

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

Xiangru Zhu, Pinglei Sun, Yaoxian Song, Yanghua Xiao, Zhixu Li,
Chengyu Wang, Jun Huang, Bei Yang, Xiaoxiao Xu

Fudan University & Alibaba Group

ICLR 2025

Problem

- Current T2I models struggle to capture semantic variations from word order changes
 - Models often treat text prompts as "bags of words"
 - Different permutations with distinct meanings yield similar images

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




Limitations of Current Evaluations

- Existing evaluation metrics rely on indirect measures but fail to reliably assess semantic understanding.

- Indirect measurement: Text-image alignment score as proxy

- High scores may obscure poor performance on complex patterns

- Focuses on frequent word combinations

Text	Image	Alignment Score
A mouse chasing a cat.		$S(T, I) = 0.65$
A cat chasing a mouse.		$S(T, I) = 0.98$
A mouse being chased by a cat.		$S(T, I) = 0.92$

Traditional Average Alignment

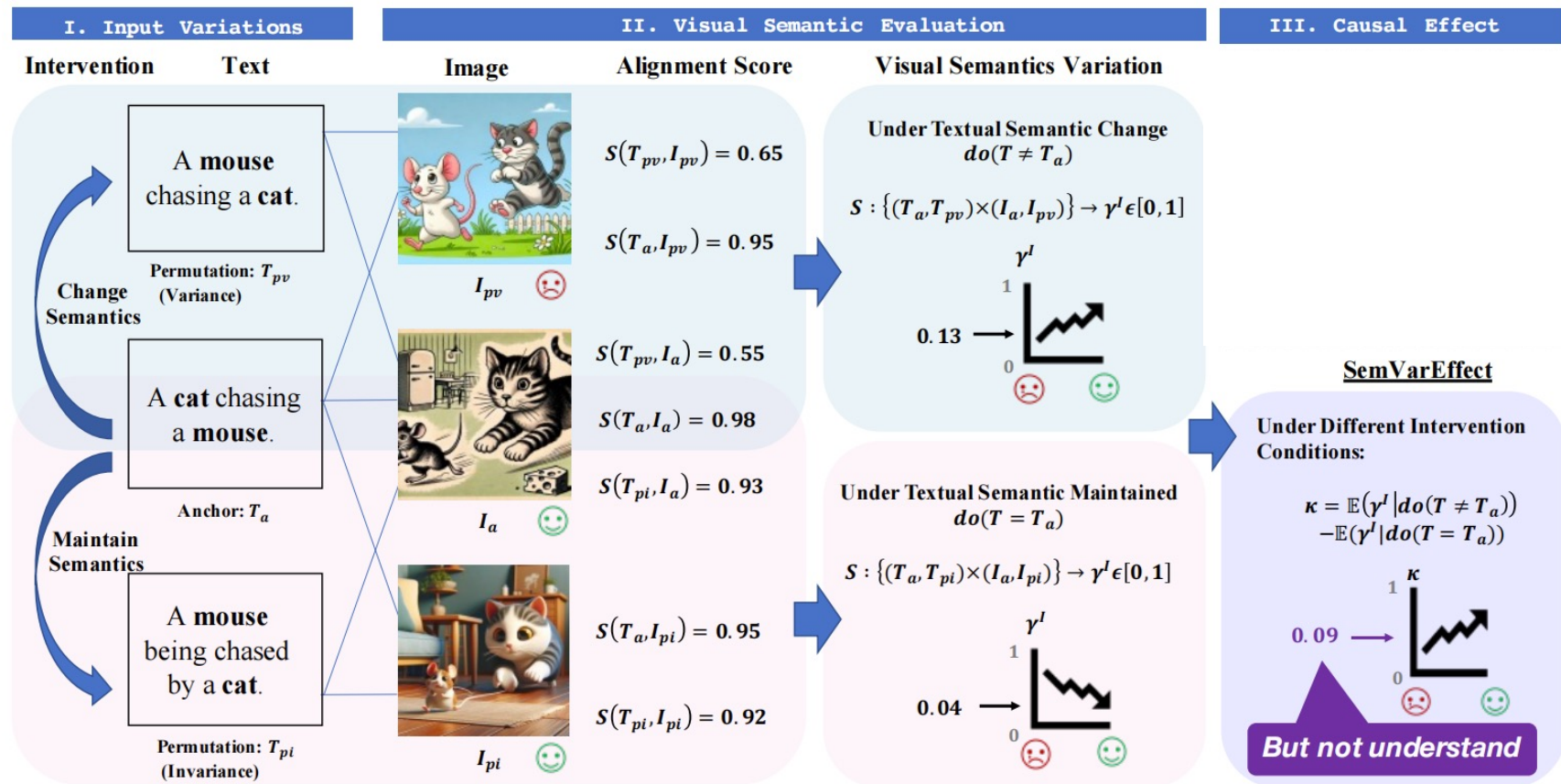
Seems Good



$$\begin{aligned}\overline{S(T, I)} &= (0.65 + 0.98 + 0.92) / 3 \\ &= 0.85\end{aligned}$$

Proposed Solution: SemVarEffect

Framework for measuring semantic variation causality



Proposed Solution: SemVarEffect

Input Variations

- **Permutation-Variance:** Different word orders → different meanings

A cat chasing
a mouse.

Anchor: T_a

Change
Semantics

A mouse
chasing a cat.

Permutation: T_{pv}

- **Permutation-Invariance:** Different word orders → same meaning

A cat chasing
a mouse.

Anchor: T_a

Maintain
Semantics

A mouse
being chased
by a cat.

Permutation: T_{pi}

Proposed Solution: SemVarEffect

Definition of Visual Semantic Variations

■ 1. Define visual semantic variations observed from a single sentence T

For each image I and its localized variation $I + \Delta I$, 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:

$$\mu_I (T, I) = S(T, I + \Delta I) - S(T, I)$$

When anchor image I_a is transformed into permutation image I_{p*} through localized changes, the total visual semantic variation is the sum across all changes:

$$\sum_{I_a}^{I_{p*}} \mu_I (T, I) = S(T, I_{p*}) - S(T, I_a)$$

Proposed Solution: SemVarEffect

Definition of Visual Semantic Variations

■ 2. Integrate visual semantic variations observed across multiple sentences

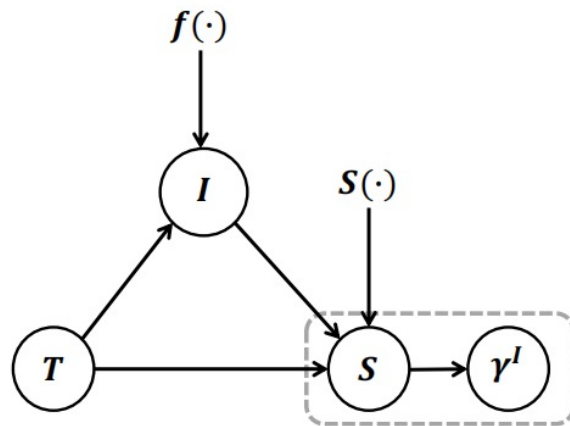
For sentences T_a and T_{p*} , the visual semantic variations have opposite directions. To measure total magnitude regardless of direction, we use absolute values:

$$\begin{aligned}\gamma^I &= \sum_{T \in \{T_a, T_{p*}\}} \left| \sum_{I_a}^{I_{p*}} \mu(T, I) \right| \\ &= |S(T_a, I_{p*}) - S(T_a, I_a)| + |S(T_{p*}, I_{p*}) - S(T_{p*}, I_a)|\end{aligned}$$

Proposed Solution: SemVarEffect

Causality Between Textual and Visual Semantic Variations

- Causal relationship between the input and the output semantic variations.



T : text input

I : generated image

S : text-image alignment score

γ^I : visual semantic variation

$f(\cdot)$: T2I model that maps T to I

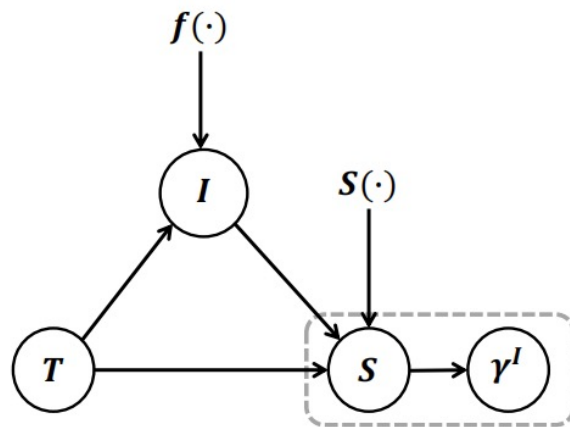
$S(\cdot)$: scoring function that maps T and I to S

γ^I is the sum of absolute differences in alignment scores S between original and permuted text-image pairs.

Proposed Solution: SemVarEffect

Causality Between Textual and Visual Semantic Variations

- Causal relationship between the input and the output semantic variations.



- $T \rightarrow$ **input variable** independent
- $I \rightarrow$ mediator
- $S \rightarrow$ intermediate result variable
- $\gamma^I \rightarrow$ **final comparison result variable**
- $f(\cdot) \rightarrow$ exogenous variable
- $S(\cdot) \rightarrow$ exogenous variable

- Define the ACE of textual semantic variations on visual semantic variations as the SemVarEffect score.

Proposed Solution: SemVarEffect

Average Causal Effect (ACE) of Textual Semantic Variations

■ Two types of interventions:

do($T \neq T_a$): Variations with semantic changes

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

do($T = T_a$): Variations without semantic changes

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

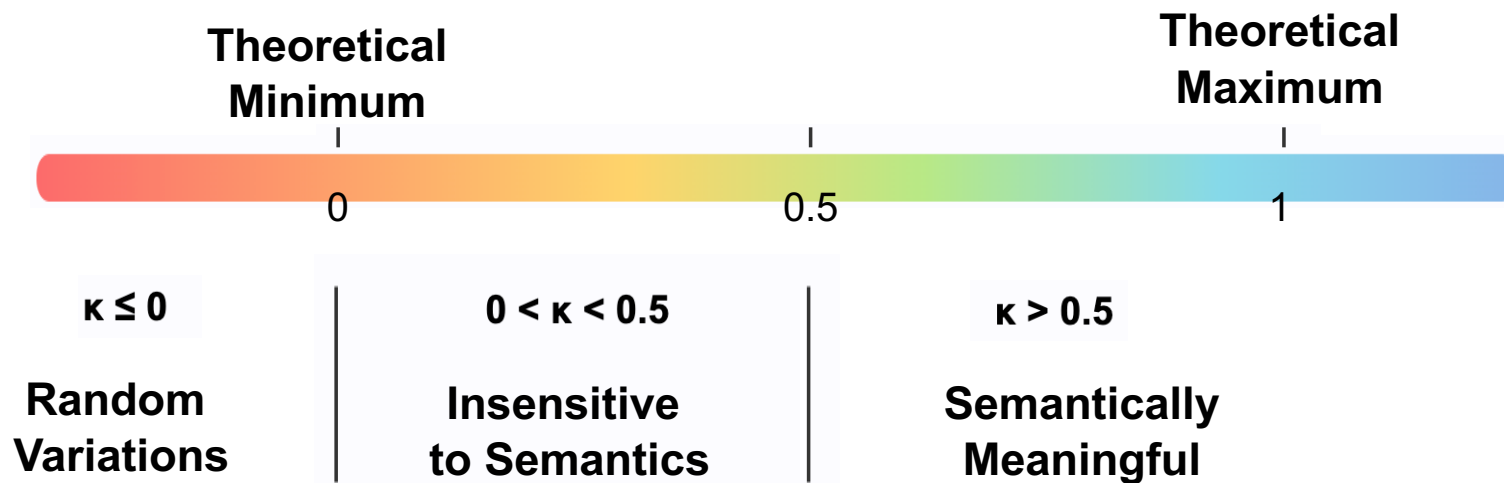
■ Average Causal Effect (SemVarEffect score):

$$\kappa = \mathbb{E}[\gamma^I | do(T \neq T_a)] - \mathbb{E}[\gamma^I | do(T = T_a)] = \gamma_{w/o}^I - \gamma_{w/o}^I$$

Proposed Solution: SemVarEffect

Understanding the SemVarEffect Score (κ)

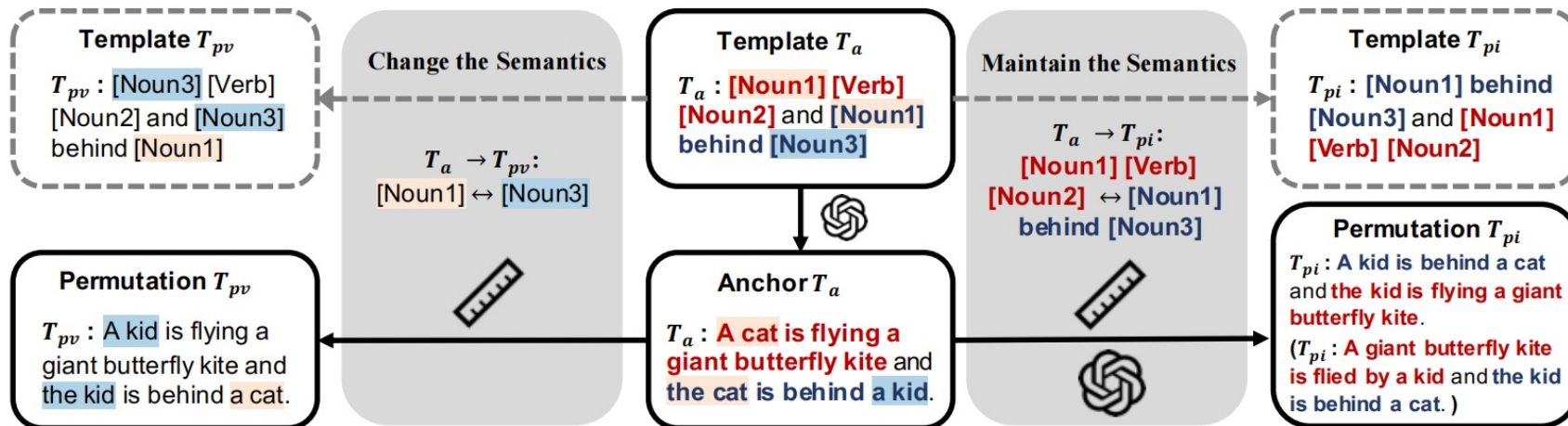
- **Components:** Alignment score $S(\cdot)$ includes object and relation (triple) components, each contributing up to 0.5 to the total score.
- **Theoretical Bounds:** The SemVarEffect score κ provides justified bounds (0.5-1.0) for evaluation.
- **Ideal Range:**



Proposed Benchmark: SemVarBench

Data Collection Process

- Extract templates from seed pairs $\rightarrow T_a$
- Generate permutations from templates $\rightarrow T_{pv}$ and T_{pi}
- Human verification and annotation
- Hard sample selection for testing



Proposed Benchmark: SemVarBench

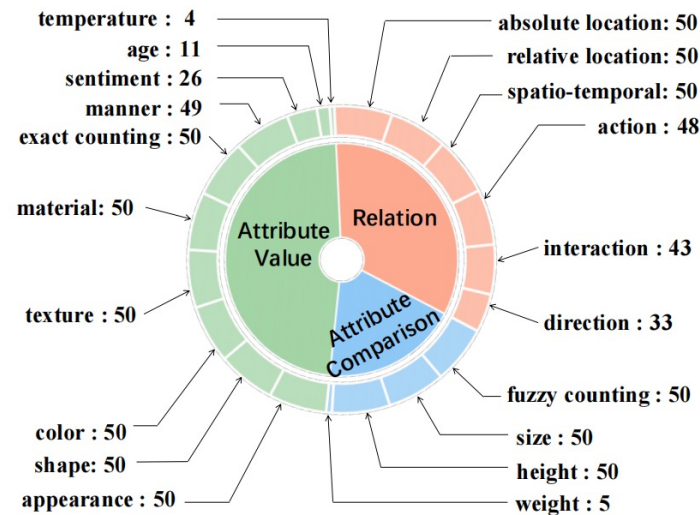
Statistics

■ Key Properties

- 11,454 samples (10,806 training, 648 testing)
- 20 Categories divided by semantic variation types
- Targets two types of linguistic permutations
- Expert-annotated for quality control

■ Categories

- Relation
- Attribute Comparison
- Attribute Value



Experimental Setup

■ Model Evaluated:

- 13 Mainstream Diffusion-based T2I models

■ Evaluation Metrics:

- S : Text-image alignment score
- κ : SemVarEffect score
 - $\gamma_{w/}^I$: Visual semantic variation with semantic change
 - $\gamma_{w/o}^I$: Visual semantic variation with meaning maintenance

■ Evaluators

- 4 MLLMs and Human

Key Findings

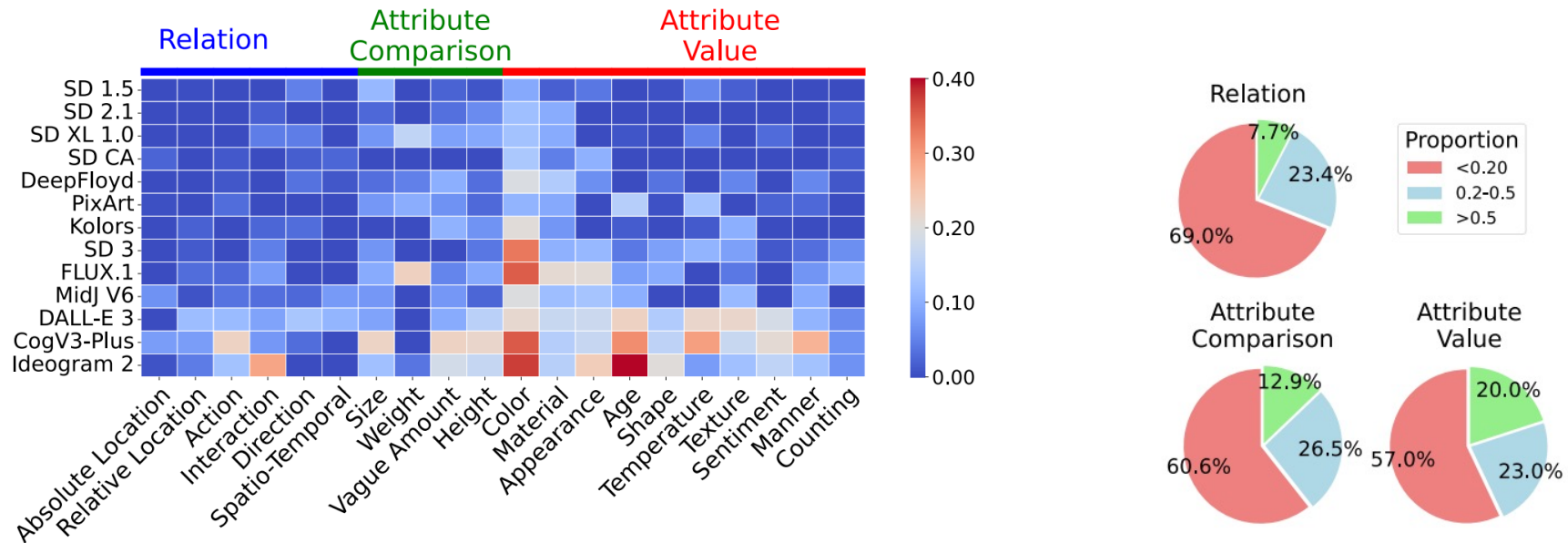
Models	Gemini 1.5 Pro				Claude 3.5 Sonnet				GPT-4o				GPT-4 Turbo			
	$\bar{S}(\uparrow)$	$\gamma_w(\uparrow)$	$\gamma_{wo}(\downarrow)$	$\kappa(\uparrow)$	$\bar{S}(\uparrow)$	$\gamma_w(\uparrow)$	$\gamma_{wo}(\downarrow)$	$\kappa(\uparrow)$	$\bar{S}(\uparrow)$	$\gamma_w(\uparrow)$	$\gamma_{wo}(\downarrow)$	$\kappa(\uparrow)$	$\bar{S}(\uparrow)$	$\gamma_w(\uparrow)$	$\gamma_{wo}(\downarrow)$	$\kappa(\uparrow)$
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	<u>0.32</u>	0.02	0.69	0.17	0.15	0.02	0.70	0.29	<u>0.26</u>	0.03	0.71	0.29	0.27	0.02
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	<u>0.17</u>	0.06	0.72	0.42	0.33	0.10	0.75	<u>0.40</u>	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	<u>0.13</u>	0.83	0.35	0.30	0.10
CogV3-Plus	<u>0.79</u>	0.52	0.35	<u>0.17</u>	0.80	0.28	0.18	0.10	<u>0.81</u>	0.49	0.28	0.20	<u>0.82</u>	0.43	<u>0.26</u>	0.17
Ideogram 2	0.80	<u>0.47</u>	0.29	0.18	<u>0.79</u>	<u>0.26</u>	<u>0.17</u>	<u>0.09</u>	<u>0.81</u>	<u>0.46</u>	0.27	0.20	0.81	<u>0.40</u>	0.24	<u>0.15</u>

■ Even SOTA models struggle with semantic variations

- Best models (CogView-3-Plus and Ideogram 2) achieve only 0.2/1.0 SemVarEffect score
- All models scored below 0.20, far from the ideal score of 1.0

Key Findings

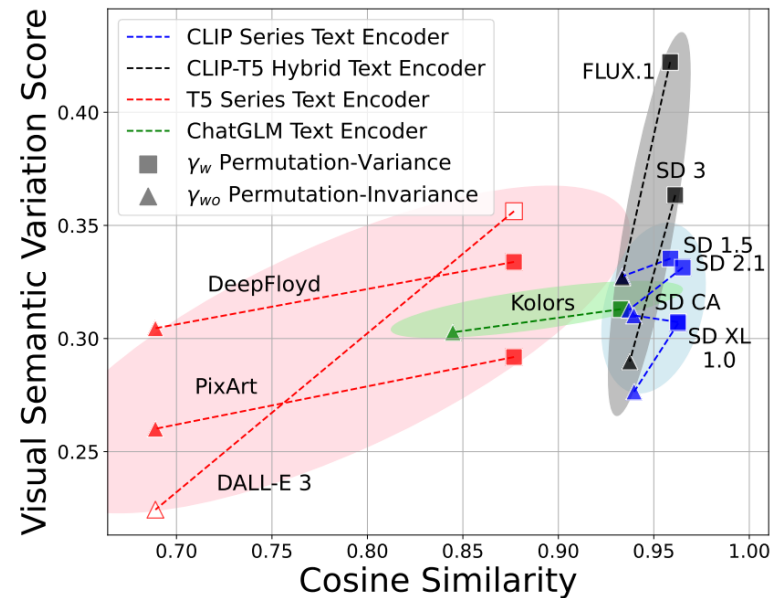
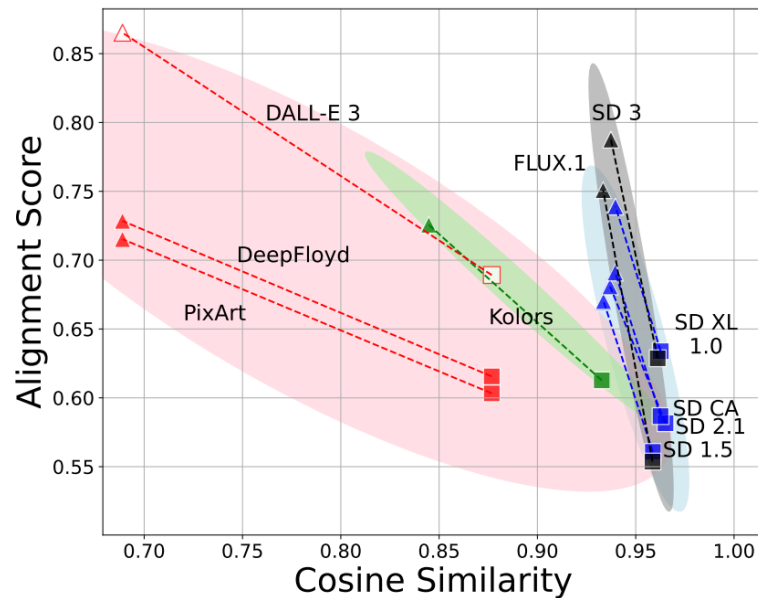
Does the influence of input semantic variations on output semantic variations vary by category?



- Semantic variations in object relations are less understood than attributes
 - Relation aspects scored 0.07/1.0 on average
 - Attribute values scored 0.17-0.19/1.0 on average
 - Color variations were best understood (0.4+ scores in top models)

Key Findings

Is a superior text encoder the exclusive solution for T2I models to grasp semantic variations?



- Cross-modal alignment in UNet or Transformers plays a crucial role
 - Superior text encoders alone are not sufficient
 - FLUX.1 with weaker text encoders outperformed some models with stronger encoders

Fine-Tuning Experiments

Does fine-tuning improve T2I model performance on semantic variations?

Category	Models	GPT-4o			
		$\bar{S}(\uparrow)$	$\gamma_w(\uparrow)$	$\gamma_{wo}(\downarrow)$	$\kappa(\uparrow)$
Color	SD XL	<u>0.73</u>	0.33	<u>0.25</u>	0.08
	+ sft-unet	0.78 (\uparrow)	0.38(\uparrow)	0.20 (\downarrow)	0.18 (\uparrow)
	+ sft-text	<u>0.73</u> ($-$)	0.40(\uparrow)	0.27(\uparrow)	0.13(\uparrow)
	+ dpo-unet	<u>0.69</u> (\downarrow)	<u>0.43</u> (\uparrow)	0.27(\uparrow)	<u>0.17</u> (\uparrow)
	+ dpo-text	0.68(\downarrow)	0.47 (\uparrow)	0.29(\uparrow)	0.18 (\uparrow)
Absolute Location	SD XL	<u>0.64</u>	0.29	0.34	-0.05
	+ sft-unet	0.65 (\uparrow)	0.34 (\uparrow)	0.32(\downarrow)	0.02 (\uparrow)
	+ sft-text	<u>0.64</u> ($-$)	0.31(\uparrow)	0.36(\uparrow)	-0.05($-$)
	+ dpo-unet	<u>0.60</u> (\downarrow)	0.29(\uparrow)	0.31 (\downarrow)	<u>-0.02</u> (\uparrow)
	+ dpo-text	0.57(\downarrow)	<u>0.33</u> (\uparrow)	0.39(\uparrow)	-0.07(\downarrow)

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(\downarrow)	0.24(\uparrow)	0.09(\downarrow)
	+ sft-text	<u>0.73</u> (\downarrow)	<u>0.39</u> (\uparrow)	0.34(\uparrow)	0.05(\downarrow)
	+ dpo-unet	0.71(\downarrow)	0.34($-$)	0.33(\uparrow)	0.02(\downarrow)
	+ dpo-text	0.66(\downarrow)	0.40 (\uparrow)	0.53(\uparrow)	-0.13(\downarrow)
Direction	SD XL	0.79	0.20	0.15	0.05
	+ sft-unet	<u>0.77</u> (\downarrow)	0.24(\uparrow)	0.23(\uparrow)	0.01(\downarrow)
	+ sft-text	<u>0.77</u> (\downarrow)	0.23(\uparrow)	0.21(\uparrow)	0.02(\downarrow)
	+ dpo-unet	0.65(\downarrow)	0.23(\uparrow)	0.26(\uparrow)	-0.03(\downarrow)
	+ dpo-text	0.70(\downarrow)	0.29 (\uparrow)	0.27(\uparrow)	0.01(\downarrow)

- Balancing sensitivity and robustness to semantic variations remains a challenge
 - DPO led to performance drops in permutation-invariance

Fine-Tuning Experiments

Does fine-tuning improve T2I model performance on semantic variations?

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 (\uparrow)	0.38(\uparrow)	0.20 (\downarrow)	0.18 (\uparrow)
	+ sft-text	0.73(—)	0.40(\uparrow)	0.27(\uparrow)	0.13(\uparrow)
	+ dpo-unet	0.69(\downarrow)	0.43(\uparrow)	0.27(\uparrow)	0.17(\uparrow)
	+ dpo-text	0.68(\downarrow)	0.47 (\uparrow)	0.29(\uparrow)	0.18 (\uparrow)
Absolute Location	SD XL	0.64	0.29	0.34	-0.05
	+ sft-unet	0.65 (\uparrow)	0.34 (\uparrow)	0.32(\downarrow)	0.02 (\uparrow)
	+ sft-text	0.64(—)	0.31(\uparrow)	0.36(\uparrow)	-0.05(—)
	+ dpo-unet	0.60(\downarrow)	0.29(\uparrow)	0.31 (\downarrow)	-0.02 (\uparrow)
	+ dpo-text	0.57(\downarrow)	0.33(\uparrow)	0.39(\uparrow)	-0.07(\downarrow)

Category	Model	Token Accuracy		
		T_a	T_{pv}	T_{pi}
Absolute	SDXL	0.709	0.640	0.716
Location	+sft-unet	0.718	0.654	0.716

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(\downarrow)	0.24(\uparrow)	0.09(\downarrow)
	+ sft-text	0.73(\downarrow)	0.39(\uparrow)	0.34(\uparrow)	0.05(\downarrow)
	+ dpo-unet	0.71(\downarrow)	0.34(—)	0.33(\uparrow)	0.02(\downarrow)
	+ dpo-text	0.66(\downarrow)	0.40 (\uparrow)	0.53(\uparrow)	-0.13(\downarrow)
Direction	SD XL	0.79	0.20	0.15	0.05
	+ sft-unet	0.77(\downarrow)	0.24(\uparrow)	0.23(\uparrow)	0.01(\downarrow)
	+ sft-text	0.77(\downarrow)	0.23(\uparrow)	0.21(\uparrow)	0.02(\downarrow)
	+ dpo-unet	0.65(\downarrow)	0.23(\uparrow)	0.26(\uparrow)	-0.03(\downarrow)
	+ dpo-text	0.70(\downarrow)	0.29 (\uparrow)	0.27(\uparrow)	0.01(\downarrow)

Category	Model	Token Accuracy		
		T_a	T_{pv}	T_{pi}
Height	SDXL	0.881	0.660	0.861
	+sft-unet	0.886	0.662	0.866

- Fine-tuning improves token-region correspondence but fails to enhance understanding of semantic relationships between tokens
- Token-level accuracy improved without enhancing semantic understanding

Fine-Tuning Experiments

Do T2I models' struggles with semantic relationships stem from training data imbalance?

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

Class	Reasons	SD XL		FT SD XL (trained on cat↔dog)	
		bull→man	man→bull	bull→man	man→bull
Wrong	Missing Objects	0	0	0	0
	No Interaction	6	12	6	13
	Wrong Interaction	5	2	0	0
	Wrong Direction	16	12	11	9
	Reversed Role	2	2	9	3
Right	Partial/Full Match	1	2	4	5

- Even with perfectly balanced training data, models fail to understand directional relationships

Resources & Contact Information

SemVarBench Dataset



github.com/zhuxiangru/SemVarBench

Contact



xrzhu19@fudan.edu.cn

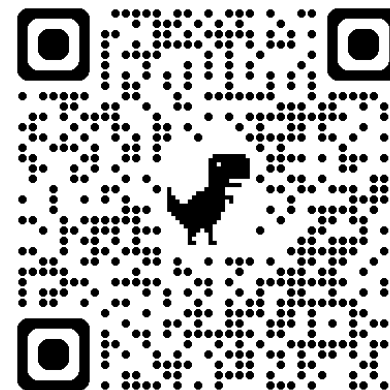
Our Lab's Previous Work



Knowledge Works Research Laboratory
@ Fudan University



<http://kw.fudan.edu.cn/>



Scan to access SemVarBench
We welcome collaborations and feedback!



Scan to access Knowledge Works