid is behind a cat and the kid is flying a giant butterfly kite.

(T_{pi} : A giant butterfly kite is flied by a kid and the kid is behind a cat.)

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".

the anchor sentence T_a. The visual semantic variation caused by this intervention is denoted as:

$$\begin{aligned}\gamma^I_{w/} &= E[\gamma^I \mid \mathrm{do}(T \neq T_a)] = E[\gamma^I \mid T = T_{pv}] \\ &= |S(T_a, I_{pv}) - S(T_a, I_a)| + |S(T_{pv}, I_{pv}) - S(T_{pv}, I_a)| \,.\end{aligned} \tag{2}$$

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:

$$\begin{aligned}\gamma^I_{w/o} &= E[\gamma^I \mid \mathrm{do}(T = T_a)] = E[\gamma^I \mid T = T_{pi}], \\ &= |S(T_a, I_{pi}) - S(T_a, I_a)| + |S(T_{pi}, I_{pi}) - S(T_{pi}, I_a)| \,.\end{aligned} \tag{3}$$

By comparing the visual semantic variations under the two interventions—one changing the meaning and the other maintaining it—we determine the ACE of textual semantic variations on visual semantic variations:

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

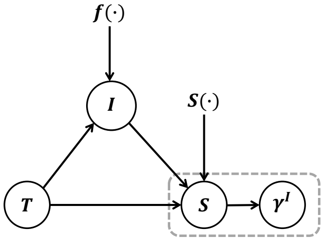


Figure 3: Causal relationship between the input and the output semantic variations.

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 $\gamma^I_{w/}$ reaches its upper bound of 1, which occurs in extreme cases where no relation between objects are identical, and $\gamma_{w/o}$ 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 collect semantic variation datasets for T2I synthesis to fill the gaps in current benchmarks and evaluation practices. The textual semantic variations are created through two typical linguistic permutations. First, we elaborate on the characteristics of the data. Then, we introduce the pipeline for data collection, annotation and statistics.

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} evoke vivid mental images.

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.

3.2 DATA COLLECTION

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

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

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

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

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

3.3 DATA ANNOTATION AND STATISTICS

LLM and Human Annotation. We establish 14 specific criteria to define what constitutes a "valid" input sample. LLMs check each sample against these criteria, labeling them as "yes" or "no" with confidence scores. Samples labeled "no" with confidence scores above 0.8 are removed. Then, 15 annotators and 3 experienced experts manually verify the remaining samples. Each sample is independently reviewed by two annotators, with an expert resolving any disagreements. After expert verification, we obtained 11,454 valid, non-duplicated samples. To rigorously evaluate T2I models, 684 challenging samples were selected based on thresholds and voting for the test set. More details on annotation and selection are provided in Appendix C.2.

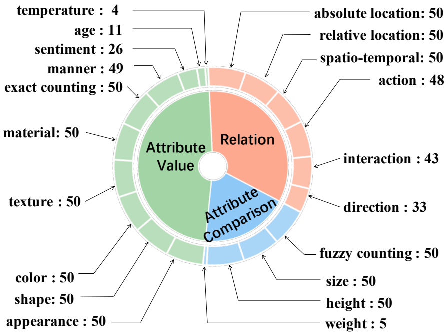


Figure 5: Distribution of semantic variations by category in the semVarBench test set.

Scale and Split. SemVarBench comprises 11,454 samples of (T_a, T_{pv}, T_{pi}), divided into a training set and a test set. The training set contains 10,806 samples, while the test set consists of 648 samples. All our evaluations are conducted on the 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.1[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	MidJ V6	Diffusion	–	–	–	–	–	–	–	–
DALL-E 3 Betker et al. (2023)	DALL-E 3	Diffusion	–	T5-XXL	4.76B	UNet	–	VAE	–	–
CogView-3-Plus	CogV3-Plus	Diffusion	–	T5-XXL[1]	4.76B[1]	Transformer	–	VAE[1]	–	–
Ideogram 2	Ideogram 2	Diffusion	–	–	–	–	–	–	–	–

[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: Diffusion 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.

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

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

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

Evaluators. We use 4 advanced MLLMs as the 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 $[0, 1]$.

Metrics. We use 4 metrics for evaluation: the alignment score $\bar{S_{ii}}$, the visual semantic variation scores $\gamma_{w/}$ and $\gamma_{w/o}$ under different interventions as defined in Eq. 2 and Eq. 3, and the SemVar-Effect score κ as defined in Eq. 4. For each sample, $\bar{S_{ii}} = \frac{1}{|K|}\sum_{i \in K} S(T_i, I_i)$, where $K = \{a, pv, pi\}$. A T2I model with high $\bar{S_{ii}}$, $\gamma_{w/}$ and κ, and low $\gamma_{w/o}$ indicates a strong understanding of semantic variations. For simplicity, we refer to $\bar{S_{ii}}$ as $\bar{S}$, $\gamma_{w/}$ as γ_w, and $\gamma_{w/o}$ as γ_{wo} in the following sections.

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

Models	Gemini-1.5-Pro				Claude-3.5-Sonnet				GPT-4o				GPT-4-Turbo			
	$\bar{S}(\uparrow)$	$\gamma_w(\uparrow)$	$\gamma_{wo}(\downarrow)$	$\kappa(\uparrow)$	$\bar{S}(\uparrow)$	$\gamma_w(\uparrow)$	$\gamma_{wo}(\downarrow)$	$\kappa(\uparrow)$	$\bar{S}(\uparrow)$	$\gamma_w(\uparrow)$	$\gamma_{wo}(\downarrow)$	$\kappa(\uparrow)$	$\bar{S}(\uparrow)$	$\gamma_w(\uparrow)$	$\gamma_{wo}(\downarrow)$	$\kappa(\uparrow)$
Open-source Models																
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
SD 2.1	0.58	0.45	0.46	-0.01	0.66	0.21	0.20	0.01	0.65	0.33	0.31	0.02	0.68	0.35	0.34	0.01
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
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
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
PixArt	0.60	0.35	0.32	0.02	0.69	0.17	**0.15**	0.02	0.70	0.29	0.26	0.03	0.71	0.29	0.27	0.02
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
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
FLUX.1	0.72	0.43	0.35	0.08	0.75	0.23	0.17	0.06	0.72	0.42	0.33	0.10	0.75	0.40	0.30	0.10
API-based Models																
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
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**	0.13	**0.83**	0.35	0.30	0.10
CogV3-Plus	0.79	**0.52**	0.35	0.17	**0.80**	**0.28**	0.18	**0.10**	0.81	**0.49**	0.28	**0.20**	0.82	**0.43**	0.26	**0.17**
Ideogram 2	**0.80**	0.47	**0.29**	**0.18**	0.79	0.26	0.17	0.09	0.81	0.46	0.27	**0.20**	0.81	0.40	**0.24**	0.15

Table 2: Evaluation results of different T2I models in understanding semantic variations.

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.

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 they still struggle to accurately differentiate between inputs with varying meanings and those with invariant meanings. 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.

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 similarity[1] 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.

---

[1] Sentences 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.

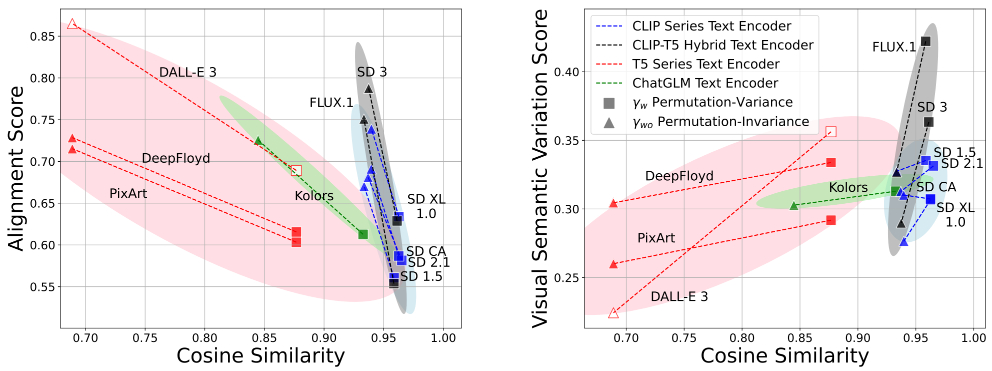


Figure 6: Illustration of the text embedding similarity between the anchor text and the permuted text. Squares represent permutation-variance results (with changed textual semantics), while triangles represent permutation-invariance results (with unchanged textual semantics). The evaluator is GPT-4o. (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.

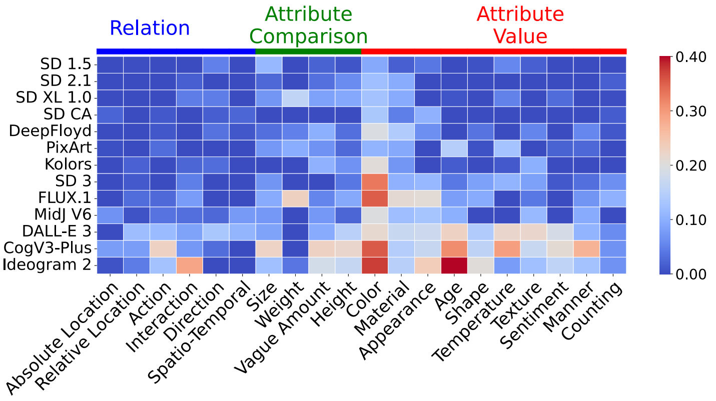


Figure 7: The distribution of categories across different T2I models based on SemVarEffect scores κ. The evaluator is GPT-4 Turbo.

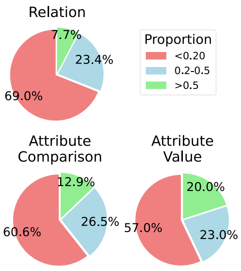


Figure 8: The distribution of SemVarEffect scores for the SOTA model Ideogram 2 across different aspects of the samples. The evaluator is GPT-4 Turbo.

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 linked to word-level alignment in the cross-attention mechanism and a lack of semantic-level constraints. Fine-tuning only improves alignment at the word level, rather than enhancing the understanding of semantic variations. The samples from permutation-variance are inherently hard negative samples, as they differ only in word order. This confuses the models and leads to performance declines, especially during DPO. Fine-tuning T2I models to better understand semantic variations caused by word order remains a formidable challenge.

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(↓)

Category	Models	GPT-4o			
		$\bar{S}(\uparrow)$	$\gamma_w(\uparrow)$	$\gamma_{wo}(\downarrow)$	$\kappa(\uparrow)$
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(↓)

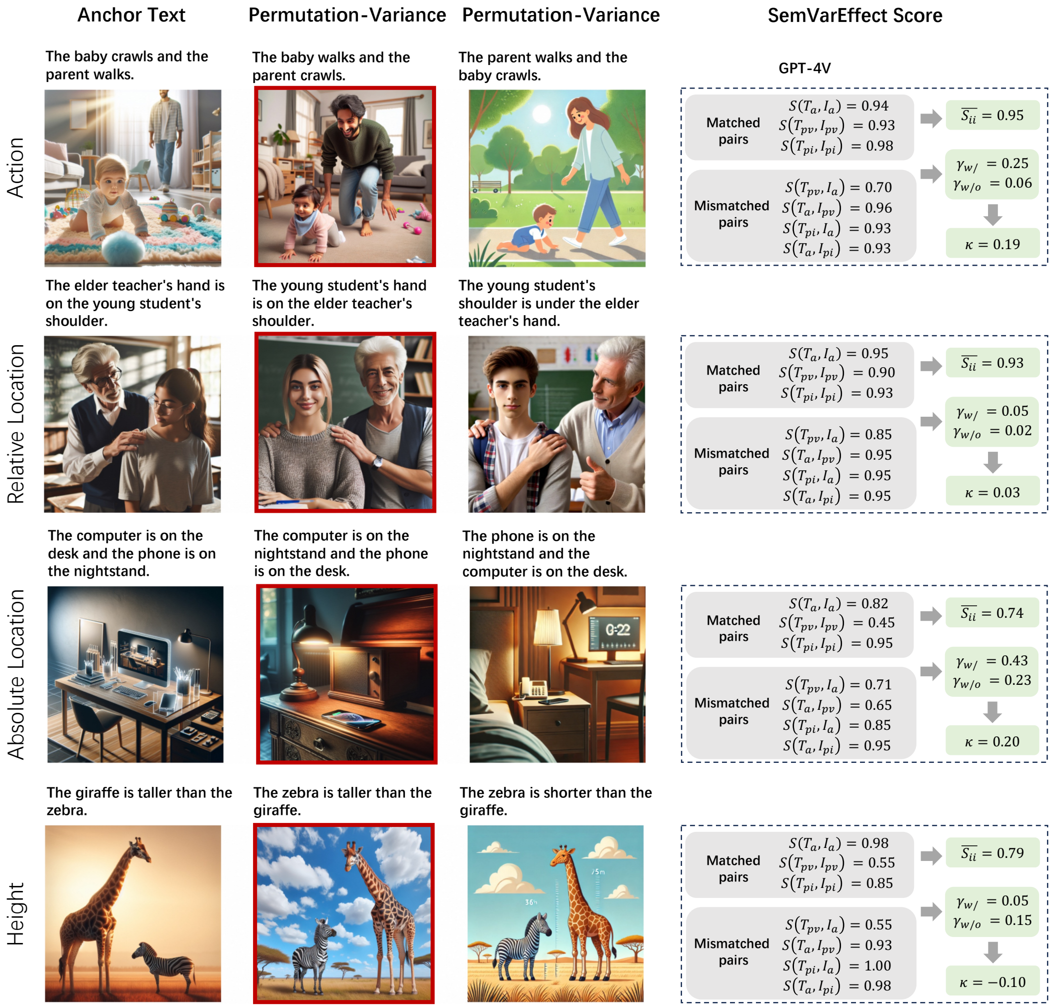
Table 3: Fine-tuned SD XL Results: Performance varies based on the quantity of high-quality samples, which are determined by category and sample size. Left Table: Color and Absolute Location with 4.4k and 1.7k candidates for training. Right Table: Height and Direction with 0.2k and 0.3k candidates for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

5 RELATED WORK

Evaluation of T2I synthesis. Benchmarks of T2I synthesis primarily focus on general alignment Saharia et al. (2022); Yu et al. (2022); Cho et al. (2022), composition Park et al. (2021); Feng et al. (2023); Park et al. (2021); Hu et al. (2023); Cho et al. (2023b); Li et al. (2024), bias and fairness Lee et al. (2023b); Luo et al. (2024b;a), common sense Fu et al. (2024) and creativity Lee et al. (2023b). In these evaluations, the quality of images is measured by detection-based or alignment-based metrics. Detection-based metrics evaluate the accuracy of object detection models applied to the generated images, while alignment-based metrics evaluate how well the visual content matches the semantic meaning of the text. Recent research on T2I synthesis has explored samples involving semantic variations caused by word orders, typically using them to evaluate reasoning abilities with alignment-based metrics Marcus et al. (2022); Lee et al. (2023b); Li et al. (2024). However, a significant gap in this research is the underexplored area of whether the generated images consistently represent subtle but important semantic variations within the input text.

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

The application of Winoground to T2I synthesis faces several limitations due to the variety and quantity of its permutations. First, the dataset, with 400 sentence pairs, provides only 171 suitable for text-image composition analysis Diwan et al. (2022), where samples are classified into three categories: object, relation, and both. This limited variety is insufficient for a comprehensive evaluation



The baby crawls and the parent walks.

The baby walks and the parent crawls.

The parent walks and the baby crawls.

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 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.

The giraffe is taller than the zebra.

The zebra is taller than the giraffe.

The zebra is shorter than the giraffe.

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

of T2I models. Second, the suitability of certain samples for T2I model evaluation is problematic. Winoground primarily focuses on semantic distinctiveness for cross-modal retrieval Yüksekgönül et al. (2023); Ma et al. (2023); Cascante-Bonilla et al. (2023). It overlooks the criteria essential for T2I synthesis, such as sentence completeness, clarity of expression, unambiguity, and specificity in referencing image elements. All of these factors have been carefully considered in the quality control of our benchmark annotations.

6 CONCLUSION

We comprehensively study the challenge of semantic variations in T2I synthesis, specifically focusing on causality between semantic variations of inputs and outputs. We propose a new metric, SemVarEffect, to quantify the influence of input semantic variations on model outputs, and a novel benchmark, SemVarBench, designed to examine T2I models' understanding of semantic variations. Our experiments reveal that SOTA T2I models, including CogView-3-Plus and Ideogram 2, struggle with semantic variations, with most scoring below 0.2 on our benchmark. This indicates that these models have yet to develop the capability to effectively handle such variations. Fine-tuning efforts also show limited success, improving sensitivity to certain variations but at the cost of robustness. These findings highlight the importance of our metric and benchmark in addressing this challenge. Future work should focus on enhancing cross-modal alignment to better manage subtle semantic changes and improve overall T2I model performance.

REFERENCES

Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. *CoRR*, abs/2304.05390, 2023. doi: 10.48550/ARXIV.2304.05390. URL `https://doi.org/10.48550/arXiv.2304.05390`.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. 2023. URL `https://cdn.openai.com/papers/dall-e-3.pdf`.

Jirayu Burapacheep, Ishan Gaur, Agam Bhatia, and Tristan Thrush. Colorswap: A color and word order dataset for multimodal evaluation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 1716–1726. Association for Computational Linguistics, 2024. URL `https://aclanthology.org/2024.findings-acl.99`.

Pu Cao, Feng Zhou, Qing Song, and Lu Yang. Controllable generation with text-to-image diffusion models: A survey. *arXiv preprint arXiv:2403.04279*, 2024.

Paola Cascante-Bonilla, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gül Varol, Aude Oliva, Vicente Ordonez, Rogério Feris, and Leonid Karlinsky. Going beyond nouns with vision & language models using synthetic data. *CoRR*, abs/2303.17590, 2023. doi: 10.48550/ARXIV.2303.17590. URL `https://doi.org/10.48550/arXiv.2303.17590`.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.

Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, et al. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? *arXiv preprint arXiv:2407.04842*, 2024.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *CoRR*, abs/2202.04053, 2022.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. *CoRR*, abs/2310.18235, 2023a. doi: 10.48550/ARXIV.2310.18235. URL `https://doi.org/10.48550/arXiv.2310.18235`.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for text-to-image generation and evaluation. *CoRR*, abs/2305.15328, 2023b. doi: 10.48550/ARXIV.2305.15328. URL `https://doi.org/10.48550/arXiv.2305.15328`.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. Why is winoground hard? investigating failures in visuolinguistic compositionality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 2236–2250. Association for Computational Linguistics, 2022.

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.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun R. Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Xingyu Fu, Muyu He, Yujie Lu, William Yang Wang, and Dan Roth. Commonsense-t2i challenge: Can text-to-image generation models understand commonsense? *arXiv preprint arXiv:2406.07546*, 2024.

Hanan Gani, Shariq Farooq Bhat, Muzammal Naseer, Salman Khan, and Peter Wonka. Llm blueprint: Enabling text-to-image generation with complex and detailed prompts. *arXiv preprint arXiv:2310.10640*, 2023.

Matthias Gerner. Predicate-induced permutation groups. *J. Semant.*, 29(1):109–144, 2012. doi: 10.1093/JOS/FFR007. URL `https://doi.org/10.1093/jos/ffr007`.

Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*, 2022.

Jonathan Gordon, David Lopez-Paz, Marco Baroni, and Diane Bouchacourt. Permutation equivariant models for compositional generalization in language. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL `https://openreview.net/forum?id=SylVNerFvr`.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36, 2024.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: accurate and interpretable text-to-image faithfulness evaluation with question answering. *CoRR*, abs/2303.11897, 2023. doi: 10.48550/ARXIV.2303.11897. URL `https://doi.org/10.48550/arXiv.2303.11897`.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *CoRR*, abs/2307.06350, 2023.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023a.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Li Fei-Fei, Jiajun Wu, Stefano Ermon, and Percy Liang. Holistic evaluation of text-to-image models. *CoRR*, abs/2311.04287, 2023b. doi: 10.48550/ARXIV.2311.04287. URL `https://doi.org/10.48550/arXiv.2311.04287`.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Evaluating and improving compositional text-to-visual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5290–5301, 2024.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pp. 423–439. Springer, 2022.

Hanjun Luo, Ziye Deng, Ruizhe Chen, and Zuozhu Liu. Faintbench: A holistic and precise benchmark for bias evaluation in text-to-image models. *arXiv preprint arXiv:2405.17814*, 2024a.

Hanjun Luo, Haoyu Huang, Ziye Deng, Xuecheng Liu, Ruizhe Chen, and Zuozhu Liu. Bigbench: A unified benchmark for social bias in text-to-image generative models based on multi-modal llm. *arXiv preprint arXiv:2407.15240*, 2024b.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. @ CREPE: can vision-language foundation models reason compositionally? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 10910–10921. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01050. URL `https://doi.org/10.1109/CVPR52729.2023.01050`.

Gary Marcus, Ernest Davis, and Scott Aaronson. A very preliminary analysis of DALL-E 2. *CoRR*, abs/2204.13807, 2022. doi: 10.48550/ARXIV.2204.13807. URL `https://doi.org/10.48550/arXiv.2204.13807`.

Wenyi Mo, Tianyu Zhang, Yalong Bai, Bing Su, Ji-Rong Wen, and Qing Yang. Dynamic prompt optimizing for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26627–26636, 2024.

Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

Pablo Pernias, Dominic Rampas, Mats L. Richter, Christopher J. Pal, and Marc Aubreville. Wuerstchen: An efficient architecture for large-scale text-to-image diffusion models, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. *CoRR*, abs/2307.01952, 2023. doi: 10.48550/ARXIV.2307.01952. URL `https://doi.org/10.48550/arXiv.2307.01952`.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10674–10685. IEEE, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. URL `http://papers.nips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html`.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34:1415–1428, 2021.

Kolors Team. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*, 2024.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 5228–5238. IEEE, 2022.

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

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

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

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

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

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

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

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

The Appendix is organized as follows:

- Section A provide a detailed illustrations for four types of semantic variations results in T2I Synthesis.
- Section B provide properties of the metric SemVarEffect relative to the implementation of the dataset and the alignment function.
- Section C provides construction details for the benchmark.
- Section D presents the implementation details of the evaluation.
- Section E presents more experimental results.
- Section F presents more analysis about the results.
- Section G visualizes more successful and failed examples in the evaluation.
- Section H presents limitation of our evaluation and benchmark.

A FOUR TYPES OF SEMANTIC VARIATIONS RESULTS IN T2I SYNTHESIS

The results of linguistic permutation in T2I synthesis in text and images can be divided into four types, as shown in Fig. 10.

- **Image Changing Semantics with Text Changing Semantics**: As shown in the first quartile, it means the model tends to understand the different semantics achieved by linguistic permutation. In this case, the value of $\gamma_{w/}$ will tend to 1.
- **Image Maintaining Semantics with Text Changing Semantics**: As shown in the forth quartile, it means the model doesn't understand the different semantics achieved by linguistic permutation. In this case, the value of $\gamma_{w/}$ will tend to 0.
- **Image Changing Semantics with Text Maintaining Semantics**: As shown in the third quartile, it means the model tends to understand the similar semantics achieved by linguistic permutation. In this case, the value of $\gamma_{w/o}$ will tend to 0.
- **Image Maintaining Semantics with Text Maintaining Semantics**: As shown in the second quartile, it means the model doesn't understand the similar semantics achieved by linguistic permutation. In this case, the value of $\gamma_{w/o}$ will tend to 1.

B PROPERTIES OF SEMVAREFFECT

B.1 PRELIMINARY

A T2I generation model f consists of one or more text encoders f_t and image decoders f_v. The T2I generation model f generates images $I \in \mathcal{I}$ for each input textual prompt $T \in \mathcal{T}$. $\mathcal{T}$ is the textual space and $\mathcal{I}$ is the visual space. $S(T, I)$ is the alignment score between T and I.

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

B.2 TEXTUAL VS. VISUAL SEMANTIC VARIATIONS

The measurement of semantic variations in the variation from (T_a, I_a) to (T_{p*}, I_{p*}) can be defined from two perspectives: (1) textual semantic variations r^T: The semantic variations measured by text changes observed from the images I_a and I_{p*}, and (2) visual semantic variations r^I: The semantic variations measured by image changes observed from the texts T_a and T_{p*}.

Specifically, we define the textual semantic variations observed from a single image I. The initial anchor sentence T_a is transformed into the a permutation T_{p*} after a series of localized linguistic permutations. For each T to $T + \Delta T$ in the text space, the textual semantic variations at position T,

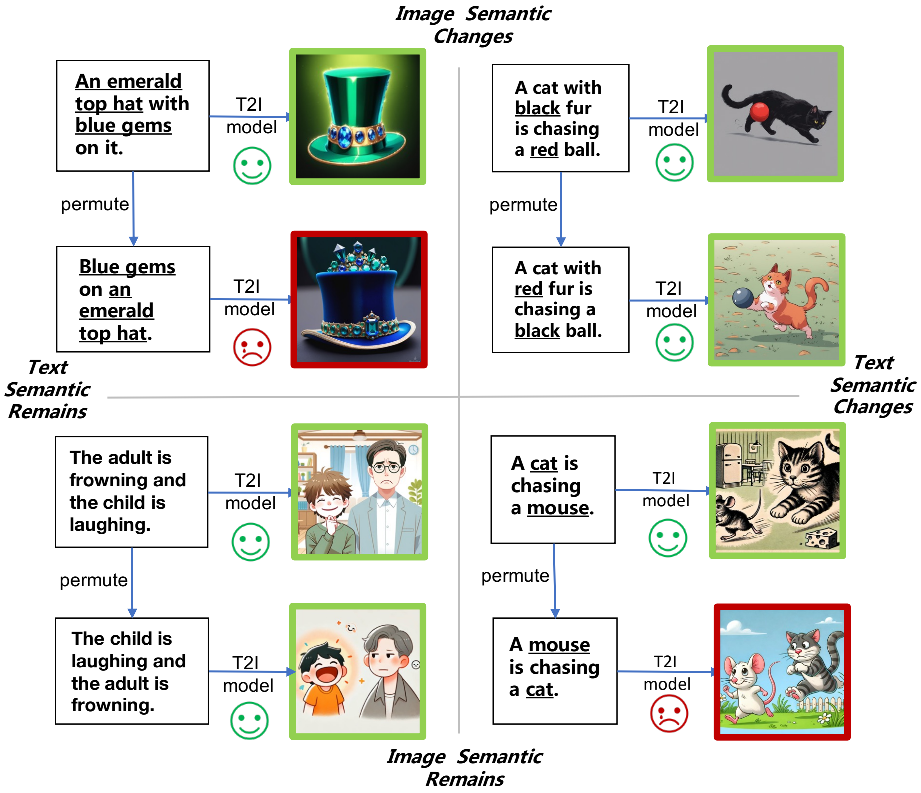


Image Semantic Changes
An emerald top hat with blue gems on it.
T2I model
permute
Blue gems on an emerald top hat.
T2I model
A cat with black fur is chasing a red ball.
T2I model
permute
A cat with red fur is chasing a black ball.
T2I model
Text Semantic Remains
Text Semantic Changes
The adult is frowning and the child is laughing.
T2I model
permute
The child is laughing and the adult is frowning.
T2I model
A cat is chasing a mouse.
T2I model
permute
A mouse is chasing a cat.
T2I model
Image Semantic Remains

Figure 10: Four Types of Linguistic Permutation Results in T2I Synthesis. The images with a red border represent an incorrect output. The images with a green border represent a correct output.

denoted as $\mu_T(T, I)$, is the difference in alignment scores between different texts and that image: $\mu_T(T, I) = S(T+\Delta T, I) - S(T, I)$. The textual semantic variations from T_a to T_{p*} is the sum of the textual semantic variations produced by each localized permutations: $\sum_{T_a}^{T_{p*}} \mu_T(T, I) = S(T_{p*}, I) - S(T_a, I)$. Therefore, the integrate textual semantic variation are $\gamma^T = \sum_{I \in \{I_a, I_{p*}\}} \left| \sum_{T_a}^{T_{p*}} \mu(T, I) \right|$.

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

We have not evaluate the influence by measuring the synchronicity of semantic changes between images and text, which has applied in VLM Wang et al. (2023). Because semantic variations bring a unique challenge in the evaluation of T2I synthesis: images are not independent; they are influenced by both the input text and the model's inherent characteristics, complicating the independent measurement of semantic changes.

Therefore, we conduct the evaluation by measuring the influence of semantic variations within their corresponding images in T2I synthesis, while avoiding directly imposing unintended interventions on images.

B.3 PROPERTIES OF ALIGNMENT SCORES S

Definition of Text-Image Alignment Score. To facilitate semantic analysis, we structured these permutations by objects and triples. The changing of word order leads to the swapping of objects or relations and the change of syntactic dependencies and semantics. Let T_{p*} represents any permutation of T_a. T_a and T_{p*} share the same objects set V and relations set R. The triple 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, yet they have the same number of triples. For example, the initial triple set of T_a contains *(apple, on, box)*, *(girl, touch, apple)* and *(girl, NULL, box)*. After swapping *box* and *apple*, the triple set of T_{p*} contains *(box, on, apple)*, *(girl, touch, box)* and *(girl, NULL, apple)*.

To calculate the fine-grained alignment scores for objects and triples, we define the alignment score S between T and I as the sum of the alignment scores of objects and triples:

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

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

$$\begin{aligned} S_{obj_i}(T, I) &= \begin{cases} w_{v_i} & \text{if the } i\text{-th object matches,} \\ 0 & \text{if the } i\text{-th object does not match,} \end{cases} \\ S_{tri_j}(T, I) &= \begin{cases} w_{e_j} & \text{if the } j\text{-th triple matches,} \\ 0 & \text{if the } j\text{-th triple does not match,} \end{cases} \end{aligned} \tag{6}$$

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

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

where V_{MA} and E_{MA} represents the exactly matched objects and triples between a text prompt and the image generated from this text, with $|V_{MA}| = |V|$ and $|E_{MA}| = |E|$. The alignment score of a mismatched text-image pair is calculated as:

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

where V_{MI} and E_{MI} represents the partially matched objects and triples between a text prompt and a mismatched image, with $|V_{MI}| = |V|$ and $0 \leq |E_{MI}| \leq |E|$.

Range of S. If f accurately depicting the text by images and S faithfully measure the semantic changes between text space and image space, any alignment score $S(T, I)$ is bounded by:

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

In our implementation, we set the value of $S(T, I)$ as a integer in $0 - 100$, where the object accuracy is in $0 - 50$ and the triple accuracy is in $0 - 50$. Then we normalize it into a real number in $[0, 1]$. Based on the assumption of f mentioned above, $0.5 \leq S(T, I) \leq 1$.

However, limitations in the capabilities of the model f and the alignment function S, often prevent the alignment score values from achieving the property in Eq. 10. For example, if a model f generated a a low-quality image I, it may fail to depict all objects within the target set V, leading to $|V_{MA}| < |V|$ and $|V_{MI}| < |V|$. This can result in an object accuracy below 0.5 (as illustrated in the bottom case of Fig. 21) and inconsistent relation accuracy (see cases in Fig. 16 and Fig. 17). Furthermore, if the scoring approach $S(\cdot)$ is inaccurate, it may incorrectly evaluate the similarity between a text prompt and a generated image. This can cause significant and unpredictable fluctuations in the semantic variation measurements (as illustrated in Fig. 23). These limitations render a direct comparison of textual and visual semantic variation scores unreliable, as attempted in Wang et al. (2023).

Identity Relation for S. In the ideal scenarios that f accurately transforms all semantic variations from text space to image space, the alignment scores would satisfy the constraints that:

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

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

However, the low-quality image I, resulting from a poorly performing model f, and the inaccuracies in the approach to scoring S, often prevent the alignment scores from meeting the criteria of Eq. 10. This limitation hinders our measurements by comparing the differences between the scores of textual semantic variations and visual semantic variations, as attempted in Wang et al. (2023).

B.4 PROPERTY OF VISUAL SEMANTIC VARIATIONS γ^I

We analyse the theoretical relationship between the visual semantic variation and model performance. According to Eq. 1, Eq. 7 and Eq. 8, we derive the results for visual semantic variations as follows:

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

For both permutation-variance and permutation-invariance, the value of $\sum_{v_i\in\{V_{MA}-V_{MI}\}} w_{v_i}$ remains constant, denote as C_1. Consequently, the visual semantic variation can be simplified to $\gamma^I = 2\left|C_1 + \sum_{e_j\in\{E_{MA}-E_{MI}\}} w_{e_j}\right|$, primarily depending on the scale of the set $E_{MA} - E_{MI}$. However, the scale of the set varies dramatically between two settings:

- In permutation-variance settings (T_a, T_{pv}), the optimal set of $E_{MA} - E_{MI}$ is its maximum set E, resulting in a positive correlation between visual semantic variation $\gamma^I_{w/}$ and model performance.
- In permutation-invariance settings (T_a, T_{pi}), the optimal set of $E_{MA} - E_{MI}$ is its minimum set $\emptyset$, resulting in a negative correlation between visual semantic variation $\gamma^I_{w/o}$ and model performance.

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

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

B.5 PROPERTY OF SEMVAREFFECT SCORE κ

The SemVarEffect score on visual semantic variations, κ, is the difference between $\gamma^I_{w/}$ and $\gamma^I_{w/o}$. It quantifies the model's ability to discriminate between significant and negligible semantic changes in the text.

- If κ is large, it indicates that the model is sensitive to semantic changes, recognizing variations in meaning. However, it does not necessarily mean the model achieves strong alignment. The model might detect changes in semantics but still struggle to fully capture all objects and relationships described in the text, reflecting a gap between sensitivity and complete alignment.
- If κ is small or close to zero, it indicates that the model either fails to reflect meaningful semantic changes or overreacts to minor text variations. No matter what the overall alignment score is, the model may generate similar images regardless of significant semantic differences in the input text.

C CONSTRUCTION DETAILS

C.1 DATA COLLECTION

Seed Sentence Pairs from Winoground	Templates & Rule
$\boldsymbol{caption\_0}$: a bird eats a snake $\boldsymbol{caption\_1}$: a snake eats a bird	$\boldsymbol{T_a}$: [Noun1] [Verb (vt)] [Noun2] $\boldsymbol{T_{pv}}$: [Noun2] [Verb (vt)] [Noun1] $\boldsymbol{T_a \rightarrow T_{pv}}$: [Noun1] ↔ [Noun2]
$\boldsymbol{caption\_0}$: a person is in a helicopter which is in a car $\boldsymbol{caption\_1}$: a person is in a car which is in a helicopter	$\boldsymbol{T_a}$: [Noun1] [Verb (vi)] [Prepositional Phrase1 (location)] which is in [Prepositional Phrase2 (location)] $\boldsymbol{T_{pv}}$: [Noun1] [Verb (vi)] [Prepositional Phrase2 (location)] which is in [Prepositional Phrase1 (location)] $\boldsymbol{T_a \rightarrow T_{pv}}$: [Prepositional Phrase1 (location)] ↔ [Prepositional Phrase2 (location)]
$\boldsymbol{caption\_0}$: there are some pineapples in boxes, and far more pineapples than boxes $\boldsymbol{caption\_1}$: there are some boxes containing pineapples, and far more boxes than pineapples	$\boldsymbol{T_a}$: ([Prepositional Phrase1 (location)],)(There be)[Noun1] [locate in] [Noun2], and far more [Noun1] than [Noun2] $\boldsymbol{T_{pv}}$: ([Prepositional Phrase1 (location)],)(There be)[Noun2] [contain] [Noun1], and far more [Noun2] than [Noun1] $\boldsymbol{T_a \rightarrow T_{pv}}$: [Noun1] ↔ [Noun2]
$\boldsymbol{caption\_0}$: the person sitting down is supporting the person standing up $\boldsymbol{caption\_1}$: the person standing up is supporting the person sitting down	$\boldsymbol{T_a}$: [Noun1] (which) [Verb1 (vi)] [Verb (vt)] [Noun2] (which) [Verb2 (vi)] $\boldsymbol{T_{pv}}$: [Noun1] (which) [Verb2 (vi)] [Verb (vt)] [Noun2] (which) [Verb1 (vi)] $\boldsymbol{T_a \rightarrow T_{pv}}$: [Verb1 (vi)] ↔ [Verb2 (vi)]
$\boldsymbol{caption\_0}$: the person with green legs is running quite slowly and the red legged one runs faster $\boldsymbol{caption\_1}$: the person with green legs is running faster and the red legged one runs quite slowly	$\boldsymbol{T_a}$: [Noun1] [Prepositional Phrase1/Relative Clause1 (appearance)] [Verb1 (vi)] slowly and [Noun2] [Prepositional Phrase2/Relative Clause2 (appearance)] [Verb2 (vi)] faster $\boldsymbol{T_{pv}}$: [Noun1] [Prepositional Phrase1/Relative Clause1 (appearance)] [Verb1 (vi)] faster and [Noun2] [Prepositional Phrase2/Relative Clause2 (appearance)] [Verb2 (vi)] slowly $\boldsymbol{T_a \rightarrow T_{pv}}$: slowly ↔ faster

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

Template Acquisition We name 171 compositional cases in Winoground Thrush et al. (2022), which are labeled as "no-tag" in subsequent research Diwan et al. (2022), as SEED_0 and SEED_1. The template of T_{pv}, the permutation with semantic changes from, is extracted from each pair of seeds by human. Then, we make the rule of T_{pi}, which is the permutation without semantic changes, as the original template of T_{pi}. An examples are illustrated as following.

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

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

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)] ↔ [Verb2 (vi)]
$T_a \rightarrow T_{pi}$: NULL

Template-guided Generation for T_a. The prompt for generating the T_a guided by the templates and seed pairs is:

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

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

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

Rule-guided Permutation for T_{pv}. The prompt for generating the T_{pv} based on the T_a and its rule is:

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

Step 1: Input some sentences structured by {Template T_a} and {Template T_{pv}}. We call each sentence as "TEXT0".
Step 2: For each "TEXT0", perform the change which is "{RULE of $T_a \rightarrow T_{pv}$}" and keep the other words unchanged as "TEXT1".

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

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

Paraphrasing-guided Permutation for T_{pi}. The prompt for generating the T_{pi} based on the T_a and its rule is:

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

Now, please generate a similar sentence for input prompt given at the end. Provide one sentence for each of the six methods. If a sentence cannot be generated using a particular method, please output "None".
Add the results as a list of JSON objects, containing 6 JSON objects. Each object should include the keys: number, modification method, and sentence.

[Prompt]
"{TEXT0}"

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

C.2 DATA ANNOTATION

Criteria for Valid Samples. The primary challenge in annotation lies in defining the criteria for what constitutes "valid". For T2I synthesis models, we define "valid" input text according to 14 specific criteria. First, we illustrate these criteria through examples of T_a and T_{pv}. Second, for T_{pi}, we require that it must apply one of the six synonymous transformations we have defined in the prompt for generating T_{pi}. Semantically, 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.

- 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.
  - Static Scene or Multiple Exposure Scene: Both sentences should be visually representable through images alone, negating the need for video, audio, or other sensory inputs like touch and smell. Temporal aspects, procedures, and comparisons in test cases must be conveyable within a single image's scope.
  - A Moderate Level of Details: Sentences should maintain a moderate level of detail with similar scales for objects and scenes. Excessive or mismatched scales can result in sentences that are challenging to depict. For example, comparing the quantity of books and magazines *in a library* is less suitable than *on a table*.
  - Quantifiable Comparison: Comparisons in both sentences should be quantifiable, using measures like counts, areas, or volumes. For example, *There are more students in the classroom than words on the blackboard* are difficult to compare quantitatively.
- Discriminative
  - Following Permutation Rules: Generated samples T_{pv} must strictly follow the designated manual template, including word swapping and moving.
  - Distinct Textual Semantics: Two sentences must have distinct textual semantics. Otherwise, the pairs are considered invalid.
  - Visually Distinguishable: Two sentences should be visually distinct, with clear differentiation regarding the visual characteristics of the objects or scenes described. Subtle differences requiring very close observation are not considered distinct visual differences.
- Recognizable
  - Item-Specific Scenes: Scenes in sentences should be identifiable, maintaining key elements for recognition. Otherwise, identification may be challenging. For instance, a sentence describing a *library* where *bookshelves outnumber books* might be unrecognizable, as we typically expect a library to contain many books.
  - Item-Specific Characters: When a sentence depicts a character through associations with specific items, these items or behaviors should remain consistent for easy identification. If not, the character may be hard to recognize. For instance, *chefs* are usually associated with *chef's attire, cooking utensils, and kitchens*.
  - Attire-Based Characters: When a sentence presents characters identifiable by their attire, such as *firefighters, police officers, soldiers, doctors, and nurses*, their clothing should remain consistent for clear recognition. Changes in attire could obscure their identities.
  - Action-Based Characters: When a sentence features characters defined by specific actions or interactions, such as bartenders (mixing drinks), businessmen (negotiating), journalists (interviewing), divers (deep-sea diving), their typical activities should be consistent. Altering distinctive features or placing characters in unusual scenarios may obscure their identities.

Automatic Annotation. We employ machine-human hybrid verification to filter out invalid samples that violate any characteristic. We use LLMs to judge whether each sample violates any of the specific characteristics, labeling them "yes" or "no" and providing confidence scores. The samples whose confidence exceed a threshold 0.8 are removed from the dataset. We initially collected 48K samples, each including 3 sentences. The automatic filtering helped eliminate over 42% of them and finally got a corpus with 27K samples.

Human Annotation. We use 15 annotators and 3 experienced experts to verify samples manually. All annotators have linguistic knowledge and are instructed with detailed annotation principles. Each sample is independently annotated by two annotators. Then an experienced expert goes over the controversial annotations and makes the final decision. After annotation, we randomly sampled 100 samples from valid samples to test the accuracy. 2 experts evaluated that 99% samples are valid. Finally, we got 11,479 valid, non-duplicated samples.

Hard Samples Selection. To effectively evaluate T2I models, it is crucial to select challenging samples rather than simple ones. Initially, we generate images using SOTA models like DALL-E3, flagging those with alignment scores below 0.7. Then we aggregate votes from those models to determine the most representative candidates, selecting those with the highest votes for additional filtering. To ensure diversity, we categorized these samples based on permutation types, as shown in Fig. 5, setting a maximum of 50 samples per category. Finally, 684 samples were included in our benchmark.

C.3 DATA STATISTICS

Category. The samples in SemVarBench are divide into 20 categories based on their permutation types, as illustrated in Fig. 5. Furthermore, these categories are classified into 3 aspects based on the type of triples. These categories are *Relation*, *Attribute Comparison* and *Attribute Value*. Specifically, *Relation* includes 6 categories: *Action, Interaction, Absolute Location, Relative Location, Spatial-Temporal, Direction*. *Attribute Contrast* includes 4 categories: *Size, Height, Weight, Vague Amount*. *Attribute Value* includes 10 categories: *Color, Counting, Texture, Material, Shape, Age, Sentiment, Temperature, Manner*, and *Appearance*.

Scale and Split. The dataset comprises 11,454 valid samples of (T_a, T_{pv}, T_{pi}) after data annotation, totaling 34,362 sentences. It is divided into a training set and a test set. The training set contains 10,806 samples, while the test set consists of 648 hard samples for effective evaluation, as shown in Table 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. Specifically, 82.75% of the permutation involves only 1 category, 14.77% involve 2 categories, and 2.49% involve 3 categories. Thus, the total count of categorized samples surpasses the actual number of samples.

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

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 v1.5 (denoted as SD 1.5), Stable Diffusion v2.1 (denoted as SD 2.1), Stable Diffusion XL v1.0 (denoted as SD XL 1.0), Stable Cascade[2] (denoted as SC), DeepFloyd IF XL[3] (denoted as DeepFloyd), PixArt-alpha XL[4](denoted as PixArt), Kolors, Stable Diffusion 3 [medium][5](denoted as SD 3), FLUX.1

---

[2] https://huggingface.co/stabilityai/stable-cascade-prior;https://huggingface.co/stabilityai/stable-cascade

[3] 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

[4] https://huggingface.co/PixArt-alpha/PixArt-XL-2-1024-MS

[5] https://huggingface.co/stabilityai/stable-diffusion-3-medium

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.

Benchmark	Capability	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$_{2D}$ $Gokhaleet\ al.$ (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	22K	LLM & Human	Train/Test

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

[dev][6] (denoted as FLUX.1), Midjourney V6[7] (denoted as MidJ V6), DALL-E 3, CogView3-Plus[8] (denoted as CogV3-Plus), Ideogram 2[9]. The schedulers in SD 1.5 and SD 2.1 are both set to DPM-Solver++. All other settings are 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. The

---

[6] https://huggingface.co/black-forest-labs/FLUX.1-dev

[7] https://www.midjourney.com/home

[8] https://www.bigmodel.cn/dev/api/image-model/cogview

[9] https://about.ideogram.ai/2.0

version of Gemini 1.5 Pro is `gemini-1.5-pro-001`. The version of Claude 3.5 Sonnet is `claude-3-5-sonnet-20240620`. The version of GPT-4o is `gpt-4o-2024-05-13`. The version of GPT-4 Turbo is `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 following.

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

After receiving outputs from LLMs, we utilize regular expressions to extract scores. In our experiments, the outputs from 4 evaluators mentioned above consistently followed the specified format as defined in the prompt. We also tested Qwen-VL-Chat, Qwen-VL-Plus, Qwen-VL-Max, and LLAVA-1.6, which exhibited poor adherence to the specified format and need complicated extractor. For the purpose of simplifying the evaluation process, we decided to adopt results exclusively from Gemini 1.5 Pro, Claude 3.5 Sonnet, GPT-4o, and GPT-4-Turbo.

D.3 TRAINING SETTING

Training Data Selection. The training set of SemVarBench comprises 10,806 samples. We investigate the improvement from the fine-tuning the T2I model Stable Diffusion XL v1.0. We select the generated images whose alignment scores meets the requirements. These constrains are as follows.

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

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

where C_2 is a threshold.

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

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

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

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

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

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

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

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

In our experiments, we utilized Stable Diffusion XL v1.0 to generate an image for each text prompt within the training set. For the selection of training data, we designated $C_2 = 0.8$ and $C_3 = 0.1$. In the end, we selected 327 samples, which equates to 981 sentences.

Supervised Fine-Tuning (SFT). Each text-image pair (T_i, I_i) is added to the training set. For every sample (T_a, T_{pv}, T_{pi}), this results in three text-image pairs: (T_a, I_a), (T_{pv}, I_{pv}) and (T_{pi}, I_{pi}), resulting in a total of 981 diverse pairs. The selected set of samples are denoted as D_s. The loss function for SFT remains unchanged Kingma et al. (2021); Song et al. (2021), which is

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

where x, y, t, z_t represent the representation of the image I_i, text prompt T_i, timestamp and the latent representation of the image at timestamp t. We conducted two separate fine-tuning processes using the diffusers library[10]: only fine-tuned the LoRA model on UNet or on text encoder for 5000 steps with the training batch size 1.

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 that of 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}), totaling 1,308 tuples. The loss function for DPO remains unchanged **?**, which is

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

where x^w, x^l, y, t, z_t^w, z_t^l, σ represent the representation of the chosen image I_i, the rejected image I_j, text prompt T_i, timestamp, the latent representation of the chosen image at timestamp t, the latent representation of the rejected image at timestamp t and the sigmoid function. We executed two independent fine-tuning processes using the DiffusionDPO[11] and diffusers library: only fine-tuned the LoRA model on UNet or on text encoder for 5000 steps with the training batch size 1.

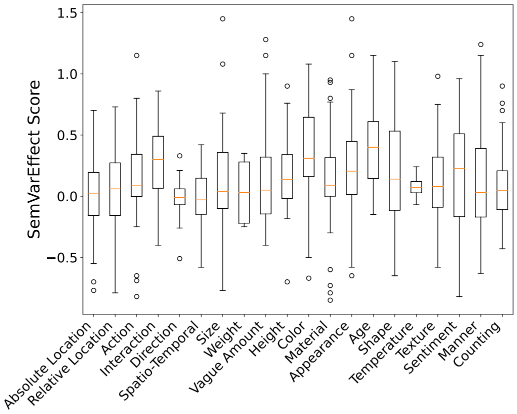


Figure 11: The distribution of SemVarEffect scores across various categories for the Ideogram 2 model, as evaluated by GPT-4 Turbo.

E More Experiment Results

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 the exact scores listed in Tab. 8 and Tab. 9. For *Relation*, most models show consistent performance with low scores, as indicated by the dark blue shading. This suggests that models handle relations like absolute location, relative location, and actions similarly but with limited accuracy. For *Attribute Value*, models like Ideogram2 perform significantly better in capturing attributes such as *Color*, as shown by the prominent red shading in Fig. 7. These models demonstrate a clear advantage in generating

---

[10] https://github.com/huggingface/diffusers/tree/main/examples /text_to_image

[11] https://github.com/SalesforceAIResearch/DiffusionDPO

Models	Relation						Attribute Comparison			
	Absolute Location	Relative Location	Action	Interaction	Direction	Spatial -Temporal	Size	Weight	Vague Amount	Height
Open Source Models										
Stable Diffusion v1.5	-0.01	0.00	-0.06	-0.11	0.05	-0.01	0.11	-0.01	-0.07	0.01
Stable Diffusion v2.1	-0.01	-0.06	-0.08	0.02	-0.02	-0.00	0.03	-0.10	0.02	0.06
Stable Diffusion XL v1.0	-0.02	-0.08	-0.01	0.05	0.05	-0.05	0.07	0.16	0.03	0.09
Stable Cascade	0.02	-0.03	0.01	-0.03	0.02	0.02	-0.02	-0.09	0.08	-0.01
DeepFloyd IF XL	-0.01	-0.00	0.01	-0.01	0.04	0.01	0.03	0.05	-0.04	0.03
PixArt-alpha XL	0.00	-0.01	0.03	0.00	-0.04	-0.03	0.07	0.10	0.10	0.03
Kolors	-0.03	0.02	-0.07	0.02	0.03	-0.02	-0.06	-0.10	0.07	0.07
Stable Diffusion 3	-0.03	0.01	-0.02	0.05	-0.08	-0.04	0.07	-0.02	0.10	0.04
FlUX.1	-0.03	0.03	0.03	**0.08**	-0.04	-0.00	0.09	**0.23**	0.05	0.09
API-based Models										
Midjourney V6	0.07	0.01	0.04	0.03	0.03	0.08	0.07	-0.12	0.07	0.02
DALL-E 3	-0.00	**0.12**	0.11	0.08	**0.13**	**0.11**	0.08	-0.00	0.09	0.15
CogView3-Plus	**0.08**	0.08	**0.23**	0.07	0.03	-0.01	**0.23**	-0.03	**0.23**	**0.22**
Ideogram 2	0.01	0.04	0.13	**0.29**	-0.02	-0.02	0.12	0.04	0.17	0.17

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

Models	Attribute Value										AVG
	Color	Material	Appearance	Age	Shape	Temperature	Texture	Sentiment	Manner	Counting	
Open Source Models											
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
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
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
Stable Cascade	0.14	0.05	0.10	-0.03	-0.02	-0.15	-0.03	-0.05	0.06	0.01	0.04
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
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
Kolors	0.21	0.07	-0.01	0.01	-0.09	0.01	0.10	-0.01	-0.04	0.00	0.01
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
FlUX.1	0.35	**0.21**	0.21	0.08	0.09	-0.13	0.04	-0.06	0.07	**0.10**	0.09
API-based Models											
Midjourney V6	0.20	0.12	0.13	0.10	-0.03	-0.21	0.11	-0.05	0.09	-0.02	0.05
DALL-E 3	0.22	0.17	0.17	0.23	0.14	0.22	**0.22**	0.19	0.11	0.06	0.12
CogView3-Plus	0.35	0.15	0.17	0.31	0.16	**0.30**	0.17	**0.21**	**0.27**	0.07	**0.15**
Ideogram 2	**0.37**	0.15	**0.24**	**0.42**	**0.20**	0.08	0.12	0.16	0.13	0.07	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.

or recognizing these attributes. Conversely, models like DALL-E 3 and CogV3-Plus display more balanced but 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*, show slightly better performance, reflected by higher median values. The Fig. 11 illustrates the distribution of SemVarEffect scores across various categories for the Ideogram 2 model, while the Fig. 12 shows the scores for different T2I models in the 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 responses, while categories such as *Sentiment* and *Texture* show more consistent effects with narrower distributions.

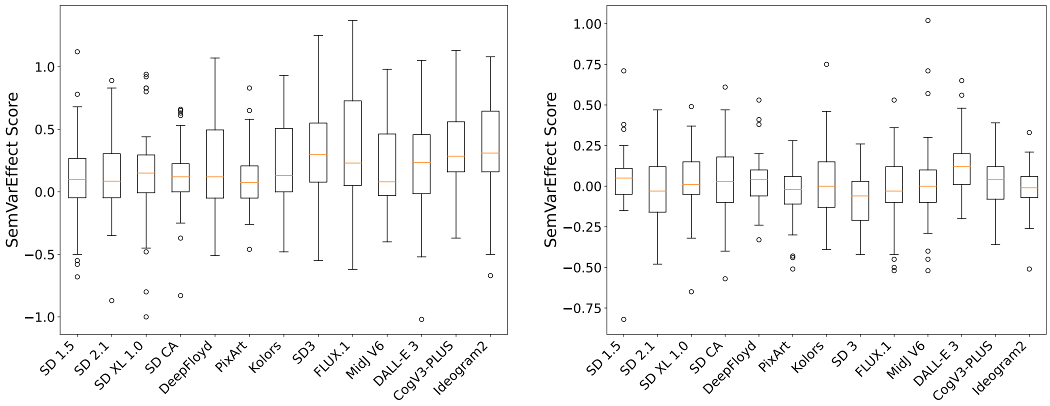


Figure 12: The distribution of SemVarEffect scores across various T2I models within the Color and Direction categories, as evaluated by GPT-4 Turbo. The top box plot is the distribution within the Color category. The bottom box plot is the distribution within the Direction category.

F MORE ANALYSIS

Is there a significant difference among various text encoders in discerning semantic nuances within language 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 CLIP as the text encoder, including SD 1.5, SD 2.1, SD XL v1.0, and SC, as well as those using T5, such as DeepFloyd and PixArt. The text encoders of Stable Diffusion v1.5, v2.1, XL v1.0, and Stable Cascade are one or two CLIP series models. The text encoders of DeepFloyd and PixArt are T5 series models. The figure depicts the similarity metric as $1 - \text{cosine}(T_a, T_{pv})$, with higher values signifying a more robust capacity 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.

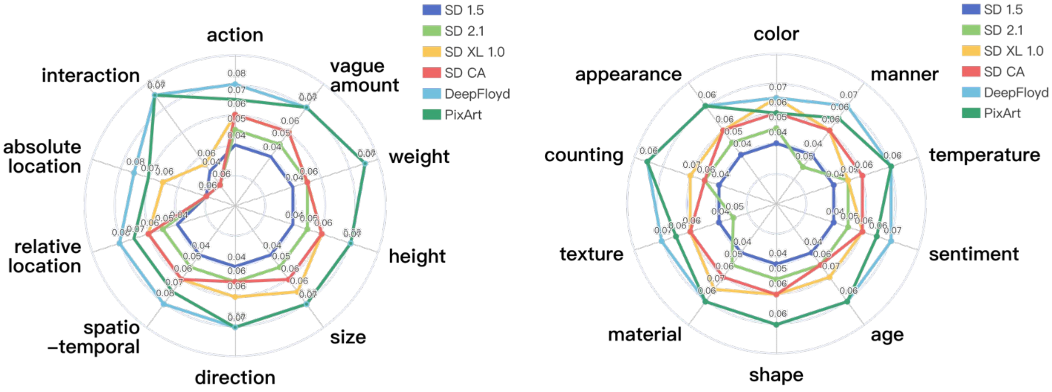


Figure 13: The semantic discrimination capabilities of different text encoders measured by $1 - \text{cosine}(T_a, T_{pv})$.

More Analysis for alignment scores vs. effect score. Fig. 14 illustrates that although the distribution of the language effect score and the alignment score are similar, the language effect 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 language effect score, it becomes evident that FLUX.1 and SD 3 are distinctly different from SD XL 1.0. SD XL 1.0 responds more similarly to semantic variations caused by word order changes as seen in 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 is more pronounced when assessed by the language effect score.

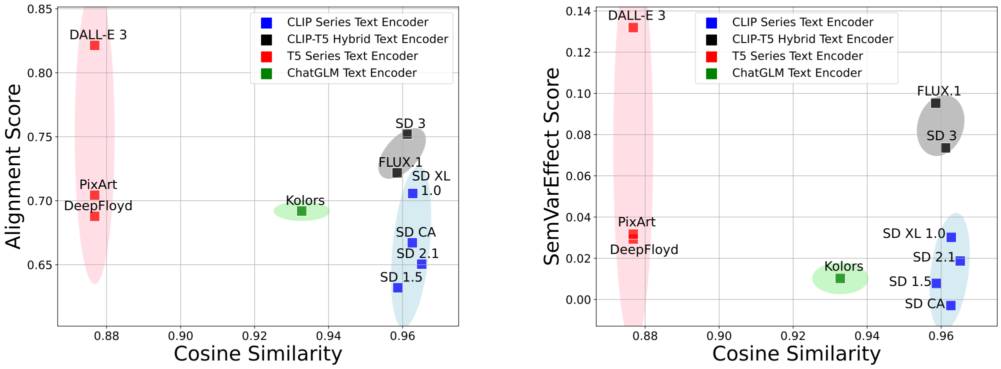


Figure 14: A comparison of alignment scores and the SemVarEffect score under the same conditions of text similarity. The squares are results of permutations of permutation-variance. The evaluator is GPT-4 Turbo.

Why do permutations without semantic changes exhibit higher text similarity scores than those with semantic changes? This phenomenon is closely tied 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 investigated that permutations with semantic changes in our benchmark show smaller edit distances from the anchor sentence compared to synonymous sentences. The average edit distances between (T_a, T_{pv}), (T_a, T_{pi}) and (T_a, T_{random}) are 13, 32 and 53. Since our analysis does not rely on similarity scores of synonymous sentences, this does not affect our previous findings.

G More Case Studies

In this section, we present examples that demonstrate an understanding of semantic variations and those that do not. Examples that grasp semantic variations typically feature high alignment scores, $\bar{S}_{ii}$, and high effect scores, κ, as illustrated in Fig. 15. Conversely, examples lacking this understanding often have high alignment scores, $\bar{S}_{ii}$, but low effect scores, κ, as depicted in Figures 16 and 17. We can distinguish the models' abilities in accurately interpreting and visually representing semantic variations based on the SemVarEffect scores. However, in practice, the evaluation accuracy can be significantly compromised by errors in generated images or evaluators' ratings. Severe errors can particularly distort the evaluation's accuracy, as evidenced in Figures 21 and 23. 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 less satisfied accuracy of LLM's validation, which incurs high costs. Furthermore, the scale of evaluation is also limited by the high costs associated with image generation and evaluation using LLMs, both in terms of time and financial expenditure, thus restricting the extent of such evaluations.

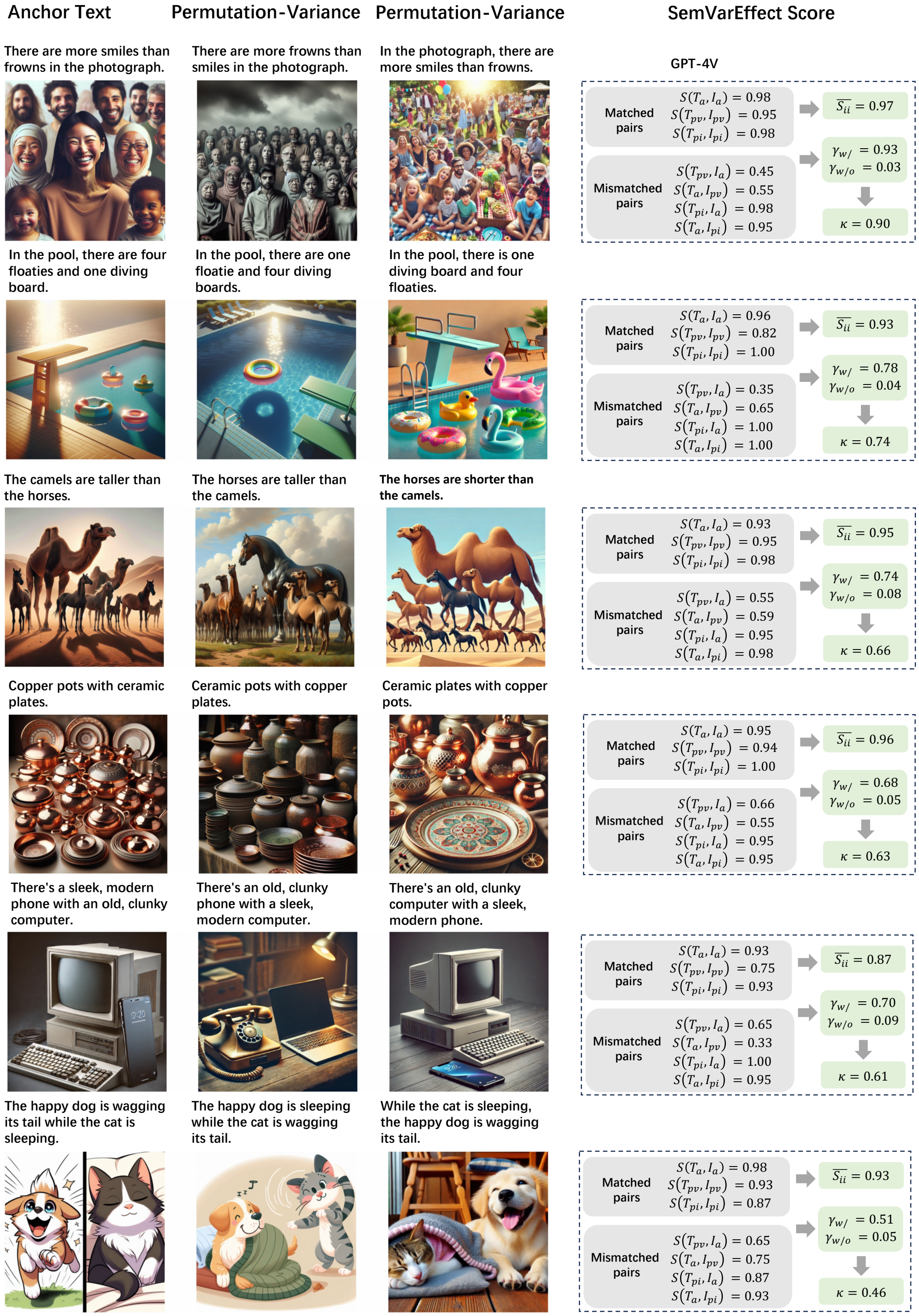
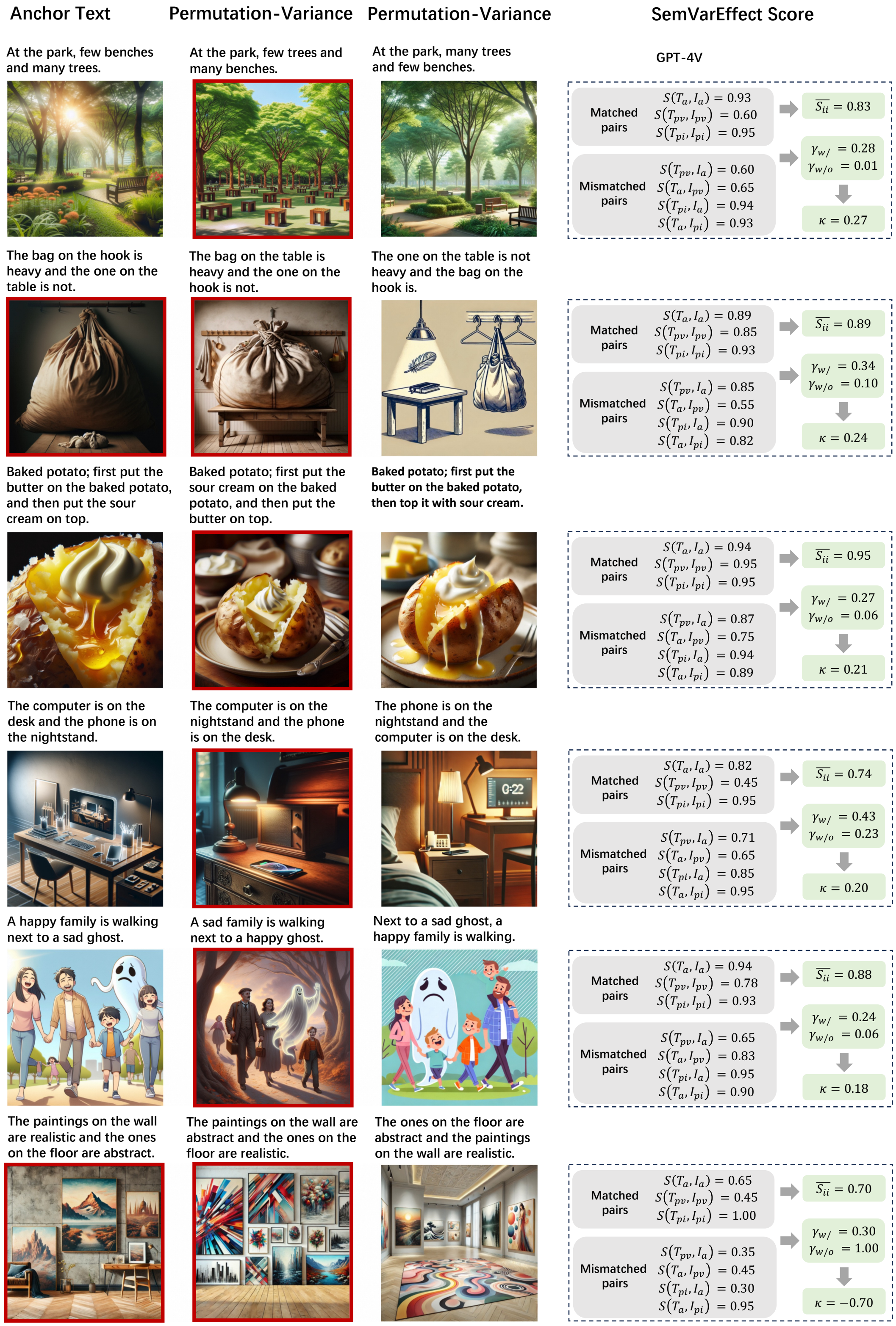

Anchor Text
Permutation-Variance
Permutation-Variance
SemVarEffect Score
There are more smiles than frowns in the photograph.
There are more frowns than smiles in the photograph.
In the photograph, there are more smiles than frowns.
GPT-4V
Matched pairs
$S(T_a, I_a) = 0.98$
$S(T_{pv}, I_{pv}) = 0.95$
$S(T_{pi}, I_{pi}) = 0.98$
Mismatched pairs
$S(T_{pv}, I_a) = 0.45$
$S(T_a, I_{pv}) = 0.55$
$S(T_{pi}, I_a) = 0.98$
$S(T_a, I_{pi}) = 0.95$
$\overline{S_{ii}} = 0.97$
$\gamma_{w/} = 0.93$
$\gamma_{w/o} = 0.03$
$\kappa = 0.90$
In the pool, there are four floaties and one diving board.
In the pool, there are one floatie and four diving boards.
In the pool, there is one diving board and four floaties.
Matched pairs
$S(T_a, I_a) = 0.96$
$S(T_{pv}, I_{pv}) = 0.82$
$S(T_{pi}, I_{pi}) = 1.00$
Mismatched pairs
$S(T_{pv}, I_a) = 0.35$
$S(T_a, I_{pv}) = 0.65$
$S(T_{pi}, I_a) = 1.00$
$S(T_a, I_{pi}) = 1.00$
$\overline{S_{ii}} = 0.93$
$\gamma_{w/} = 0.78$
$\gamma_{w/o} = 0.04$
$\kappa = 0.74$
The camels are taller than the horses.
The horses are taller than the camels.
The horses are shorter than the camels.
Matched pairs
$S(T_a, I_a) = 0.93$
$S(T_{pv}, I_{pv}) = 0.95$
$S(T_{pi}, I_{pi}) = 0.98$
Mismatched pairs
$S(T_{pv}, I_a) = 0.55$
$S(T_a, I_{pv}) = 0.59$
$S(T_{pi}, I_a) = 0.95$
$S(T_a, I_{pi}) = 0.98$
$\overline{S_{ii}} = 0.95$
$\gamma_{w/} = 0.74$
$\gamma_{w/o} = 0.08$
$\kappa = 0.66$
Copper pots with ceramic plates.
Ceramic pots with copper plates.
Ceramic plates with copper pots.
Matched pairs
$S(T_a, I_a) = 0.95$
$S(T_{pv}, I_{pv}) = 0.94$
$S(T_{pi}, I_{pi}) = 1.00$
Mismatched pairs
$S(T_{pv}, I_a) = 0.66$
$S(T_a, I_{pv}) = 0.55$
$S(T_{pi}, I_a) = 0.95$
$S(T_a, I_{pi}) = 0.95$
$\overline{S_{ii}} = 0.96$
$\gamma_{w/} = 0.68$
$\gamma_{w/o} = 0.05$
$\kappa = 0.63$
There's a sleek, modern phone with an old, clunky computer.
There's an old, clunky phone with a sleek, modern computer.
There's an old, clunky computer with a sleek, modern phone.
Matched pairs
$S(T_a, I_a) = 0.93$
$S(T_{pv}, I_{pv}) = 0.75$
$S(T_{pi}, I_{pi}) = 0.93$
Mismatched pairs
$S(T_{pv}, I_a) = 0.65$
$S(T_a, I_{pv}) = 0.33$
$S(T_{pi}, I_a) = 1.00$
$S(T_a, I_{pi}) = 0.95$
$\overline{S_{ii}} = 0.87$
$\gamma_{w/} = 0.70$
$\gamma_{w/o} = 0.09$
$\kappa = 0.61$
The happy dog is wagging its tail while the cat is sleeping.
The happy dog is sleeping while the cat is wagging its tail.
While the cat is sleeping, the happy dog is wagging its tail.
Matched pairs
$S(T_a, I_a) = 0.98$
$S(T_{pv}, I_{pv}) = 0.93$
$S(T_{pi}, I_{pi}) = 0.87$
Mismatched pairs
$S(T_{pv}, I_a) = 0.65$
$S(T_a, I_{pv}) = 0.75$
$S(T_{pi}, I_a) = 0.87$
$S(T_a, I_{pi}) = 0.93$
$\overline{S_{ii}} = 0.93$
$\gamma_{w/} = 0.51$
$\gamma_{w/o} = 0.05$
$\kappa = 0.46$

Figure 15: The cases which understand semantic variations.

Aspect	Category	Example
Relation	Action	T_a: A dog sits and a cat stands. T_{pv}: A dog stands and a cat sits. T_{pi}: A cat stands and a dog sits.
	Interaction	T_a: An old person kisses a young person. T_{pv}: A young person kisses an old person. T_{pi}: A young person is kissed by an old person.
	Absolute Location	T_a: The soft teddy bear is on the bed and the hard toy car is on the shelf. T_{pv}: The soft teddy bear is on the shelf and the hard toy car is on the bed. T_{pi}: The hard toy car is on the shelf and the soft teddy bear is on the bed.
	Relative Location	T_a: A green apple sits atop a red leaf. T_{pv}: A red leaf sits atop a green apple. T_{pi}: A red leaf sits below a green apple.
	Spatial-Temporal	T_a: Sushi roll; first put the fish on the seaweed, and then put the rice on top. T_{pv}: Sushi roll; first put the rice on the seaweed, and then put the fish on top. T_{pi}: Sushi roll; first apply the fish on the seaweed, and then place the rice on top.
	Direction	T_a: A boy jumps away from the fence and towards the river. T_{pv}: A boy jumps away from the river and towards the fence. T_{pi}: A boy towards the river and jumps away from the fence.
Attribute Comparison	Size	T_a: The cake and the plate; the cake is too big for the plate. T_{pv}: The cake and the plate; the plate is too big for the cake. T_{pi}: The plate and the cake; the place is too small for the cake.
	Height	T_a: A dinosaur towering over a human. T_{pv}: A human towering over a dinosaur. T_{pi}: A human being towered over by a dinosaur.
	Weight	T_a: The athlete with a heavy backpack is walking quite slowly and the one with a light bag is running faster. T_{pv}: The athlete with a light backpack is walking quite slowly and the one with a heavy bag is running faster. T_{pi}: The athlete with a light bag is running faster and the one with a heavy backpack is walking quite slowly.
	Vague Amount	T_a: A cake with more frosting on the top than on the slides. T_{pv}: A cake with more frosting on the slides than on the top. T_{pi}: A cake with less frosting on the slides than on the top.
Attribute Values	Color	T_a: A man in a purple shirt is carrying a brown suitcase. T_{pv}: A man in a brown shirt is carrying a purple suitcase. T_{pi}: A brown suitcase is being carried by a man in a purple shirt.
	Counting	T_a: Four dogs in a doghouse and one dog barking outside. T_{pv}: One dogs in a doghouse and four dog barking outside. T_{pi}: One dog barking outside and four dogs in a doghouse.
	Texture	T_a: Two fish; the one in the tank has stripes and the one in the bowl doesn't. T_{pv}: Two fish; the one in the bowl has stripes and the one in the tank doesn't. T_{pi}: Two fish; the one in the bowl has no stripes and the one in the tank does.
	Material	T_a: There's a satin teddy bear with a furry bow. T_{pv}: There's a furry teddy bear with a satin bow. T_{pi}: A satin teddy bear has a furry bow.
	Shape	T_a: The circular suitcase has an oblong lock. T_{pv}: The oblong suitcase has an circular lock. T_{pi}: An oblong lock is on the circular suitcase.
	Age	T_a: The person on the left is old and the person on the right is young. T_{pv}: The person on the right is old and the person on the left is young. T_{pi}: The person on the right is young and the person on the left is old.
	Sentiment	T_a: The happy child is playing next to a sad clown. T_{pv}: The sad child is playing next to a happy clown. T_{pi}: Next to a sad clown, a happy child is playing.
	Temperature	T_a: Iced coffee and steaming tea. T_{pv}: Steaming coffee and iced tea. T_{pi}: Steaming tea and iced coffee.
	Manner	T_a: The building on the corner has a modern design and the monument in the park has a classic design. T_{pv}: The building on the corner has a classic design and the monument in the park has a modern design. T_{pi}: The monument in the park has a classic design and the building on the corner has a modern design.
	Appearance	T_a: The boy with a blue shirt has long hair and the girl in the pink dress has short hair. T_{pv}: The boy with a blue shirt has short hair and the girl in the pink dress has long hair. T_{pi}: The girl in the pink dress has short hair and the boy with a blue shirt has long hair.

Table 10: Permutation-based valid sentences (T_a, T_{pv}, T_{pi}) in diverse categories.



At the park, few benches and many trees.

At the park, few trees and many benches.

At the park, many trees and few benches.

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.

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.

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.

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.

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.

Figure 16: The cases which don’t understand semantic variations.

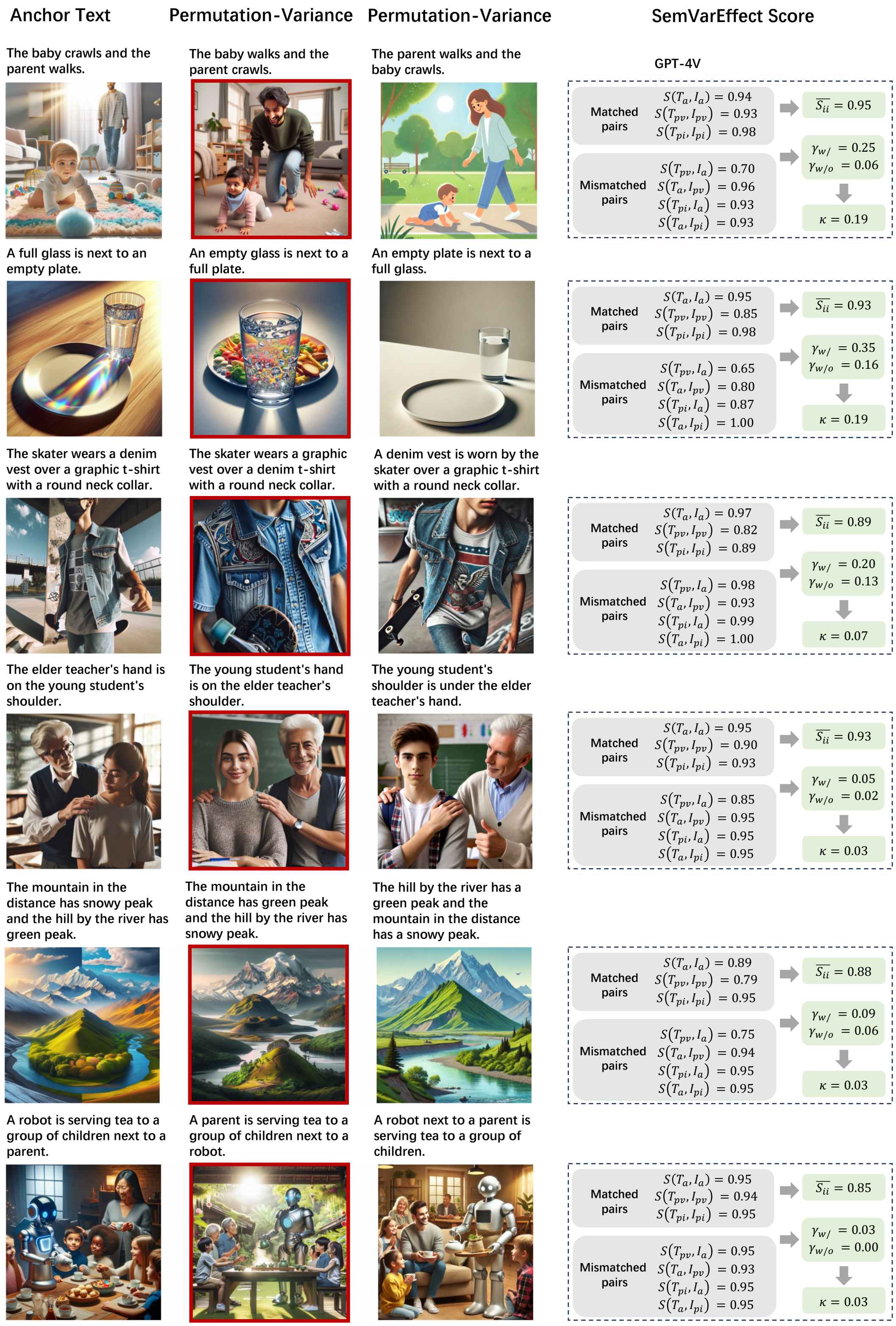
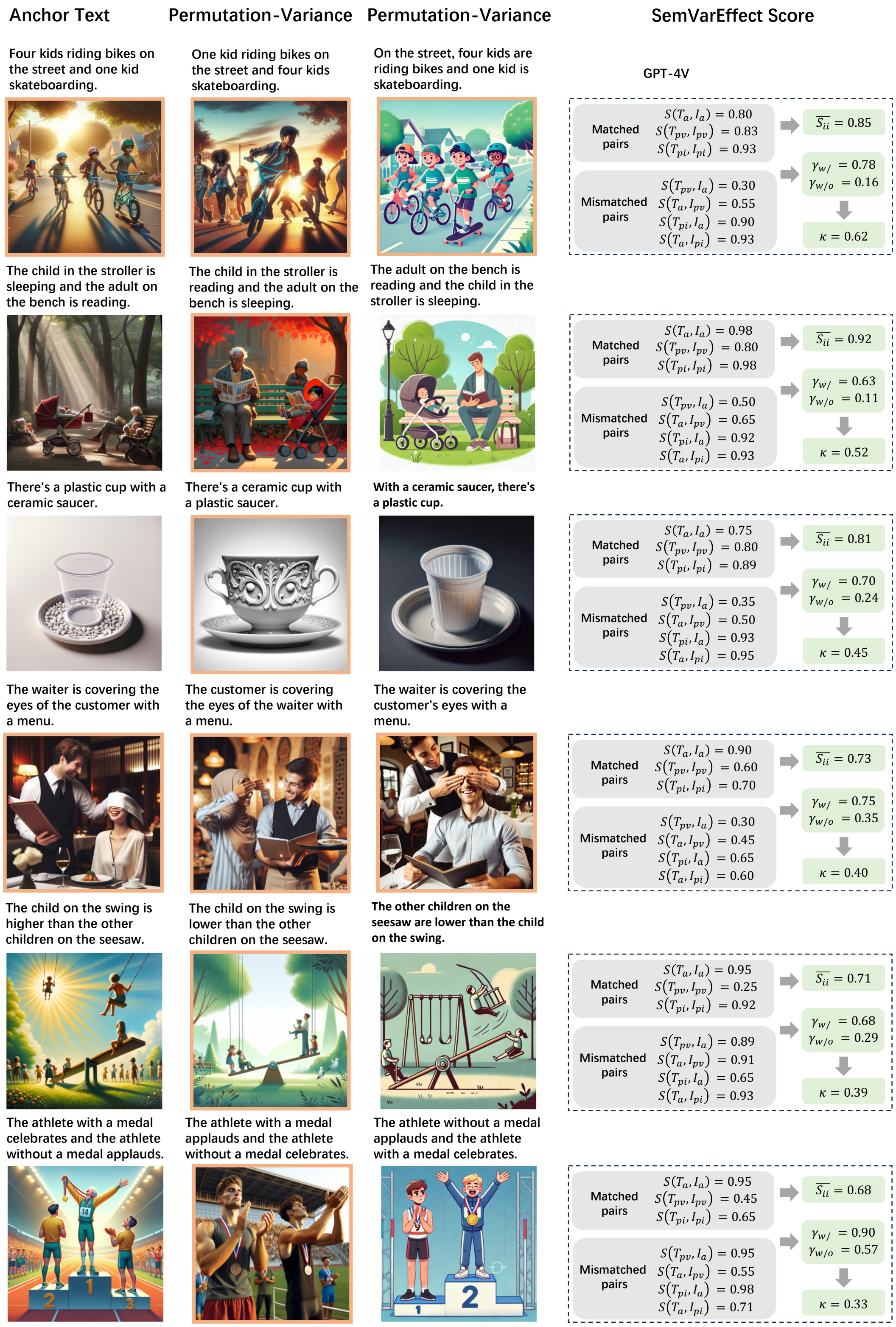

Anchor Text
Permutation-Variance
Permutation-Variance
SemVarEffect Score
GPT-4V
The baby crawls and the parent walks.
The baby walks and the parent crawls.
The parent walks and the baby crawls.
Matched pairs
$S(T_a, I_a) = 0.94$
$S(T_{pv}, I_{pv}) = 0.93$
$S(T_{pi}, I_{pi}) = 0.98$
Mismatched pairs
$S(T_{pv}, I_a) = 0.70$
$S(T_a, I_{pv}) = 0.96$
$S(T_{pi}, I_a) = 0.93$
$S(T_a, I_{pi}) = 0.93$
$\overline{S_{ii}} = 0.95$
$\gamma_{w/} = 0.25$
$\gamma_{w/o} = 0.06$
$\kappa = 0.19$
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.
Matched pairs
$S(T_a, I_a) = 0.95$
$S(T_{pv}, I_{pv}) = 0.85$
$S(T_{pi}, I_{pi}) = 0.98$
Mismatched pairs
$S(T_{pv}, I_a) = 0.65$
$S(T_a, I_{pv}) = 0.80$
$S(T_{pi}, I_a) = 0.87$
$S(T_a, I_{pi}) = 1.00$
$\overline{S_{ii}} = 0.93$
$\gamma_{w/} = 0.35$
$\gamma_{w/o} = 0.16$
$\kappa = 0.19$
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.
Matched pairs
$S(T_a, I_a) = 0.97$
$S(T_{pv}, I_{pv}) = 0.82$
$S(T_{pi}, I_{pi}) = 0.89$
Mismatched pairs
$S(T_{pv}, I_a) = 0.98$
$S(T_a, I_{pv}) = 0.93$
$S(T_{pi}, I_a) = 0.99$
$S(T_a, I_{pi}) = 1.00$
$\overline{S_{ii}} = 0.89$
$\gamma_{w/} = 0.20$
$\gamma_{w/o} = 0.13$
$\kappa = 0.07$
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.
Matched pairs
$S(T_a, I_a) = 0.95$
$S(T_{pv}, I_{pv}) = 0.90$
$S(T_{pi}, I_{pi}) = 0.93$
Mismatched pairs
$S(T_{pv}, I_a) = 0.85$
$S(T_a, I_{pv}) = 0.95$
$S(T_{pi}, I_a) = 0.95$
$S(T_a, I_{pi}) = 0.95$
$\overline{S_{ii}} = 0.93$
$\gamma_{w/} = 0.05$
$\gamma_{w/o} = 0.02$
$\kappa = 0.03$
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.
Matched pairs
$S(T_a, I_a) = 0.89$
$S(T_{pv}, I_{pv}) = 0.79$
$S(T_{pi}, I_{pi}) = 0.95$
Mismatched pairs
$S(T_{pv}, I_a) = 0.75$
$S(T_a, I_{pv}) = 0.94$
$S(T_{pi}, I_a) = 0.95$
$S(T_a, I_{pi}) = 0.95$
$\overline{S_{ii}} = 0.88$
$\gamma_{w/} = 0.09$
$\gamma_{w/o} = 0.06$
$\kappa = 0.03$
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.
Matched pairs
$S(T_a, I_a) = 0.95$
$S(T_{pv}, I_{pv}) = 0.94$
$S(T_{pi}, I_{pi}) = 0.95$
Mismatched pairs
$S(T_{pv}, I_a) = 0.95$
$S(T_a, I_{pv}) = 0.93$
$S(T_{pi}, I_a) = 0.95$
$S(T_a, I_{pi}) = 0.95$
$\overline{S_{ii}} = 0.85$
$\gamma_{w/} = 0.03$
$\gamma_{w/o} = 0.00$
$\kappa = 0.03$

Figure 17: More cases which don't understand semantic variations.



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.

Matched pairs: $S(T_a, I_a) = 0.80$, $S(T_{pv}, I_{pv}) = 0.83$, $S(T_{pi}, I_{pi}) = 0.93$ → $\overline{S_{ii}} = 0.85$

Mismatched pairs: $S(T_{pv}, I_a) = 0.30$, $S(T_a, I_{pv}) = 0.55$, $S(T_{pi}, I_a) = 0.90$, $S(T_a, I_{pi}) = 0.93$ → $\gamma_{w/} = 0.78$, $\gamma_{w/o} = 0.16$ → $\kappa = 0.62$

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.

Matched pairs: $S(T_a, I_a) = 0.98$, $S(T_{pv}, I_{pv}) = 0.80$, $S(T_{pi}, I_{pi}) = 0.98$ → $\overline{S_{ii}} = 0.92$

Mismatched pairs: $S(T_{pv}, I_a) = 0.50$, $S(T_a, I_{pv}) = 0.65$, $S(T_{pi}, I_a) = 0.92$, $S(T_a, I_{pi}) = 0.93$ → $\gamma_{w/} = 0.63$, $\gamma_{w/o} = 0.11$ → $\kappa = 0.52$

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.

Matched pairs: $S(T_a, I_a) = 0.75$, $S(T_{pv}, I_{pv}) = 0.80$, $S(T_{pi}, I_{pi}) = 0.89$ → $\overline{S_{ii}} = 0.81$

Mismatched pairs: $S(T_{pv}, I_a) = 0.35$, $S(T_a, I_{pv}) = 0.50$, $S(T_{pi}, I_a) = 0.93$, $S(T_a, I_{pi}) = 0.95$ → $\gamma_{w/} = 0.70$, $\gamma_{w/o} = 0.24$ → $\kappa = 0.45$

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.

Matched pairs: $S(T_a, I_a) = 0.90$, $S(T_{pv}, I_{pv}) = 0.60$, $S(T_{pi}, I_{pi}) = 0.70$ → $\overline{S_{ii}} = 0.73$

Mismatched pairs: $S(T_{pv}, I_a) = 0.30$, $S(T_a, I_{pv}) = 0.45$, $S(T_{pi}, I_a) = 0.65$, $S(T_a, I_{pi}) = 0.60$ → $\gamma_{w/} = 0.75$, $\gamma_{w/o} = 0.35$ → $\kappa = 0.40$

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.

Matched pairs: $S(T_a, I_a) = 0.95$, $S(T_{pv}, I_{pv}) = 0.25$, $S(T_{pi}, I_{pi}) = 0.92$ → $\overline{S_{ii}} = 0.71$

Mismatched pairs: $S(T_{pv}, I_a) = 0.89$, $S(T_a, I_{pv}) = 0.91$, $S(T_{pi}, I_a) = 0.65$, $S(T_a, I_{pi}) = 0.93$ → $\gamma_{w/} = 0.68$, $\gamma_{w/o} = 0.29$ → $\kappa = 0.39$

The athlete with a medal celebrates and the athlete without a medal applauds.

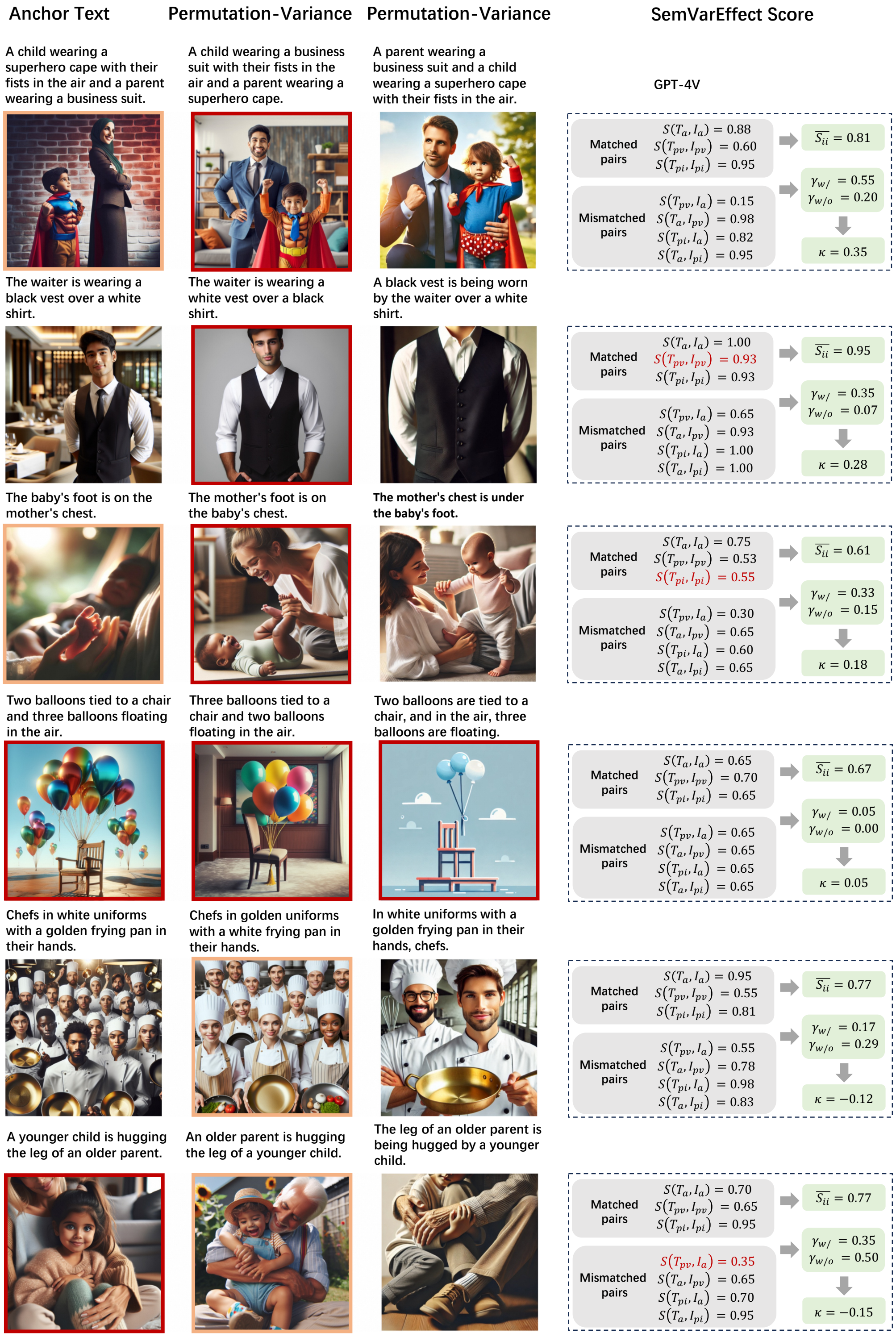
The athlete with a medal applauds and the athlete without a medal celebrates.

The athlete without a medal applauds and the athlete with a medal celebrates.

Matched pairs: $S(T_a, I_a) = 0.95$, $S(T_{pv}, I_{pv}) = 0.45$, $S(T_{pi}, I_{pi}) = 0.65$ → $\overline{S_{ii}} = 0.68$

Mismatched pairs: $S(T_{pv}, I_a) = 0.95$, $S(T_a, I_{pv}) = 0.55$, $S(T_{pi}, I_a) = 0.98$, $S(T_a, I_{pi}) = 0.71$ → $\gamma_{w/} = 0.90$, $\gamma_{w/o} = 0.57$ → $\kappa = 0.33$

Figure 18: Cases with minor errors which understand semantic variations.



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.

$S(T_a, I_a) = 0.88$, $S(T_{pv}, I_{pv}) = 0.60$, $S(T_{pi}, I_{pi}) = 0.95$
$S(T_{pv}, I_a) = 0.15$, $S(T_a, I_{pv}) = 0.98$, $S(T_{pi}, I_a) = 0.82$, $S(T_a, I_{pi}) = 0.95$
$\overline{S_{ii}} = 0.81$, $\gamma_{w/} = 0.55$, $\gamma_{w/o} = 0.20$, $\kappa = 0.35$

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.

$S(T_a, I_a) = 1.00$, $S(T_{pv}, I_{pv}) = 0.93$, $S(T_{pi}, I_{pi}) = 0.93$
$S(T_{pv}, I_a) = 0.65$, $S(T_a, I_{pv}) = 0.93$, $S(T_{pi}, I_a) = 1.00$, $S(T_a, I_{pi}) = 1.00$
$\overline{S_{ii}} = 0.95$, $\gamma_{w/} = 0.35$, $\gamma_{w/o} = 0.07$, $\kappa = 0.28$

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.

$S(T_a, I_a) = 0.75$, $S(T_{pv}, I_{pv}) = 0.53$, $S(T_{pi}, I_{pi}) = 0.55$
$S(T_{pv}, I_a) = 0.30$, $S(T_a, I_{pv}) = 0.65$, $S(T_{pi}, I_a) = 0.60$, $S(T_a, I_{pi}) = 0.65$
$\overline{S_{ii}} = 0.61$, $\gamma_{w/} = 0.33$, $\gamma_{w/o} = 0.15$, $\kappa = 0.18$

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.

$S(T_a, I_a) = 0.65$, $S(T_{pv}, I_{pv}) = 0.70$, $S(T_{pi}, I_{pi}) = 0.65$
$S(T_{pv}, I_a) = 0.65$, $S(T_a, I_{pv}) = 0.65$, $S(T_{pi}, I_a) = 0.65$, $S(T_a, I_{pi}) = 0.65$
$\overline{S_{ii}} = 0.67$, $\gamma_{w/} = 0.05$, $\gamma_{w/o} = 0.00$, $\kappa = 0.05$

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.

$S(T_a, I_a) = 0.95$, $S(T_{pv}, I_{pv}) = 0.55$, $S(T_{pi}, I_{pi}) = 0.81$
$S(T_{pv}, I_a) = 0.55$, $S(T_a, I_{pv}) = 0.78$, $S(T_{pi}, I_a) = 0.98$, $S(T_a, I_{pi}) = 0.83$
$\overline{S_{ii}} = 0.77$, $\gamma_{w/} = 0.17$, $\gamma_{w/o} = 0.29$, $\kappa = -0.12$

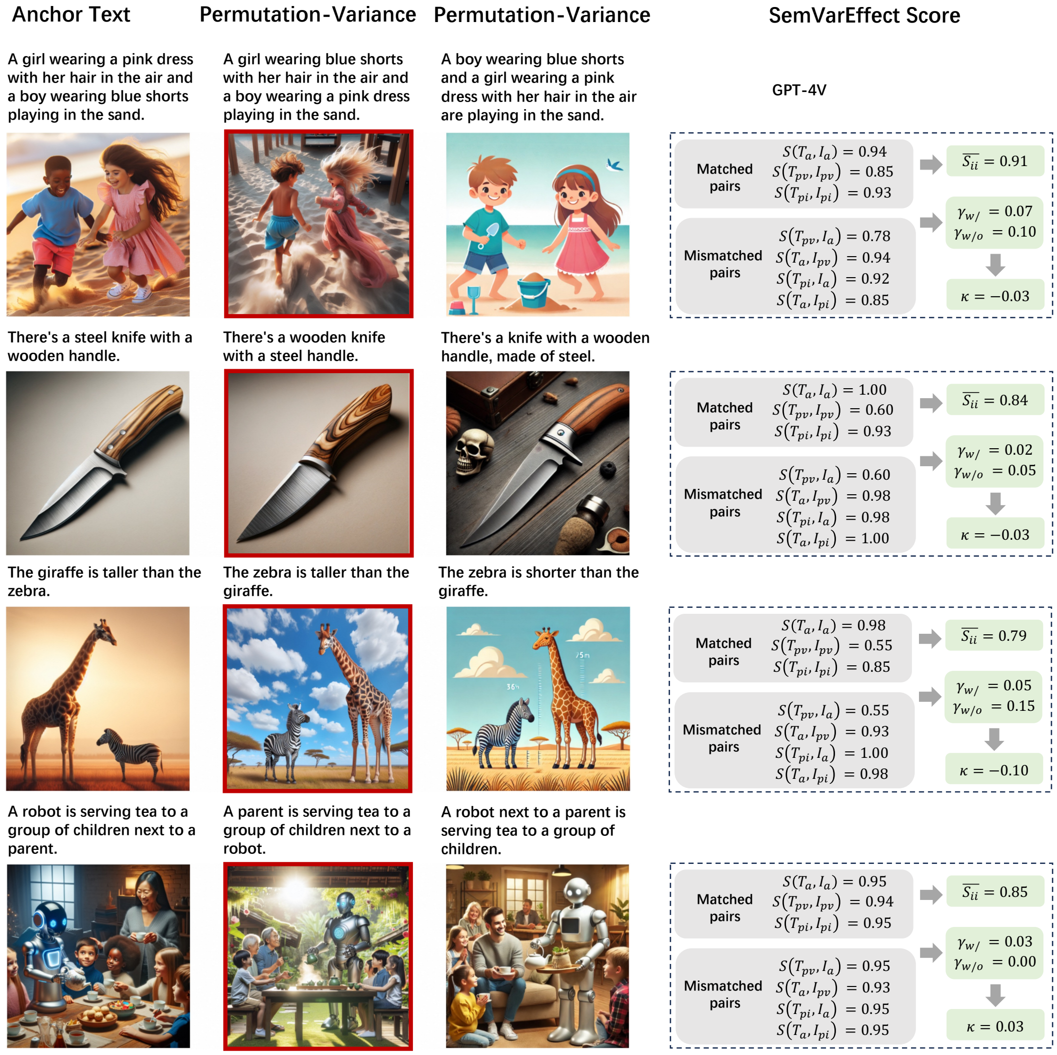
A younger child is hugging the leg of an older parent.

An older parent is hugging the leg of a younger child.

The leg of an older parent is being hugged by a younger child.

$S(T_a, I_a) = 0.70$, $S(T_{pv}, I_{pv}) = 0.65$, $S(T_{pi}, I_{pi}) = 0.95$
$S(T_{pv}, I_a) = 0.35$, $S(T_a, I_{pv}) = 0.65$, $S(T_{pi}, I_a) = 0.70$, $S(T_a, I_{pi}) = 0.95$
$\overline{S_{ii}} = 0.77$, $\gamma_{w/} = 0.35$, $\gamma_{w/o} = 0.50$, $\kappa = -0.15$

Figure 19: Cases with minor errors which don't understand semantic variations. Several alignment scores, which are incorrect according to GPT-4V, are labeled in red.



A girl wearing a pink dress with her hair in the air and a boy wearing blue shorts playing in the sand.

A girl wearing blue shorts with her hair in the air and a boy wearing a pink dress playing in the sand.

A boy wearing blue shorts and a girl wearing a pink dress with her hair in the air are playing in the sand.

There's a steel knife with a wooden handle.

There's a wooden knife with a steel handle.

There's a knife with a wooden handle, made of steel.

The giraffe is taller than the zebra.

The zebra is taller than the giraffe.

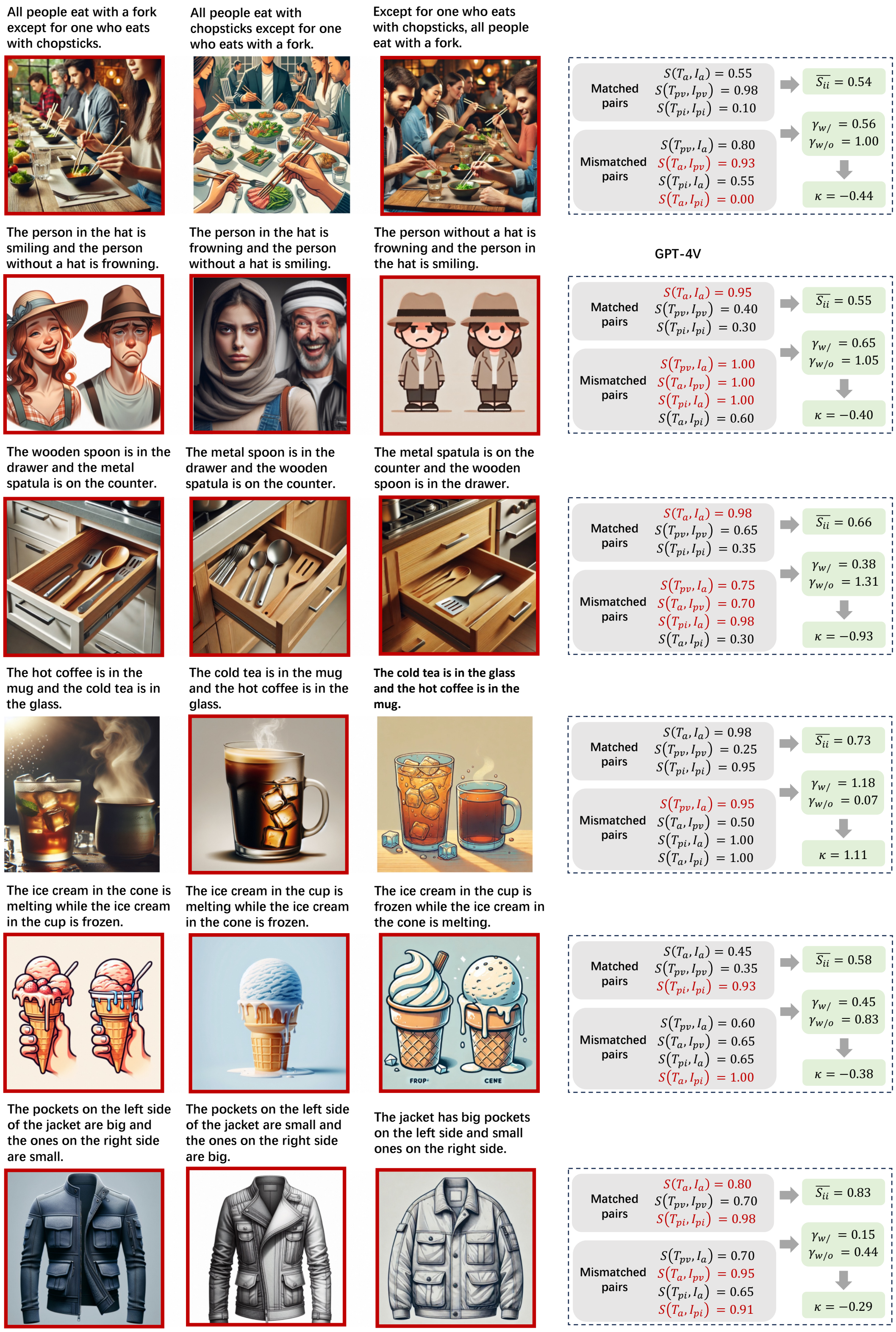
The zebra is shorter than the giraffe.

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: Examples of acceptable outliers include negative SemVarEffect (κ) values that are close to zero. Outliers with a SemVarEffect score (κ) slightly below 0 are acceptable.


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.

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.

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.

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.

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.

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.

Figure 21: 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.

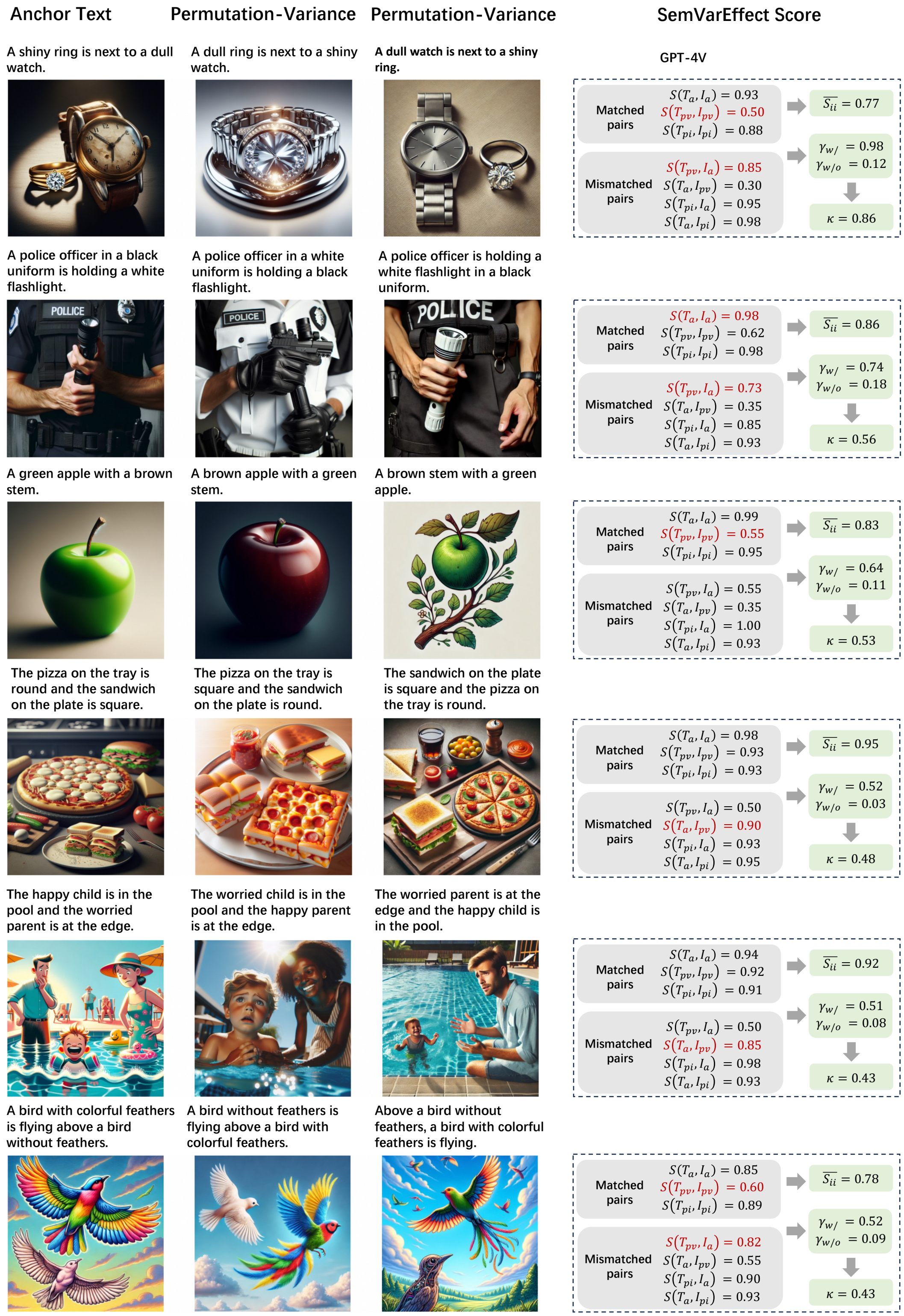
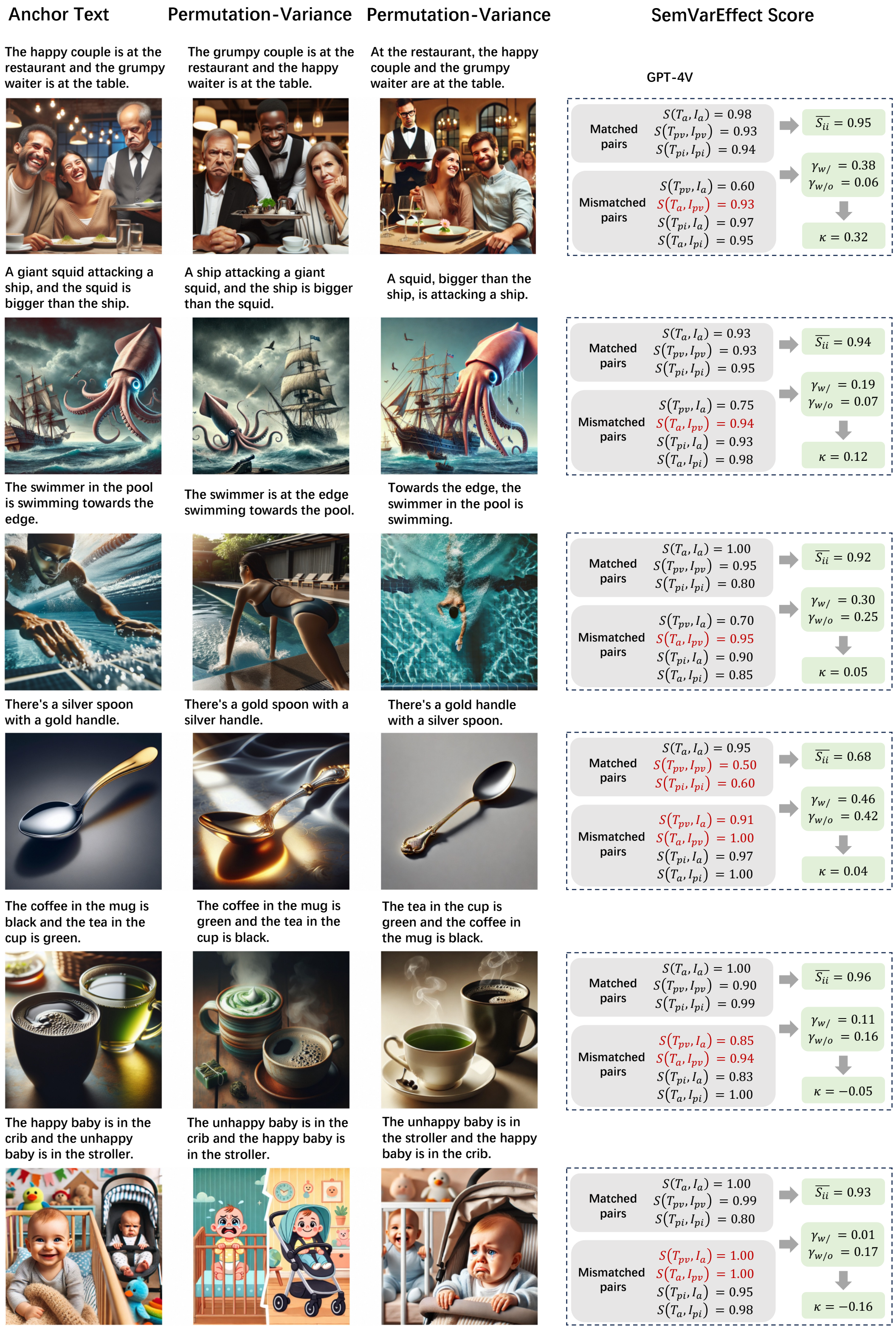

Anchor Text
Permutation-Variance
Permutation-Variance
SemVarEffect Score
GPT-4V
A shiny ring is next to a dull watch.
A dull ring is next to a shiny watch.
A dull watch is next to a shiny ring.
Matched pairs
$S(T_a, I_a) = 0.93$
$S(T_{pv}, I_{pv}) = 0.50$
$S(T_{pi}, I_{pi}) = 0.88$
Mismatched pairs
$S(T_{pv}, I_a) = 0.85$
$S(T_a, I_{pv}) = 0.30$
$S(T_{pi}, I_a) = 0.95$
$S(T_a, I_{pi}) = 0.98$
$\overline{S_{ii}} = 0.77$
$\gamma_{w/} = 0.98$
$\gamma_{w/o} = 0.12$
$\kappa = 0.86$
A police officer in a black uniform is holding a white flashlight.
A police officer in a white uniform is holding a black flashlight.
A police officer is holding a white flashlight in a black uniform.
Matched pairs
$S(T_a, I_a) = 0.98$
$S(T_{pv}, I_{pv}) = 0.62$
$S(T_{pi}, I_{pi}) = 0.98$
Mismatched pairs
$S(T_{pv}, I_a) = 0.73$
$S(T_a, I_{pv}) = 0.35$
$S(T_{pi}, I_a) = 0.85$
$S(T_a, I_{pi}) = 0.93$
$\overline{S_{ii}} = 0.86$
$\gamma_{w/} = 0.74$
$\gamma_{w/o} = 0.18$
$\kappa = 0.56$
A green apple with a brown stem.
A brown apple with a green stem.
A brown stem with a green apple.
Matched pairs
$S(T_a, I_a) = 0.99$
$S(T_{pv}, I_{pv}) = 0.55$
$S(T_{pi}, I_{pi}) = 0.95$
Mismatched pairs
$S(T_{pv}, I_a) = 0.55$
$S(T_a, I_{pv}) = 0.35$
$S(T_{pi}, I_a) = 1.00$
$S(T_a, I_{pi}) = 0.93$
$\overline{S_{ii}} = 0.83$
$\gamma_{w/} = 0.64$
$\gamma_{w/o} = 0.11$
$\kappa = 0.53$
The pizza on the tray is round and the sandwich on the plate is square.
The pizza on the tray is square and the sandwich on the plate is round.
The sandwich on the plate is square and the pizza on the tray is round.
Matched pairs
$S(T_a, I_a) = 0.98$
$S(T_{pv}, I_{pv}) = 0.93$
$S(T_{pi}, I_{pi}) = 0.93$
Mismatched pairs
$S(T_{pv}, I_a) = 0.50$
$S(T_a, I_{pv}) = 0.90$
$S(T_{pi}, I_a) = 0.93$
$S(T_a, I_{pi}) = 0.95$
$\overline{S_{ii}} = 0.95$
$\gamma_{w/} = 0.52$
$\gamma_{w/o} = 0.03$
$\kappa = 0.48$
The happy child is in the pool and the worried parent is at the edge.
The worried child is in the pool and the happy parent is at the edge.
The worried parent is at the edge and the happy child is in the pool.
Matched pairs
$S(T_a, I_a) = 0.94$
$S(T_{pv}, I_{pv}) = 0.92$
$S(T_{pi}, I_{pi}) = 0.91$
Mismatched pairs
$S(T_{pv}, I_a) = 0.50$
$S(T_a, I_{pv}) = 0.85$
$S(T_{pi}, I_a) = 0.98$
$S(T_a, I_{pi}) = 0.93$
$\overline{S_{ii}} = 0.92$
$\gamma_{w/} = 0.51$
$\gamma_{w/o} = 0.08$
$\kappa = 0.43$
A bird with colorful feathers is flying above a bird without feathers.
A bird without feathers is flying above a bird with colorful feathers.
Above a bird without feathers, a bird with colorful feathers is flying.
Matched pairs
$S(T_a, I_a) = 0.85$
$S(T_{pv}, I_{pv}) = 0.60$
$S(T_{pi}, I_{pi}) = 0.89$
Mismatched pairs
$S(T_{pv}, I_a) = 0.82$
$S(T_a, I_{pv}) = 0.55$
$S(T_{pi}, I_a) = 0.90$
$S(T_a, I_{pi}) = 0.93$
$\overline{S_{ii}} = 0.78$
$\gamma_{w/} = 0.52$
$\gamma_{w/o} = 0.09$
$\kappa = 0.43$

Figure 22: Errors only due to incorrect scoring by GPT-4V, where images are essentially correct.



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.

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.

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.

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.

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.

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.

Figure 23: Errors only due to incorrect scoring by GPT-4V, where images are essentially correct. The errors heavily influence the SemVarEffect scores.