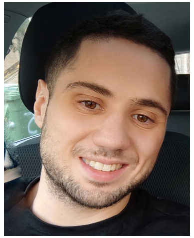
Linguistic Binding in Diffusion Models: Enhancing Attribute Correspondence through Attention Map Alignment

Royi Rassin



Eran Hirsch

Daniel Glickman

Shauli Ravfogel

Yoav Goldberg

Gal Chechik

Improper Binding

Improper Binding

A yellow flamingo and a pink sunflower

Improper Binding

A yellow <u>flamingo</u> and a pink <u>sunflower</u>

2 entities

Improper Binding

A yellow flamingo and a pink sunflower

2 modifiers

Improper Binding

pink *yellow*

A ~~yellow~~ flamingo and a ~~pink~~ sunflower

Improper Binding

pink *yellow*

A ~~yellow~~ flamingo and a ~~pink~~ sunflower

Leak “in” Prompt

Improper Binding

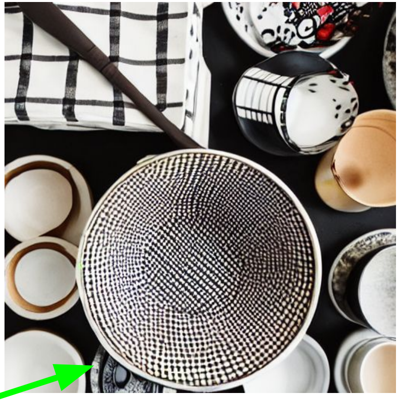
A checkered bowl in a cluttered room

Improper Binding

A checkered bowl in a cluttered room

Improper Binding

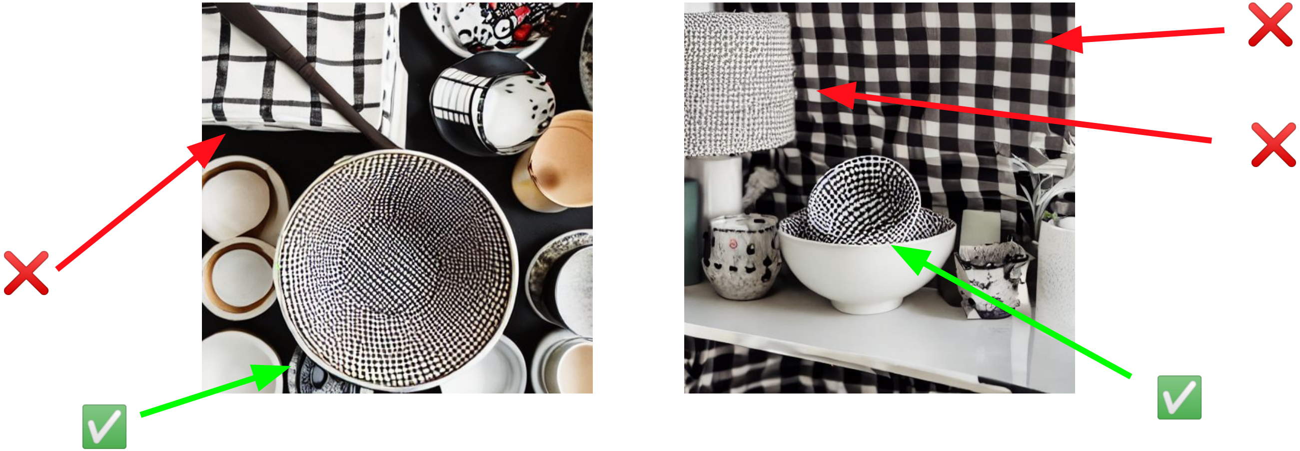
A ***checkered*** *bowl in a* ***cluttered*** *room*



✅ ✅

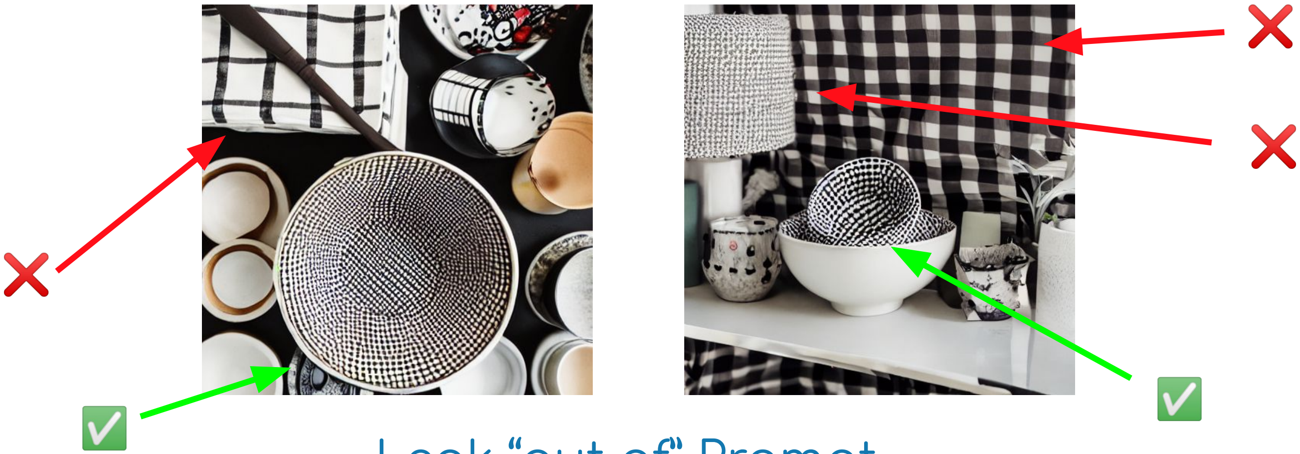
Improper Binding

A checkered bowl in a cluttered room



Improper Binding

A checkered bowl in a cluttered room



Leak “out of” Prompt

Improper Binding

A <u>horned</u> lion and a <u>spotted</u> monkey

Improper Binding

A ~~horned~~ lion and a ~~spotted~~ monkey

Improper Binding

A ~~horned~~ lion and a ~~spotted~~ monkey

Improper Binding

A ~~horned~~ lion and a ~~spotted~~ monkey

Attribute Neglect

Improper Binding | MidJourney-5

A yellow flamingo and a pink sunflower

a checkered bowl in a cluttered room

a horned lion and a spotted monkey

Improper Binding | DALL-E 3

A pink sunflower and a yellow flamingo

a checkered bowl in a cluttered room

a horned lion and a spotted monkey

Improper Binding | DALL-E 3

A pink sunflower and a yellow flamingo

a checkered bowl in a cluttered room

a horned lion and a spotted monkey

Why does it happen?

- The underlying model does not represent the relations between words
- The text encoder acts to a large extent as a bag of words

How do we solve this?

- Use parser to inject linguistic knowledge
- Uncover semantic constraints
- Enforce the constraints by intervening in the generation process

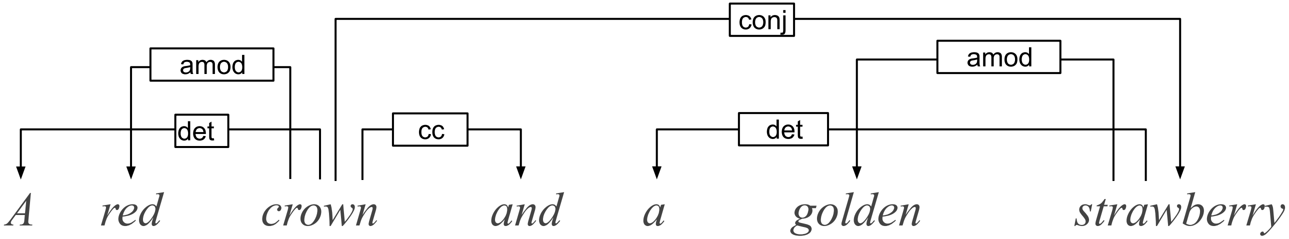
SynGen | Our goal

- We seek to fix **all three leakage types**
- In **inference-time** (no training or fine-tuning)

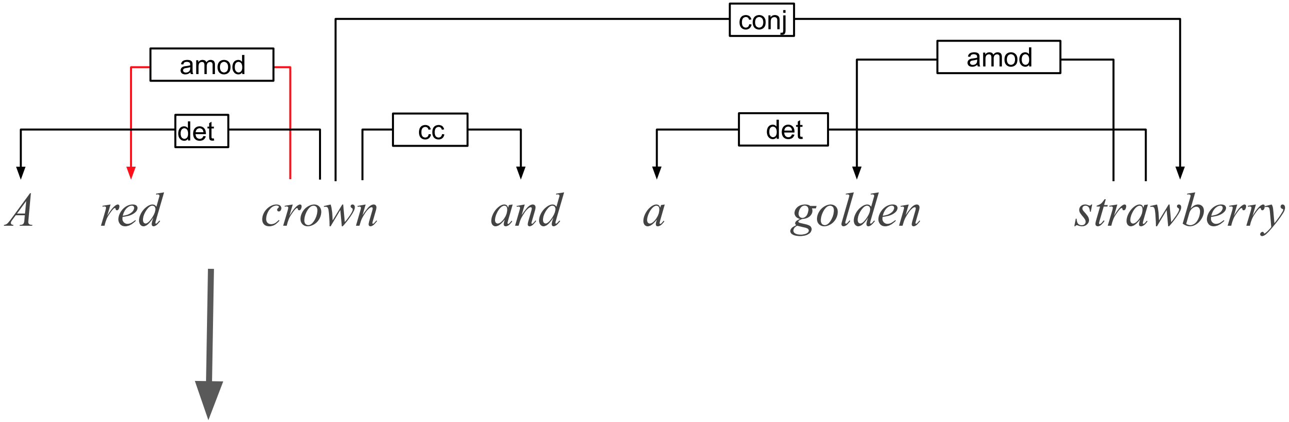
SynGen | Our approach

- Obtain the syntactic structure of the prompt
- Guide the diffusion on the prompt's syntax
- Steer the cross-attention using syntax in inference-time

SynGen | Syntactic structure

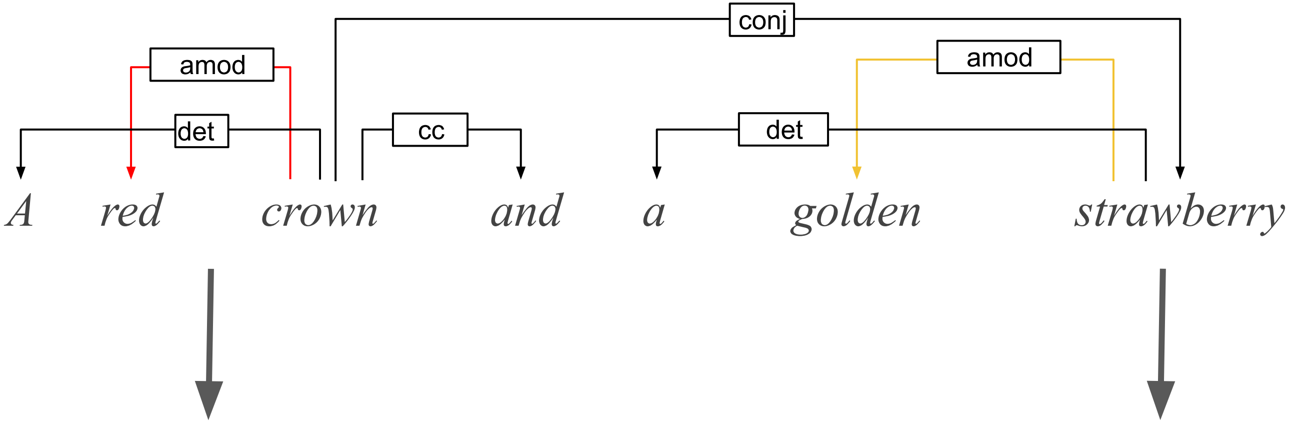


SynGen | Syntactic structure



S_1 = {crown, red}

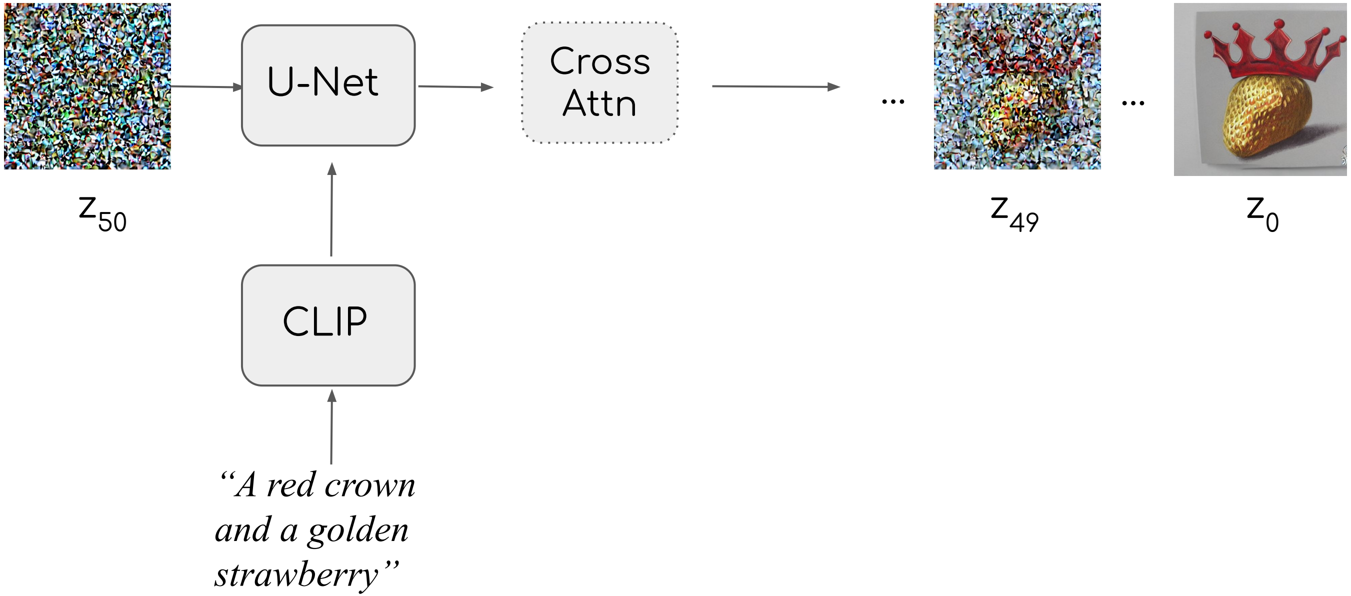
SynGen | Syntactic structure



S_1 = {crown, red}

S_2 = {strawberry, golden}

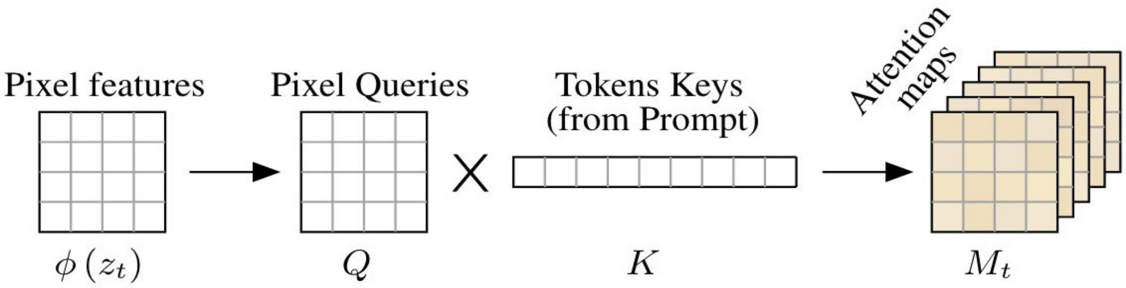
SynGen | Obtaining Cross Attention Maps



"Prompt-to-Prompt Image Editing with Cross Attention Control" by Hertz et al., 2022

SynGen | Obtaining cross-attention maps

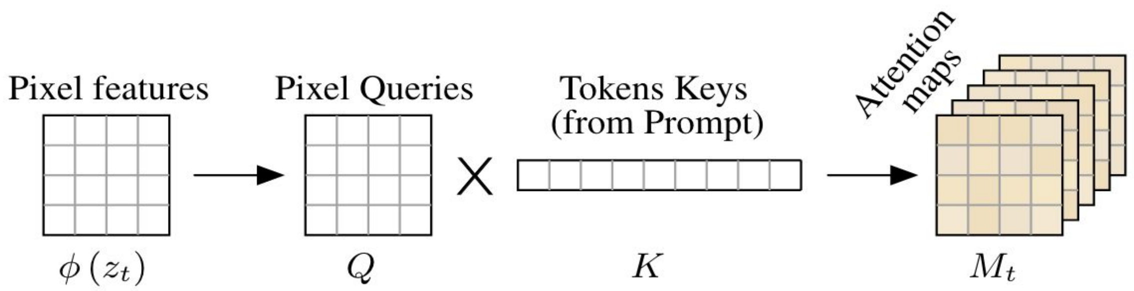
Cross-attention



The figure is taken from "Prompt-to-Prompt Image Editing with Cross Attention Control"

SynGen | Obtaining cross-attention maps

Cross-attention



crown

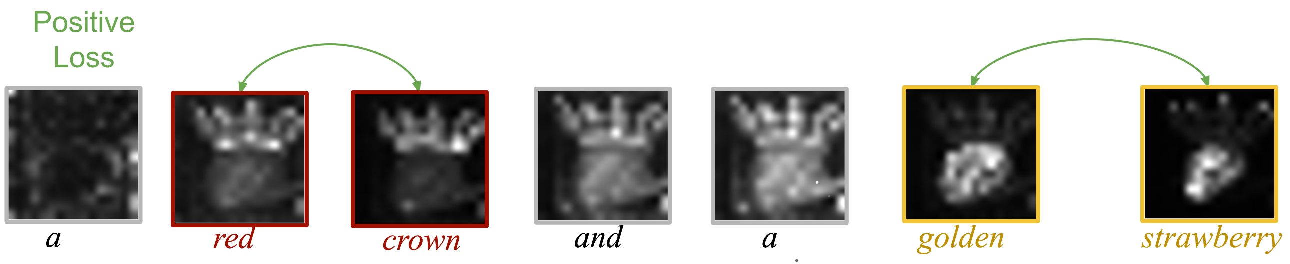
The figure is taken from "Prompt-to-Prompt Image Editing with Cross Attention Control"

SynGen | Aligning the denoising process

- Cross-attention maps are (token,patch) pairs and are derived from the latent
- We can define a loss that updates the latent (noise)

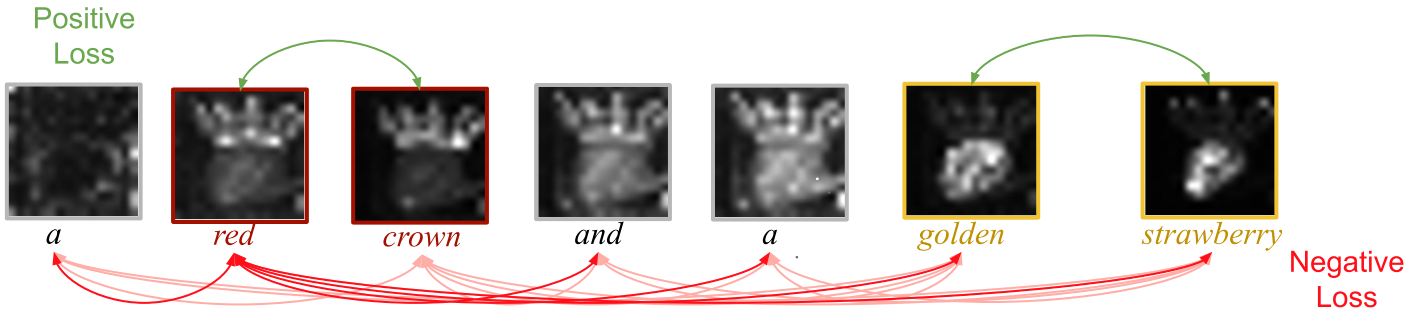
SynGen | Aligning the denoising process

- Cross-attention maps are (token,patch) pairs and are derived from the latent
- We can define a loss that updates the latent (noise)
  - encourage overlap of maps corresponding to entities and their modifiers



SynGen | Aligning the denoising process

- Cross-attention maps are (token,patch) pairs and are derived from the latent
- We can define a loss that updates the latent (noise)
  - encourage overlap of maps corresponding to entities and their modifiers
  - discourage overlap with all other maps



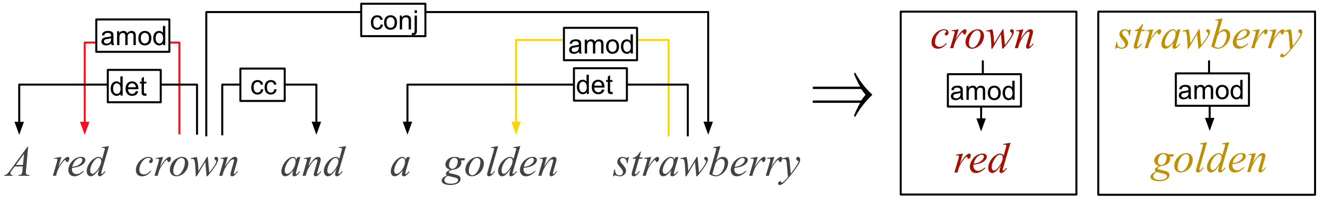
SynGen | Computing the loss

- Minimize distance over related (entity, modifier) pairs
  - Normalize maps
  - Compute Symmetric KL
- Maximize distance over non-related (entity, modifier) pairs
  - Normalize maps
  - Compute Symmetric KL
  - Negate result
- Adding the terms: $L = L_{pos} + L_{neg}$

SynGen | Workflow

SynGen | Workflow

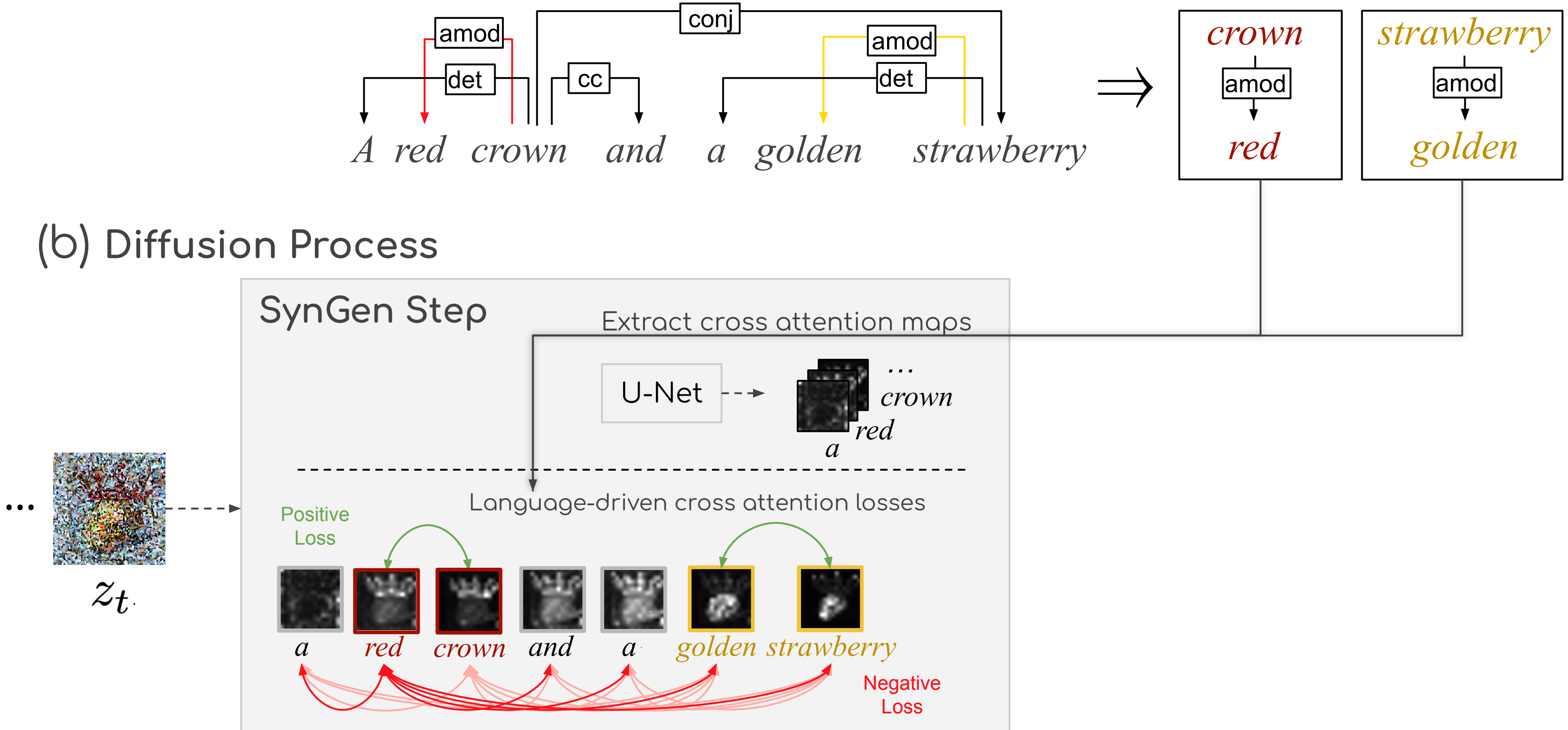
(a) Extract Entities and Modifiers



SynGen | Workflow

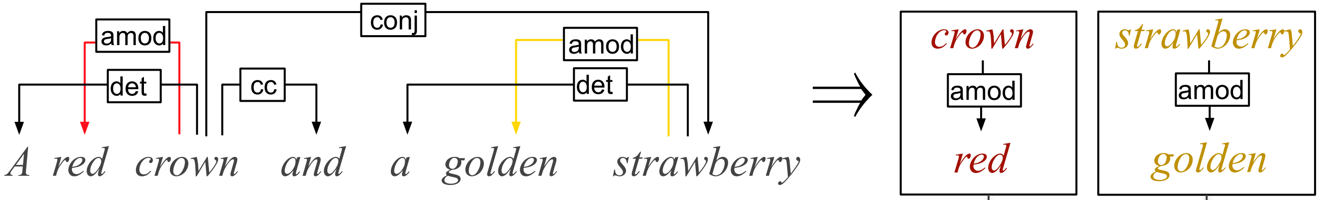
(a) Extract Entities and Modifiers

(b) Diffusion Process

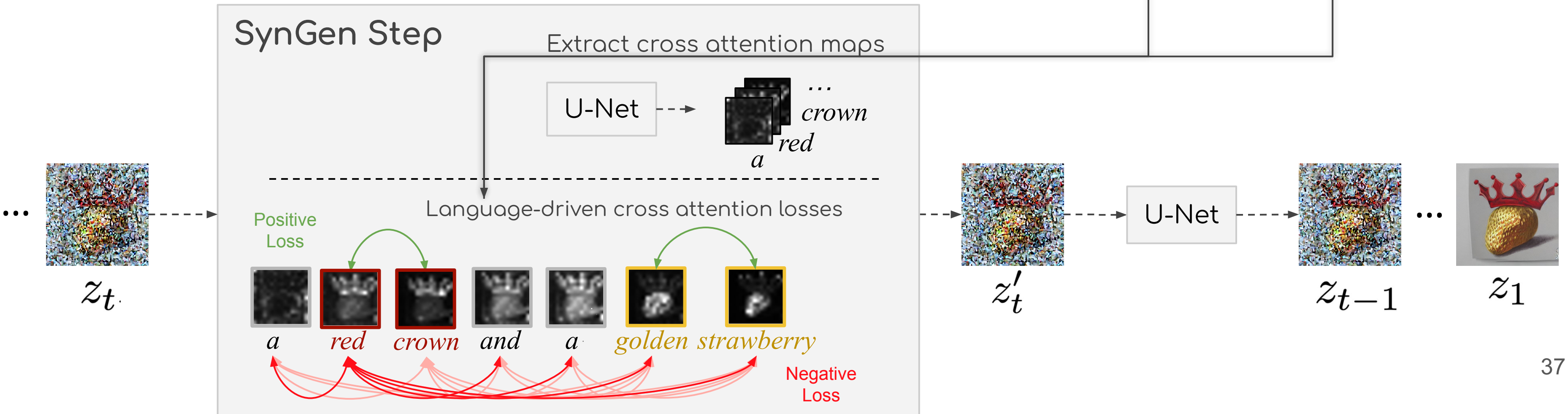


SynGen | Workflow

(a) Extract Entities and Modifiers



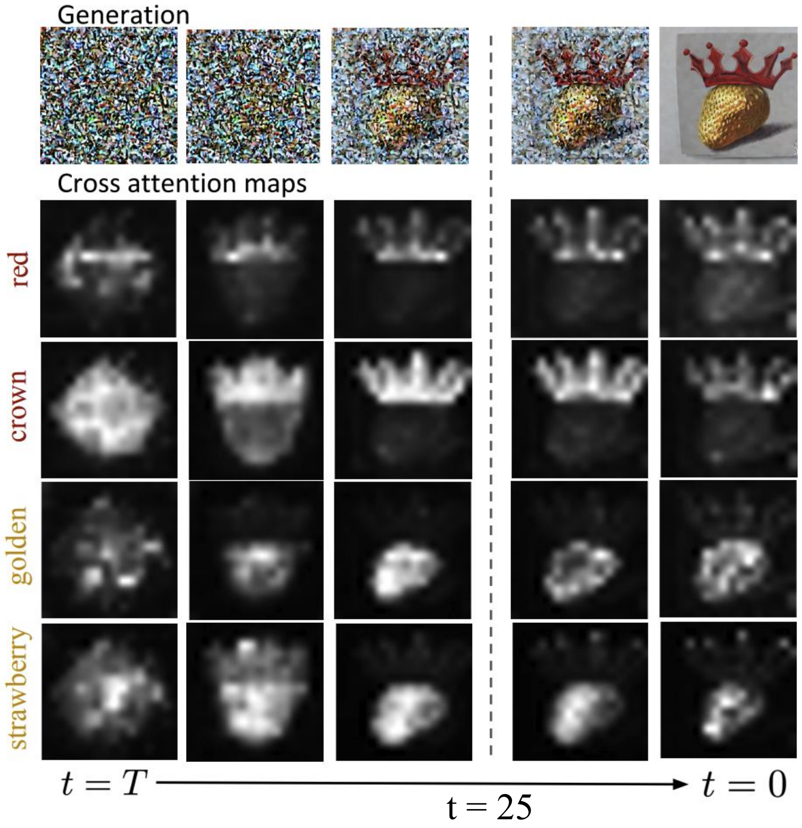
(b) Diffusion Process



SynGen | Evolution of Cross-attention Maps

Prompt

a <u>red</u> crown and a <u>golden</u> strawberry



"Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models" by Cheffer and Alaluf et al., 2023

Semantic Leak in Prompt

*"A yellow flamingo
and
a pink sunflower"*

Semantic Leak out of Prompt

*"A checkered bowl
in
a cluttered room"*

Attribute Neglect

*"A horned lion
and
a spotted monkey"*

Semantic Leak in Prompt

*"A yellow flamingo
and
a pink sunflower"*

Semantic Leak out of Prompt

*"A checkered bowl
in
a cluttered room"*

Attribute Neglect

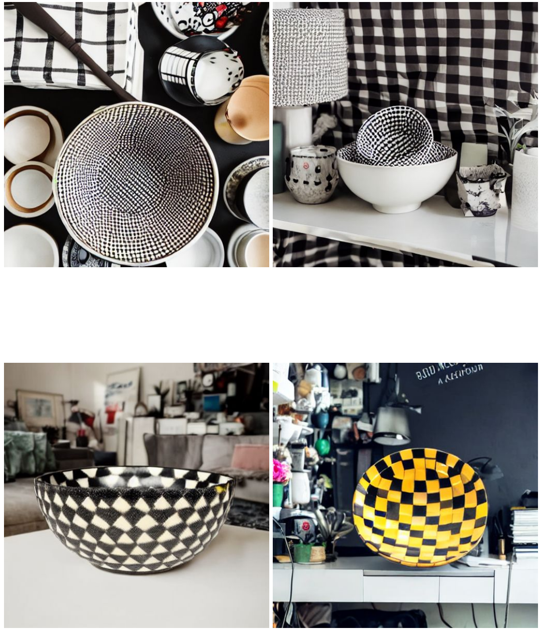
*"A horned lion
and
a spotted monkey"*

Semantic Leak in Prompt

*"A yellow flamingo
and
a pink sunflower"*

Semantic Leak out of Prompt

*"A checkered bowl
in
a cluttered room"*



Attribute Neglect

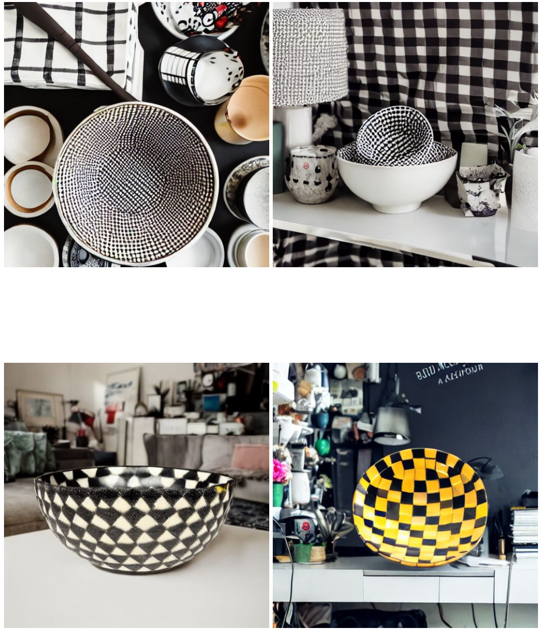
*"A horned lion
and
a spotted monkey"*

Semantic Leak in Prompt

*"A yellow flamingo
and
a pink sunflower"*

Semantic Leak out of Prompt

*"A checkered bowl
in
a cluttered room"*



Attribute Neglect

*"A horned lion
and
a spotted monkey"*

Experiments

We compare our method to **three baselines**

Experiments

We compare our method to **three baselines**

- Attend-and-Excite, StructureDiffusion, Stable Diffusion

Experiments

We compare our method to **three baselines**

- Attend-and-Excite, StructureDiffusion, Stable Diffusion
- Across **two existing** datasets and a **novel challenging one** by us

Experiments

We compare our method to three baselines

- Attend-and-Excite, StructureDiffusion, Stable Diffusion
- Across two existing datasets and a novel challenging one by us
- Using human raters on two metrics

Experiments | Datasets

	ABC-6K	Attend-and-Excite	DVMP (ours)
Key Challenges	* Subset of MSCOCO (human authored) * Contains contrastive examples	* Entities are objects or animals * Only colors as modifiers	* More objects and animals * Many types of modifiers * Much harder sentences
Format	Free-form text	A {color-1} {entity-1} and a {color-2} {entity-2}	A {modifier-1} … {entity-1} and a {modifier-2} … {entity-2} …
Examples	*A white fire hydrant sitting in a field next to a red building*	*A monkey and a black bow*	*a wooden crown and a furry baby rabbit and a pink metal bench*
# Examples	600	177	600

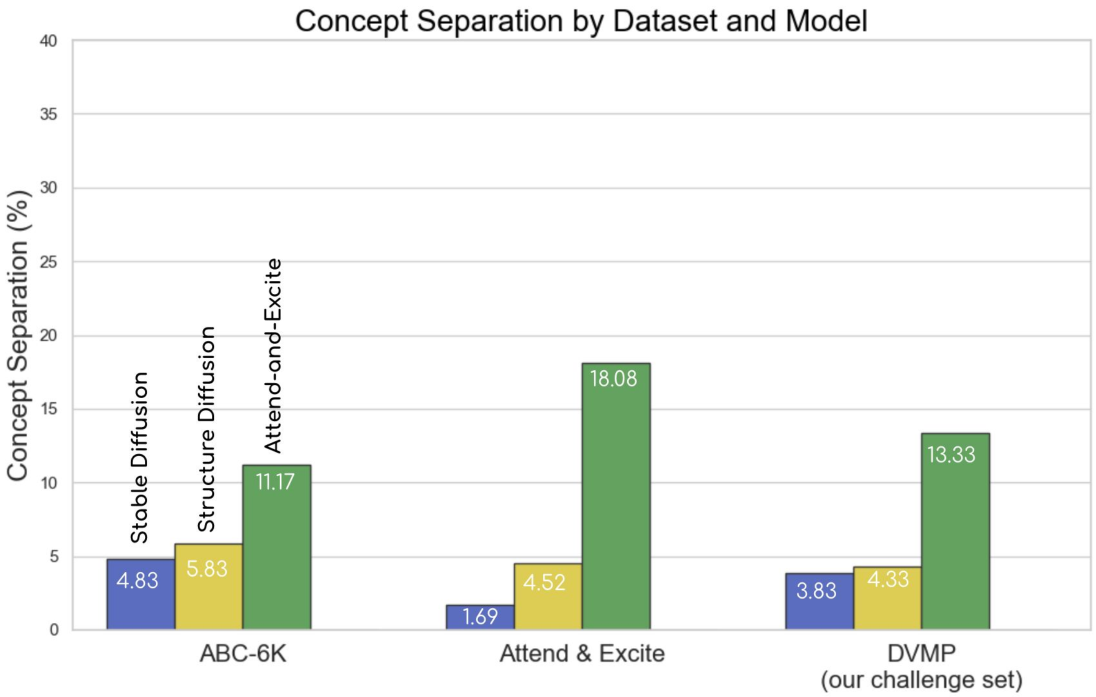
Experiments | Human Evaluation

- Concept Separation: "Which image best matches the description?"
- Visual Appeal: "Which image looks overall better or more natural?"
- Select a winning model or "no winner"

- Raters on Mechanical Turk
  - 3 raters
  - 100% on qualification test, ≥ 99% approval, ≥ 5000 HITs
- The majority decision was selected

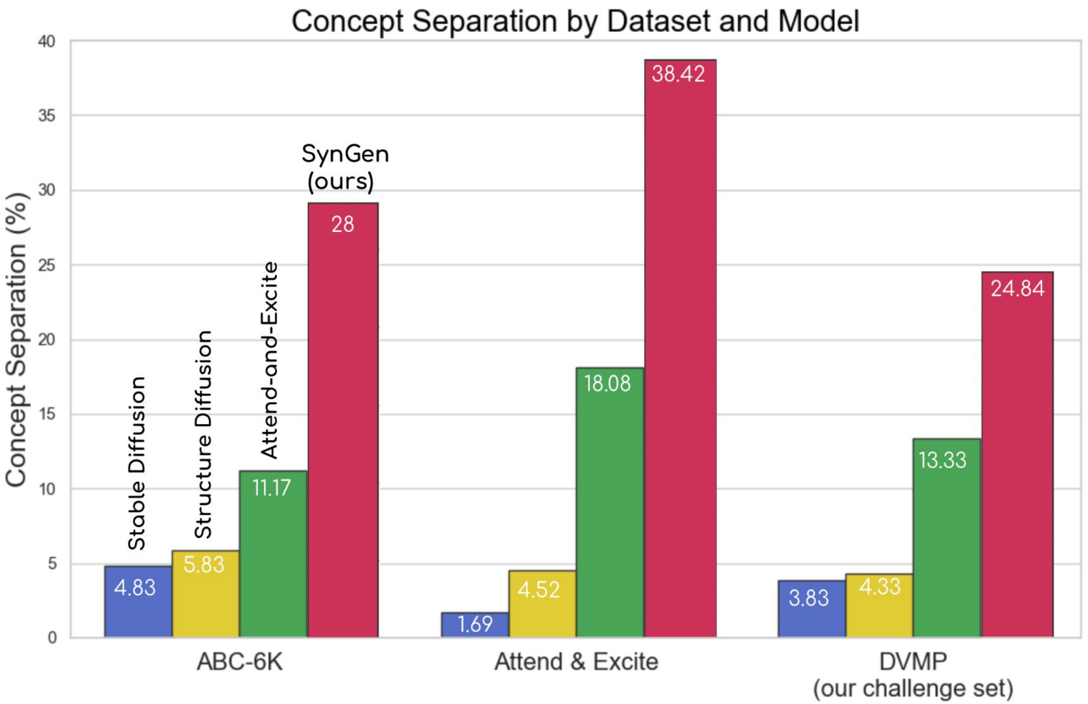
Results | Quantitative

"Which output best matches the prompt?"



Results | Quantitative

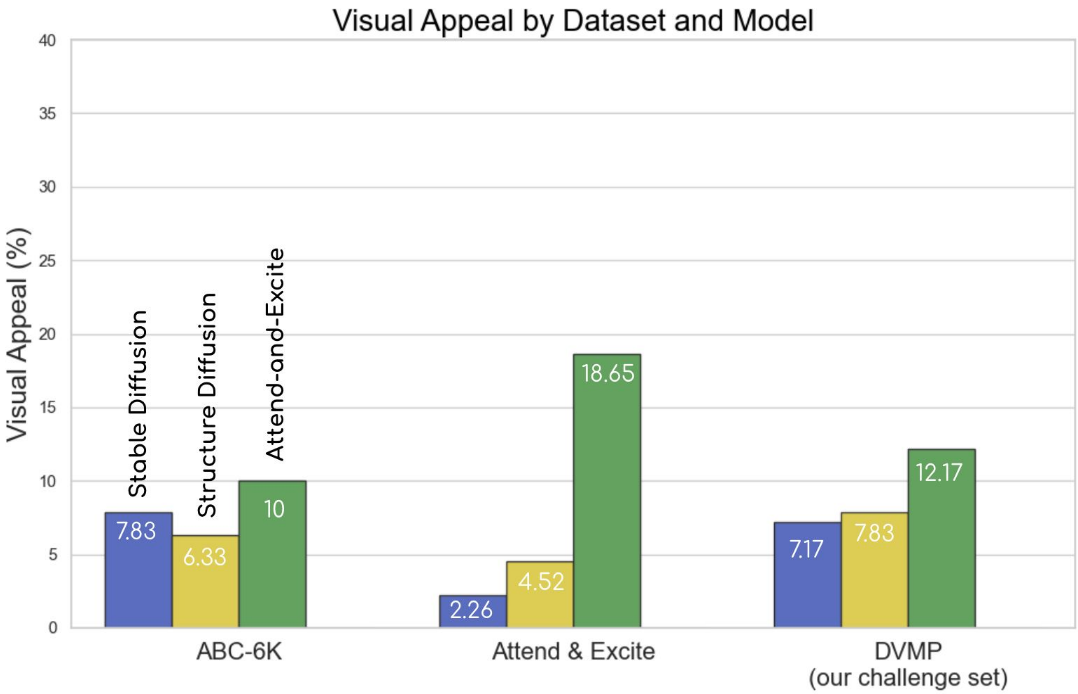
"Which output best matches the prompt?"



Concept Separation improvement by **117%** on average

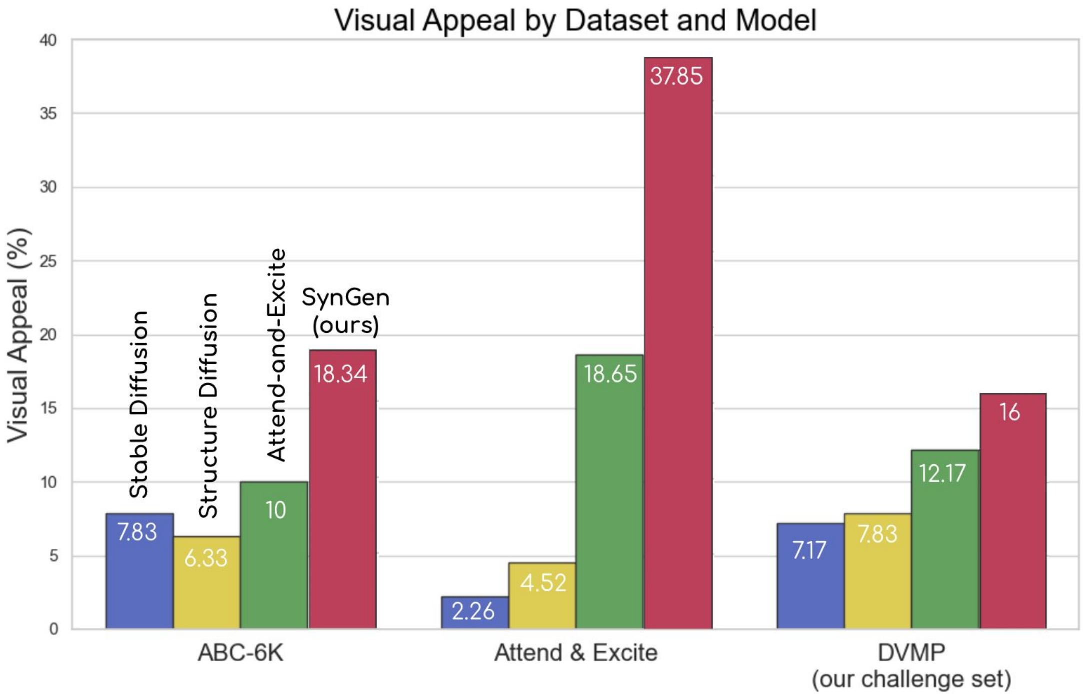
Results | Quantitative

"Which output looks best?"



Results | Quantitative

"Which output looks best?"



Visual Appeal improvement by **63%** on average

Conclusion

- We tackle improper binding, where visual interpretation doesn't match the prompt

- We propose SynGen, to improve image-text alignment
  - An inference-time method (no training or fine-tuning!)
  - Incorporates a linguistic-driven objective function to steer cross-attention
  - SOTA performance on all three datasets

Take SynGen for a ride!

Say hi @ poster sess!

Today 5:15-7:15 PM.

Great Hall & B1 + B2
Poster No. 615

Paper

Demo

Thank you!

@RoyiRassin

rassinroyi@gmail.com