Layout guidance related paper discussion

Capstone meeting 7.23.2023

CHEN ZEYU

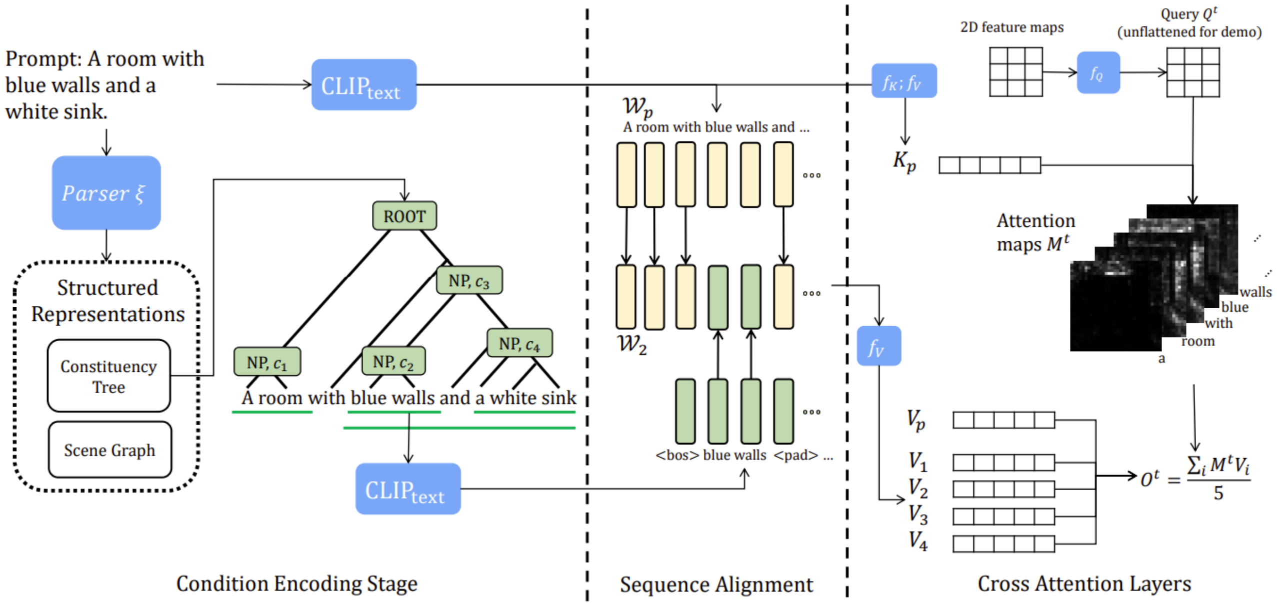
Attention map editing methods: Structure diffusion

- Focus on attribute binding problem
- Attention maps control layout structure and value matrix V controls object semantics mapped into attended regions
- Embed each noun phrase (np) separately to form multiple value V1,···Vk to address contextualization problem of CLIP (i.e. tokens in the later part of a sequence are blended with the token semantics before them)

A yellow apple and a red banana
A yellow apple / a red banana

CLIP

Prompt embedding
Individual np embedding



Original attention map:

Use the whole text prompt to generate value matrix V_p.

The output cross-attention map $a = AV_p$ where A is the attention score map.

Adjusted attention map:

Use individual nps to generate value matrix $V_1, V_2, \cdots, V_k$.

The output attention map $a = \frac{1}{k+1} \sum_i AV_i, i = p, 1, 2, \cdots, k.$

Training-Free Structured Diffusion Guidance Compositional Text-to-Image Synthesis
https://weixi-feng.github.io/structure-diffusion-guidance/

Attention map editing methods: Structure diffusion

Stable diffusion Structure diffusion

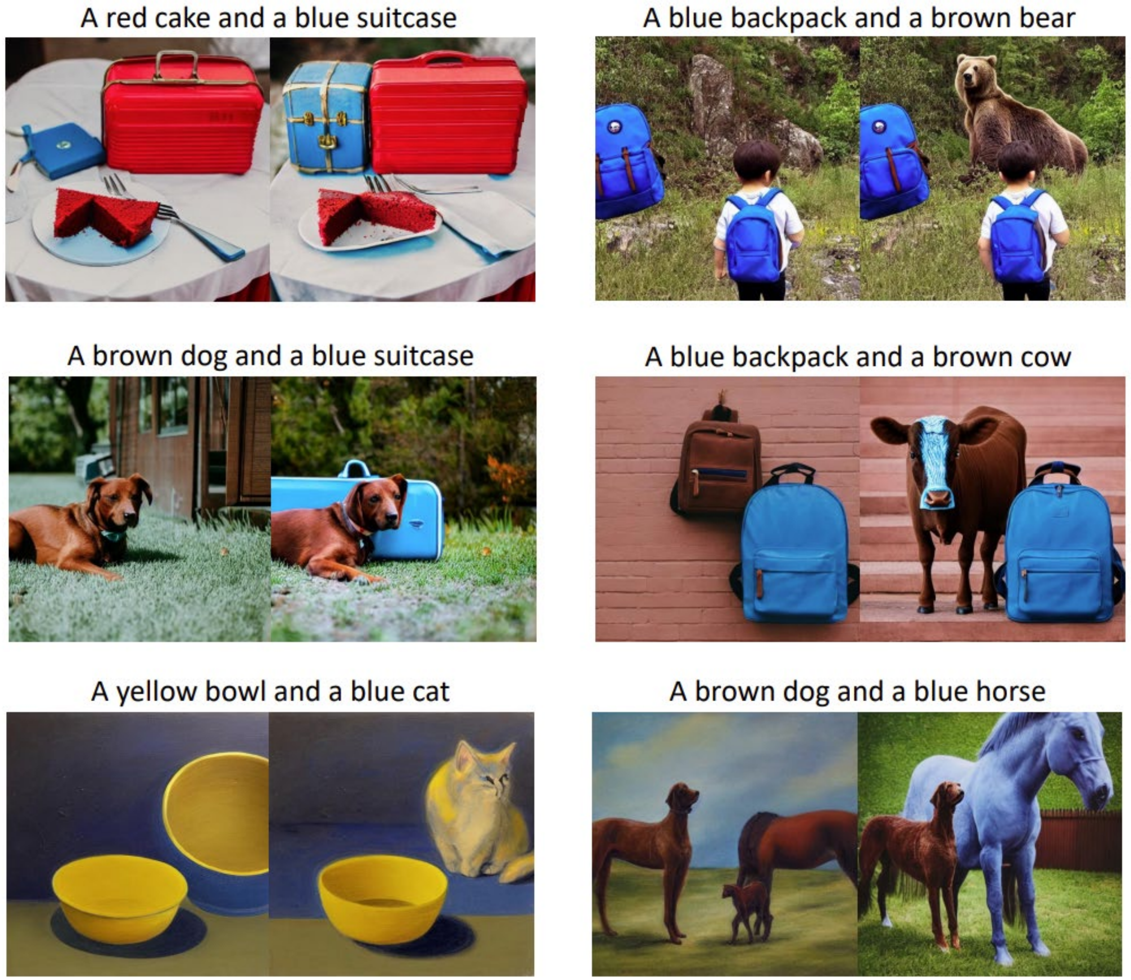


Figure 13: Qualitative results on CC-500

- Evaluation dataset:
  - Attribute binding contrast (ABC-6k): prompts from MSCOCO with contrast caption
  - Concept Conjunction 500 (CC-500): two objects conjunction
- Evaluation metrics:
  - Human evaluation
  - GLIP (phrase grounding prediction model)
  - Percentage of incomplete/ complete / complete with correct attribute images

	CC-500 (Prompt format: "*a* [*colorA*] [*objectA*] and *a* [*colorB*] [*objectB*]")					
	Human Annotations			GLIP		
Methods	Zero/One obj. (↓)	Two obj.	Two obj. w/ correct colors	Zero/One obj. (↓)	Two obj.	Human-GLIP Consistency
Stable Diffusion	65.5	34.5	19.2	69.0	31.0	46.4
Composable Diffusion	69.7	30.3	20.6	74.2	25.8	48.9
StructureDiffusion (Ours)	**62.0**	**38.0**	**22.7**	**68.8**	**31.2**	47.6

Attention map editing methods: Attend and Excite

- Focus on neglect object
- Design a loss function to strengthen the attention of the most neglected tokens. (similar to backward guidance in layout guidance paper)

Algorithm 1 A Single Denoising Step using Attend-and-Excite

Input: A text prompt $\mathcal{P}$, a set of subject token indices $\mathcal{S}$, a timestep t, a set of iterations for refinement $\{t_1, \ldots, t_k\}$, a set of thresholds $\{T_1, \ldots, T_k\}$, and a trained Stable Diffusion model SD.
Output: A noised latent z_{t-1} for the next timestep

1: $\_, A_t \leftarrow SD(z_t, \mathcal{P}, t)$
2: $A_t \leftarrow \text{Softmax}(A_t - \langle sot \rangle)$
3: **for** $s \in \mathcal{S}$ **do**
4: $\quad A_t^s \leftarrow A_t[:, :, s]$
5: $\quad A_t^s \leftarrow \text{Gaussian}(A_t^s)$
6: $\quad \mathcal{L}_s \leftarrow 1 - \max(A_t^s)$
7: **end for**
8: $\mathcal{L} \leftarrow \max_s(\mathcal{L}_s)$
9: $z_t' \leftarrow z_t - \alpha_t \cdot \nabla_{z_t} \mathcal{L}$
10: **if** $t \in \{t_1, \ldots, t_k\}$ **then** ▷ If performing iterative refinement at t
11: $\quad$ **if** $\mathcal{L} > 1 - T_t$ **then**
12: $\quad\quad z_t \leftarrow z_t'$
13: $\quad\quad$ **Go to** Step 1
14: $\quad$ **end if**
15: **end if**
16: $z_{t-1}, \_ \leftarrow SD(z_t', \mathcal{P}, t)$
17: **Return** z_{t-1}

Loss function design:

$$L = max_i L_i \quad L_i = 1 - max_u A_{ui}$$

Encourage z_t to generate along the direction of increasing $min_i max_u A_{ui}$.

Find the minimum of the highest attention score of all tokens and try to increase this value during sampling.

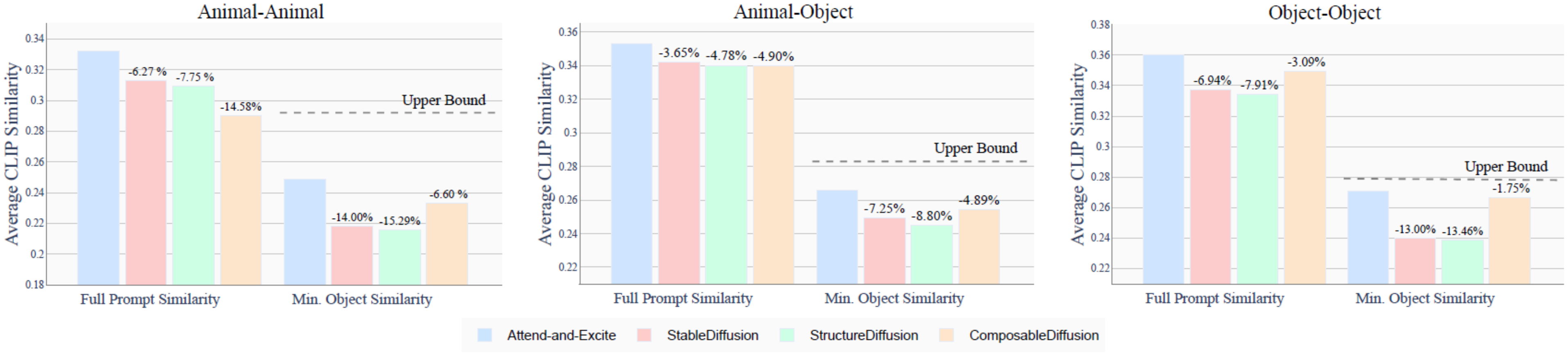
Ignore the [SOT] attention score map

Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models
https://yuval-alaluf.github.io/Attend-and-Excite/

Attention map editing methods: Attend and Excite

Evaluation:

- CLIP similarity:
  - Image & text: full prompt / sub prompt with generated image
  - Text & text: predicted caption using generated image & original text prompt
- User study



Attention map editing methods: Directed diffusion

- Re-weighting attention score maps on corresponding token and padding tokens (similar to forward guidance in layout guidance) risk for overly aggressive guidance

$$D_{ui} = A_{ui} \cdot W + S,\ W = \begin{cases} 1 & u \in B \\ c < 1 & u \notin B' \end{cases} \quad S = \begin{cases} gaussian \quad f(u) & u \in B \\ 0 & u \notin B' \end{cases} \quad i \in target\ token \cup padding\ tokens$$

W: weaken mask S: strengthen mask

- Find the best weighted combination of padding token maps s.t. target token maps Ai best match desired map Di

Find adjustment weights $a_t^* \in \mathbb{R}^{77-|P|-1}$ to minimize $L_{a_t} = \sum_i ||A_i^{t-1}(Diag(a_t) \cdot A^t_{|P|+1:77}) - D_i||^2, \quad i \in target\ token.$

Only change padding tokens attention maps instead of changing target token maps directly: $A^t_{|P|+1:77} = Diag(a_t^*) \cdot D^t_{|P|+1:77}.$

Change attention of padding maps and see their impact on the next iteration attention maps for target token

Directed Diffusion: Direct Control of Object Placement through Attention Guidance
https://hohonu-vicml.github.io/DirectedDiffusion.Page/

Attention map editing methods: Directed diffusion

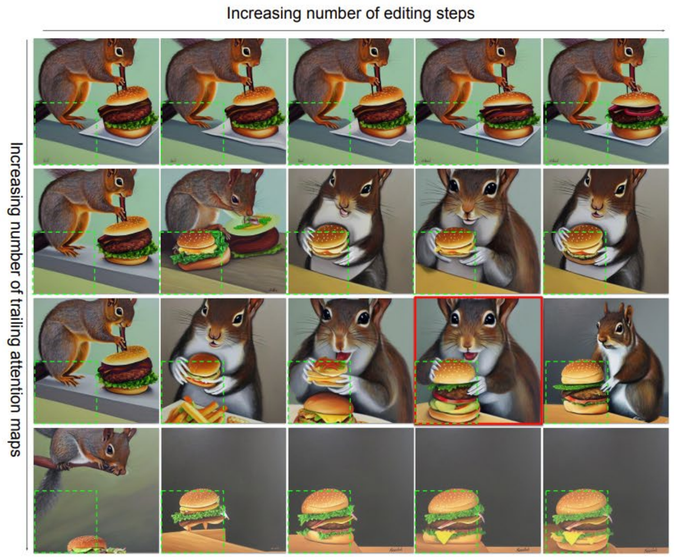


Figure 11: The number of trailing attention maps (5, 10, 15, 20, maps on the vertical axis) versus the number of attention map editing steps (1, 3, 5, 10, and 15 steps on the horizontal axis). The prompt is *"A photo of a squirrel eating a burger"* with the directed object "burger" positioned at the bottom left. The best results are obtained with an intermediate number of editing steps and edited trailing attention maps, as in the case of the image with the red border

- An interesting application: placement finetuning

- Impact of adjusting padding tokens attention maps

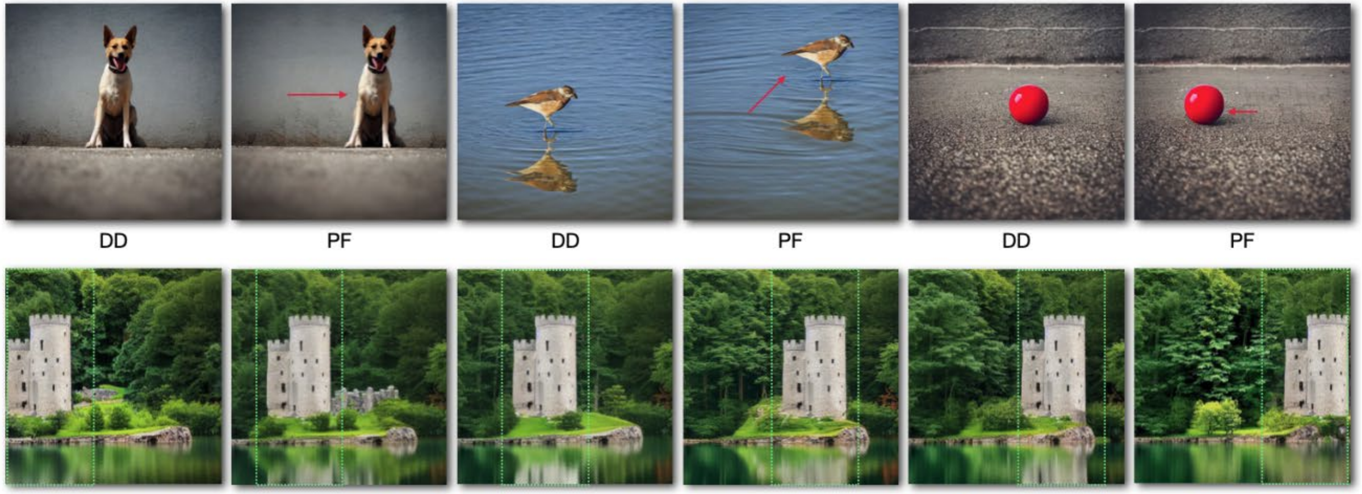


Figure 9: (Top) placement finetuning (PF) allows the position of an object to be changed while largely preserving the object identity and existing background, and without requiring network optimization or fine-tuning. (Bottom) PF is used to explore different locations for a desired castle. Compare to Fig. 7, top, where repositioning the DD bounding box results in a different castle at each location.

Summary

- Structure diffusion and Attend & Excite don’t need additional bounding boxes. They mainly focus on two major problems in compositional image generation: neglect objects and attribute binding rather than layout control.

- Layout control can also help to alleviate such problems. But layout guidance doesn’t provide analysis on attribute binding.

- Role of [SOT] and padding tokens attention maps is still not clear.

- Evaluation of compositionality is mainly based on human inspection.

Gradient guidance methods

Classifier guidance:

$$\epsilon_\theta(z_t, t) \leftarrow \epsilon_\theta(z_t, t) - \sqrt{1-\alpha_t}\nabla_{z_t} \log p(c|z_t)$$

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(z_t, t)\right) + \sigma_t z$$

Encourage z_t to generate along the direction of increasing classification probability.

Additional prompt

Extension:

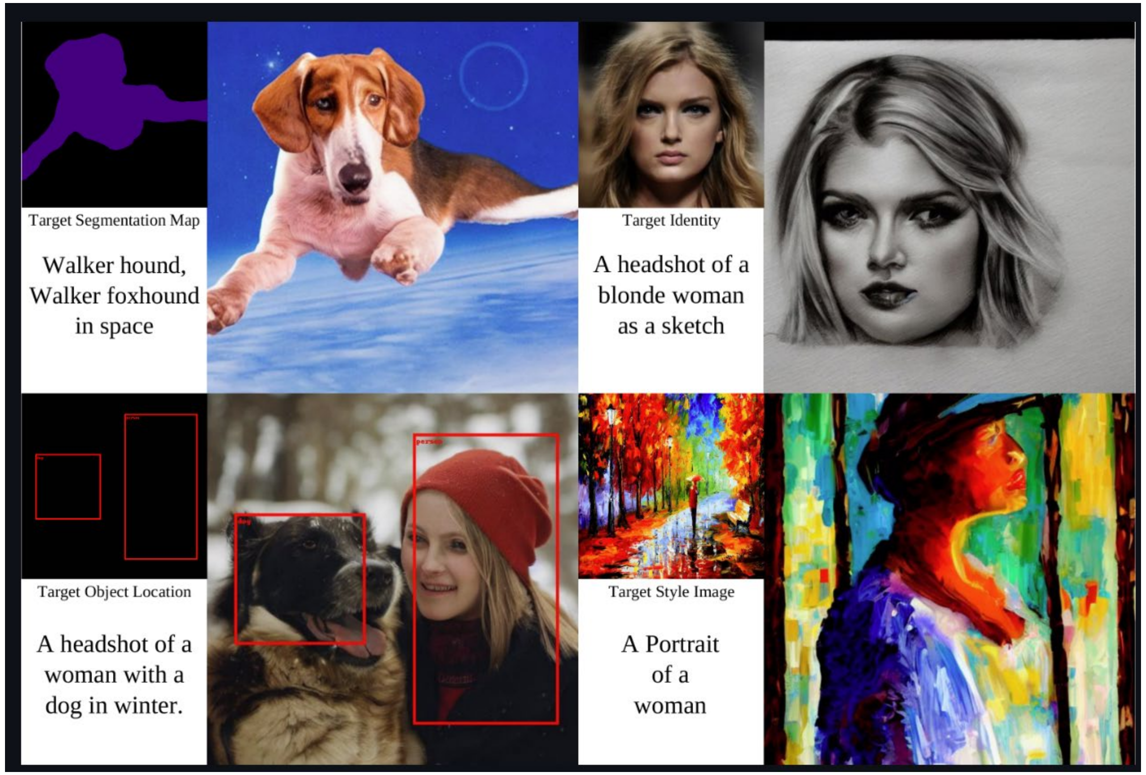
$$\epsilon_\theta(z_t, t) \leftarrow \epsilon_\theta(z_t, t) + \sqrt{1-\alpha_t}\nabla_{z_t} l(c, f(z_t))$$

Loss function to measure the distance between guided sample and given prompt

Guidance function

- Use a text prompt c to generate an image
  - f: CLIP image encoder
  - l: dot product similarity
- Use a bounding box prompt c to generate an image
  - f: object detector
  - l: bounding box regression loss and classification loss
- Use a segmentation prompt c to generate an image
  - f: segmentation network
  - l: per-pixel cross entropy loss

Gradient guidance methods



Universal Guidance for Diffusion Models
https://github.com/arpitbansal297/Universal-Guided-Diffusion

More Control for Free! Image Synthesis with Semantic Diffusion Guidance
https://xh-liu.github.io/sdg/

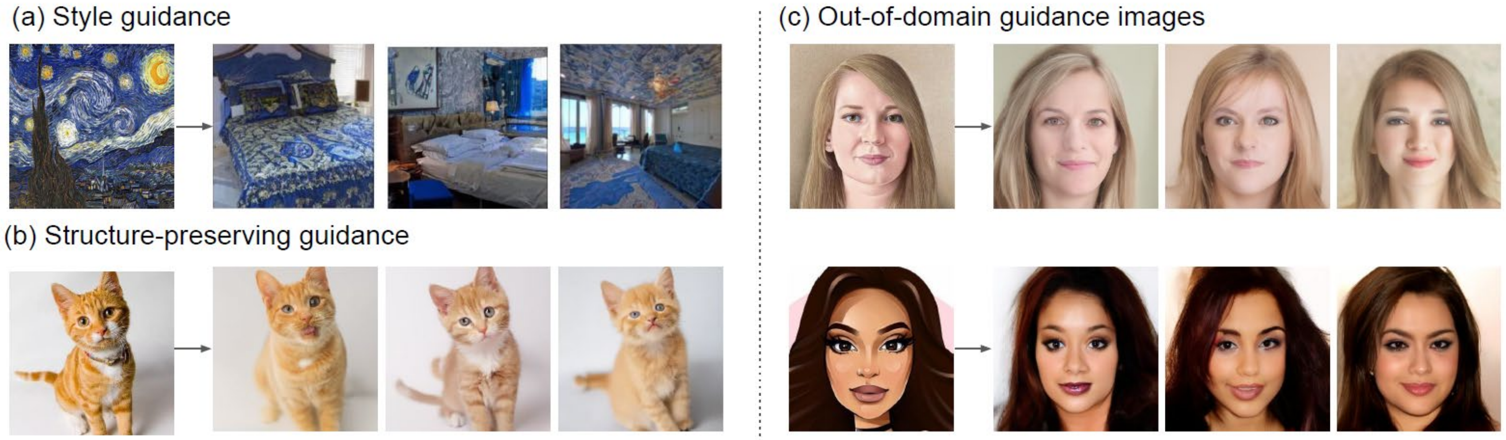
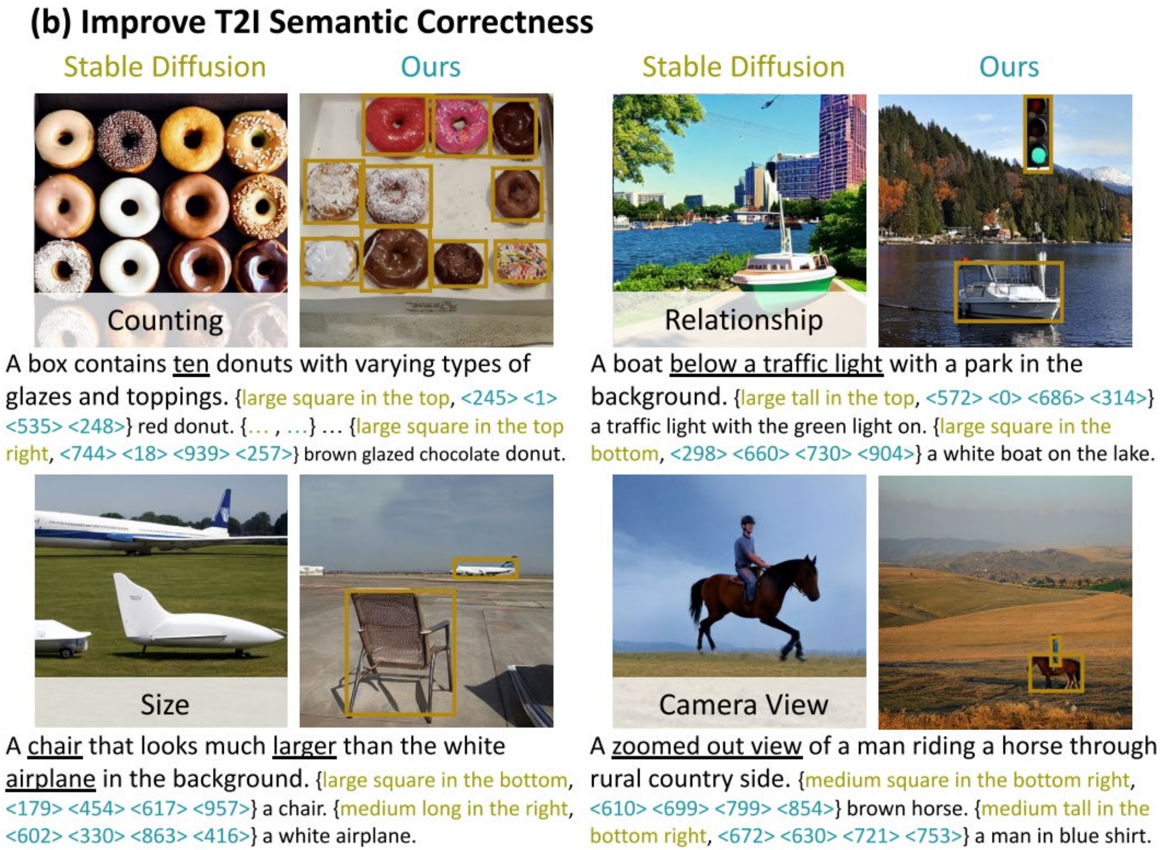


Figure 7: Different applications of SDG. (a) Style-guided synthesis. (b) Structure-preserving synthesis when the user does not want to generate diverse structures. (c) Synthesizing photo-realistic images with out-of-domain image guidance.

Summary

- Conduct other guidance rather than bounding box guidance like segmentation guidance through attention map editing
- More detailed text prompt for objects inside bounding boxes to increase compositionality



ReCo: Region-Controlled Text-to-Image Generation
https://github.com/microsoft/ReCo