

Role Bias in Diffusion Models: Diagnosing and Mitigating through Intermediate Decomposition

NeurIPS 2025

Sina Malakouti and Adriana Kovashka

Computer Science Department, University of Pittsburgh

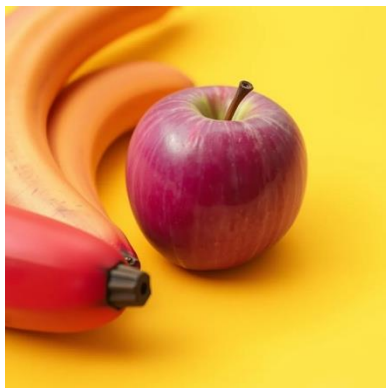


Limitation in compositional generalization

Attribute binding and spatial relation



Pink box on top of
red box on top of
blue box



Red banana and
blue apple

Rare concept generation



Furry Frog



Mouse chasing cat

Action-Based Relation Generation

Cat chasing mouse



DALL-E 3



SDXL

Mouse chasing cat



DALL-E 3



SDXL

- T2I models fail in **rare action-based relations**
- **Role collapse**
 - They often **default to their frequent counter parts**

Overview

Mouse chasing cat



DALL-E 3



SDXL

Mouse chasing boy



DALL-E 3



SDXL

Mouse chasing cat



Ours (SDXL)



Ours (SDXL)

Observation

- Models can generate *similarly rare* relations that their frequent is not as dominant

Hypothesize

- Over representation of frequent** compositions impedes *rare generation*
- Similar but plausible composition can reinforce role binding

RoleBench

Benchmark

- 10 relations
 - chasing, riding, lifting, etc.
- Animate-animate interactions
- Rare relation whose reverse is frequent
 - *Frequent:* $c_F = (s_F, r, o_F)$ *Rare:* $c_R = (s_R = o_F, r, o_R = s_F)$
 - $p_D(c_F) \gg p_D(c_R)$
 - We use semantic plausibility as a proxy
- **Role-binding & Direction**
 - Requires **assigning correct roles** and **direction**

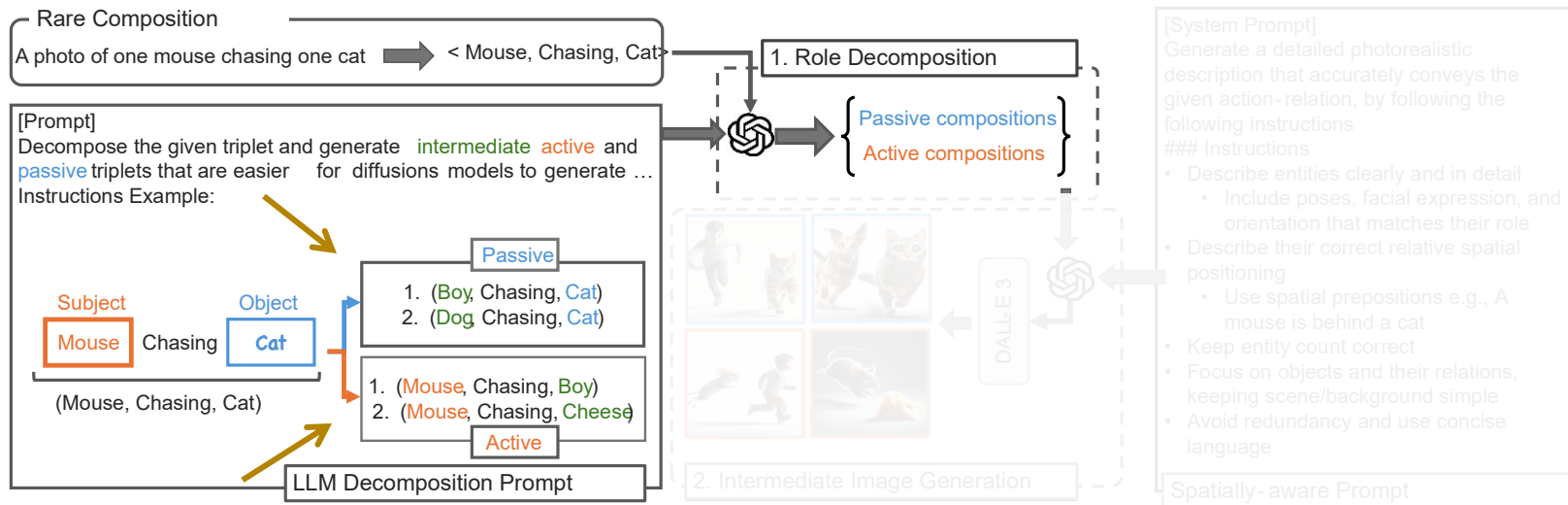


ReBind

➤ Step1: Role Decomposition

- Rare composition is decomposed into
 - **Active:** *subject* remains the same
 - **Passive:** *object* remains the same

- Intermediates must be plausibility (i.e., generatable)
 - LLM semantic plausibility
- Use active/passives to improve rare concept generation



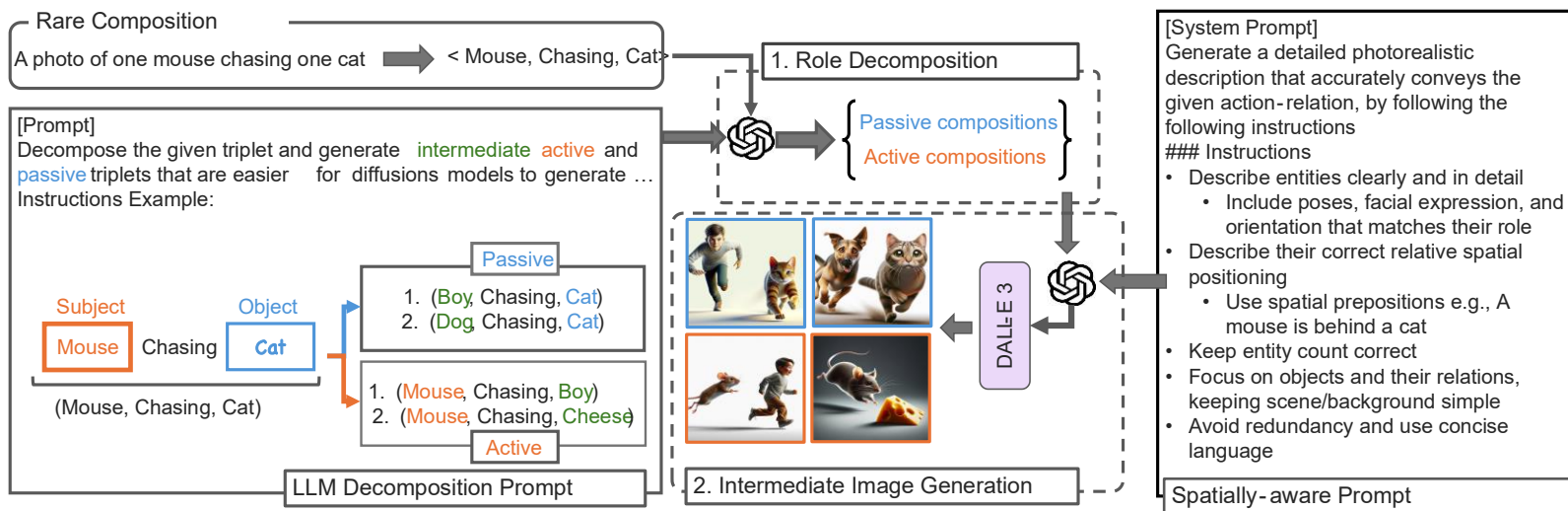
ReBind

Step1: Role Decomposition

- Rare composition is decomposed into active/passive intermediates

➤ Step2: Intermediate Generation

- Simple description (mouse chasing cat) is abstract
- Generate spatially-aware description of relation
 - Objects, spatial and non spatial attributes info
- Synthesize images via Dalle-3
 - We filter low quality images



ReBind

Step 1: Role Decomposition

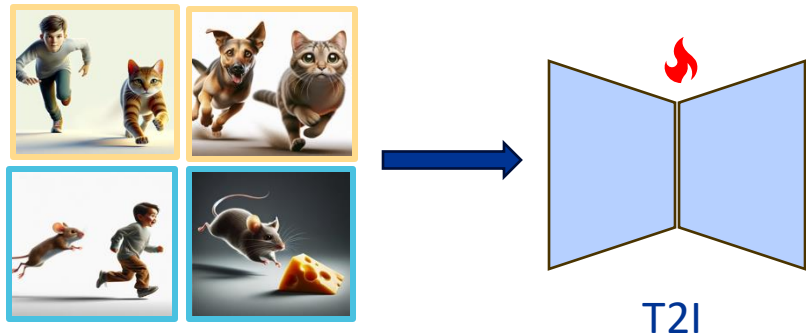
- Rare composition is decomposed into active/passive intermediates

➤ Step 3: Finetuning

- Finetuning the model on synthetic images with LoRA
- λ : active role weight

Step 2: Intermediate Generation

- Synthesize images via Dalle-3



$$\mathcal{L}(\theta) = \mathbb{E}_{(x_0, p) \sim \mathcal{D}, \epsilon, t} [\|\epsilon - \epsilon_{\theta}(z_t, p, t)\|_2^2]$$

$$\mathcal{L}_{\text{compos}} = \lambda \cdot \mathcal{L}_{\text{active}} + \mathcal{L}_{\text{passive}}$$

Evaluation

Role Bias β for composition c

- Image-Text Alignment (ITA)
 - We use VQAScore (Lin et al., 2024)
- E.g., $c = (\text{mouse, chasing, cat})$

$$\beta = \text{ITA}(I, c^{rev}) - \text{ITA}(I, c)$$



$$\beta = \text{ITA} \left(\text{Image of a cat chasing a mouse}, \text{cat chasing mouse} \right) - \text{ITA} \left(\text{Image of a cat chasing a mouse}, \text{mouse chasing cat} \right)$$

Role-Specific Questions

- Facial Expression, Spatial, Orientation, Pose
- E.g., Is the cat behind the mouse? (Yes/No)

Role collapse in pre-trained T2I models

- T2I models default to the freq. composition
 - Positive β on rare compositions

Category	T2I Model	Size (B)	Spatial			Orientation			Pose			Facial Expression			VQAScore			Negative β
			M	U	$\beta \downarrow$	M	U	$\beta \downarrow$	M	U	$\beta \downarrow$	M	U	$\beta \downarrow$	M	U	$\beta \downarrow$	
Frequent	SDXL	3.5	82.70	67.80	-14.90	76.10	68.60	-7.50	74.20	69.10	-5.10	81.80	70.30	-11.50	84.00	57.70	-26.30	
	SD3	2	80.20	65.10	-15.10	74.10	68.90	-5.20	71.50	64.20	-7.30	78.40	68.20	-10.20	82.60	53.30	-29.30	
	SD3.5	2.5	83.30	66.10	-17.20	76.10	68.90	-7.20	73.00	64.70	-8.30	81.20	69.60	-11.60	84.90	51.30	-33.60	
	AuraFlow2	6.8	87.20	69.10	-18.10	78.50	71.30	-7.20	76.90	69.10	-7.80	83.20	70.40	-12.80	84.40	60.80	-23.60	
	DALL-E 3	12	86.90	70.50	-16.40	78.30	70.20	-8.10	77.50	67.20	-10.30	80.60	67.80	-12.80	88.30	60.60	-27.70	
Rare	SDXL	3.5	68.20	80.30	12.10	68.60	74.30	5.70	70.30	73.90	3.60	71.70	80.00	8.30	56.60	79.70	23.10	
	SD3	2	65.10	75.40	10.30	66.40	70.30	3.90	64.80	66.40	1.60	68.50	75.30	6.80	56.40	73.60	17.20	
	SD3.5	2.5	66.90	77.80	10.90	68.30	72.60	4.30	65.90	69.00	3.10	72.70	79.10	6.40	59.50	80.40	20.90	
	AuraFlow2	6.8	72.90	83.70	10.80	71.80	77.60	5.80	75.10	77.90	2.80	76.60	83.60	7.00	74.60	84.70	10.10	
	DALL-E 3	12	74.90	84.10	9.20	71.30	75.60	4.30	73.00	76.50	3.50	73.60	78.10	4.50	68.90	84.90	16.00	

$\beta = \text{VQAScore (unmatching)} - \text{VQAScore (matching)}$

M: matching, U: unmatching

Role collapse through attentions

Token: <chasing>



rare



frequent

Token: <riding>



rare



frequent

Images are generated by Stable Diffusion 3 Medium

Role Collapse in compositional methods

- Compositional generation methods are not effective
 - Expensive: Spatial prior, cross-attention manipulation, LLM guided generation
 - Don't resolve existing bias and model may not follow spatial prior
- ReBind is more effective
 - Especially on non-spatial (Facial Expression)

Model	Spatial			Orientation			Pose			Facial Expression			VQAScore		
	M	U	$\beta \downarrow$	M	U	$\beta \downarrow$	M	U	$\beta \downarrow$	M	U	$\beta \downarrow$	M	U	$\beta \downarrow$
DALL-E 3	74.9	84.1	9.2 -	71.3	75.6	4.3 -	73.0	76.5	3.5 -	73.6	78.1	4.5 -	68.9	84.9	16.0 -
SDXL	68.2	80.3	12.1 -	68.6	74.3	5.7 -	70.3	73.9	3.6 -	71.7	80.0	8.3 -	56.6	79.7	23.1 -
IterComp	66.8	80.3	13.5 (-1.4)	68.4	75.3	6.9 (-1.2)	68.5	72.5	4.0 (-0.4)	70.0	80.3	10.3 (-2.0)	59.2	81.4	22.2 (0.9)
RRNet	68.2	76.8	8.6 (3.5)	67.5	71.2	3.7 (2.0)	69.1	72.2	3.1 (0.5)	71.0	74.7	3.7 (4.6)	61.2	76.3	15.1 (8.0)
R2F	68.0	79.4	10.7 (1.4)	68.0	73.4	5.4 (0.3)	71.0	74.4	3.4 (0.2)	71.9	79.3	7.4 (0.9)	43.4	59.1	15.7 (7.4)
RPG	65.4	74.3	8.9 (3.2)	66.6	70.6	4.0 (1.7)	63.7	66.7	3.0 (0.6)	67.3	75.8	8.5 (-0.2)	56.6	73.3	16.7 (6.4)
SLD	68.4	75.1	6.7 (5.0)	67.2	70.1	2.9 (2.8)	67.2	70.0	2.8 (0.8)	69.8	74.9	5.0 (3.3)	56.8	73.4	16.6 (6.5)
ReBind	68.5	75.5	7.0 (5.1)	66.0	69.6	3.6 (2.1)	73.3	76.5	3.2 (0.4)	67.7	70.2	2.5 (5.8)	57.9	65.8	7.9 (15.2)

Low improvement

β = VQAScore (unmatching) – VQAScore (matching)

M: matching, U: unmatching

Is ReBind effective?

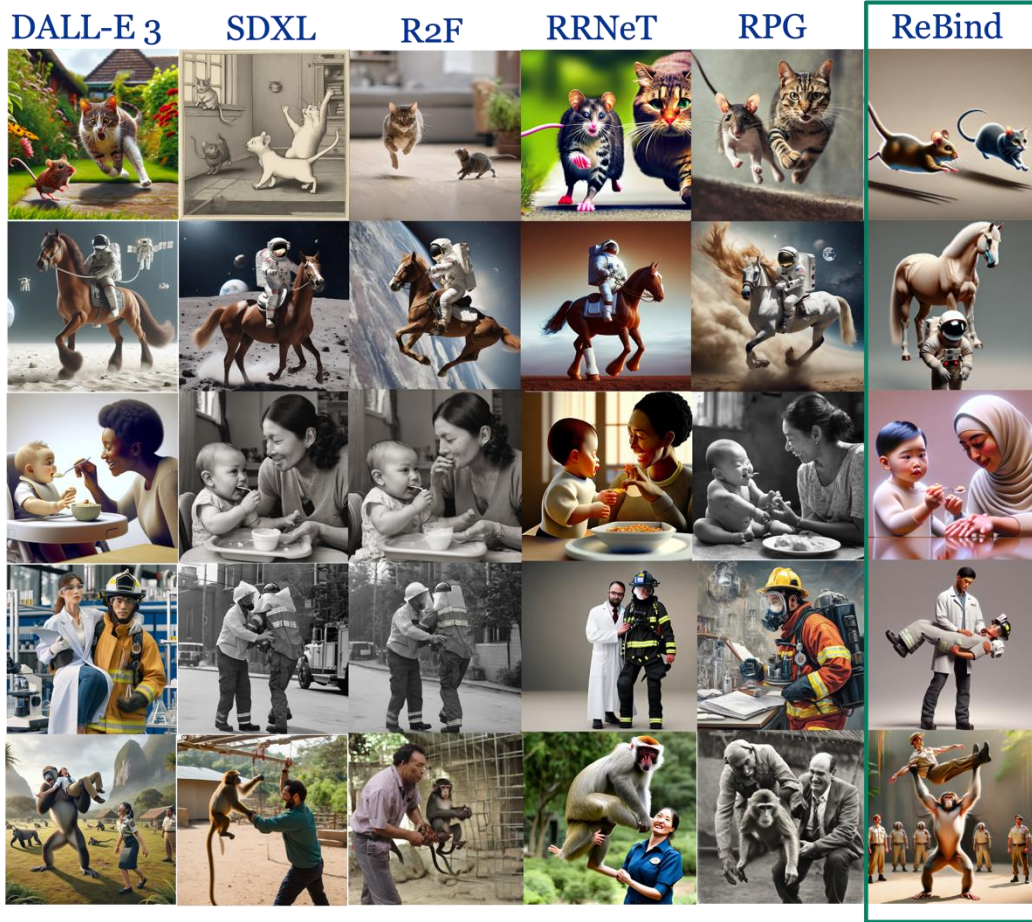
Mouse chasing cat

Horse riding astronaut

Baby feeding food to a woman

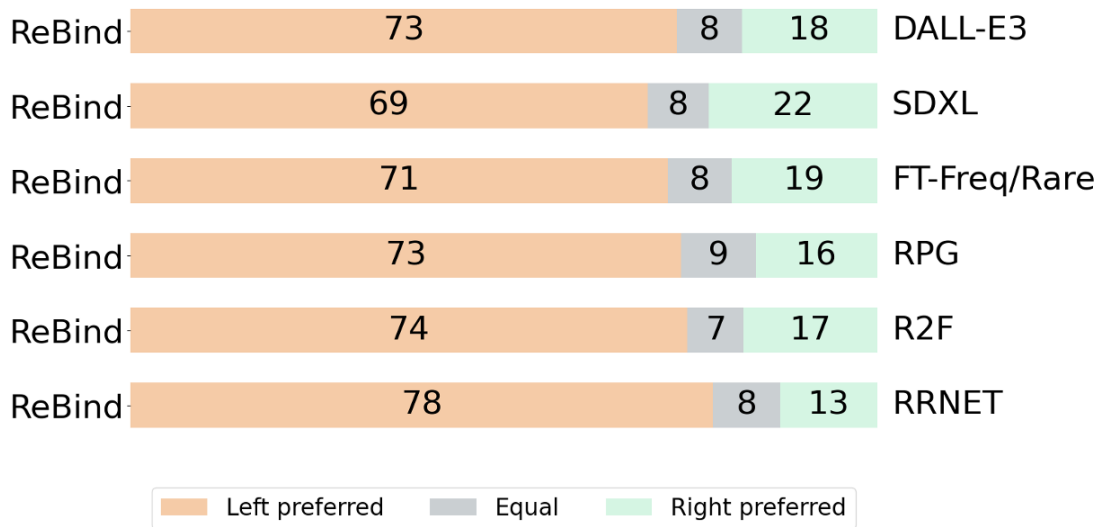
Scientist carrying fireman

Monkey lifting zoo trainer



Is ReBind effective?

- Humans compare ReBind and baselines in a head-to-head manner
 - Compare top-3 output of models selected by β
- Training on intermediate compositions is significantly more effective



Thank you so much for your attention!

