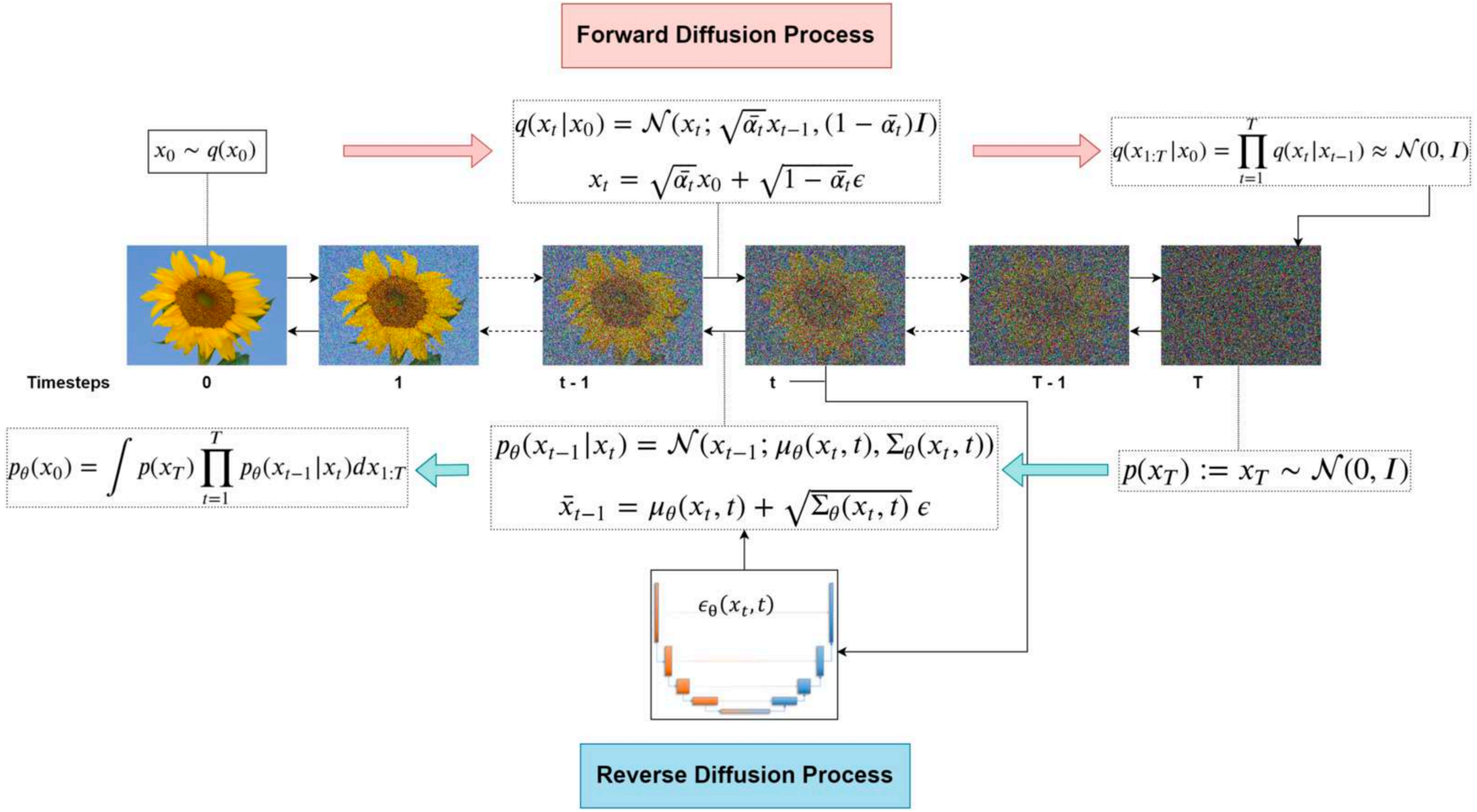
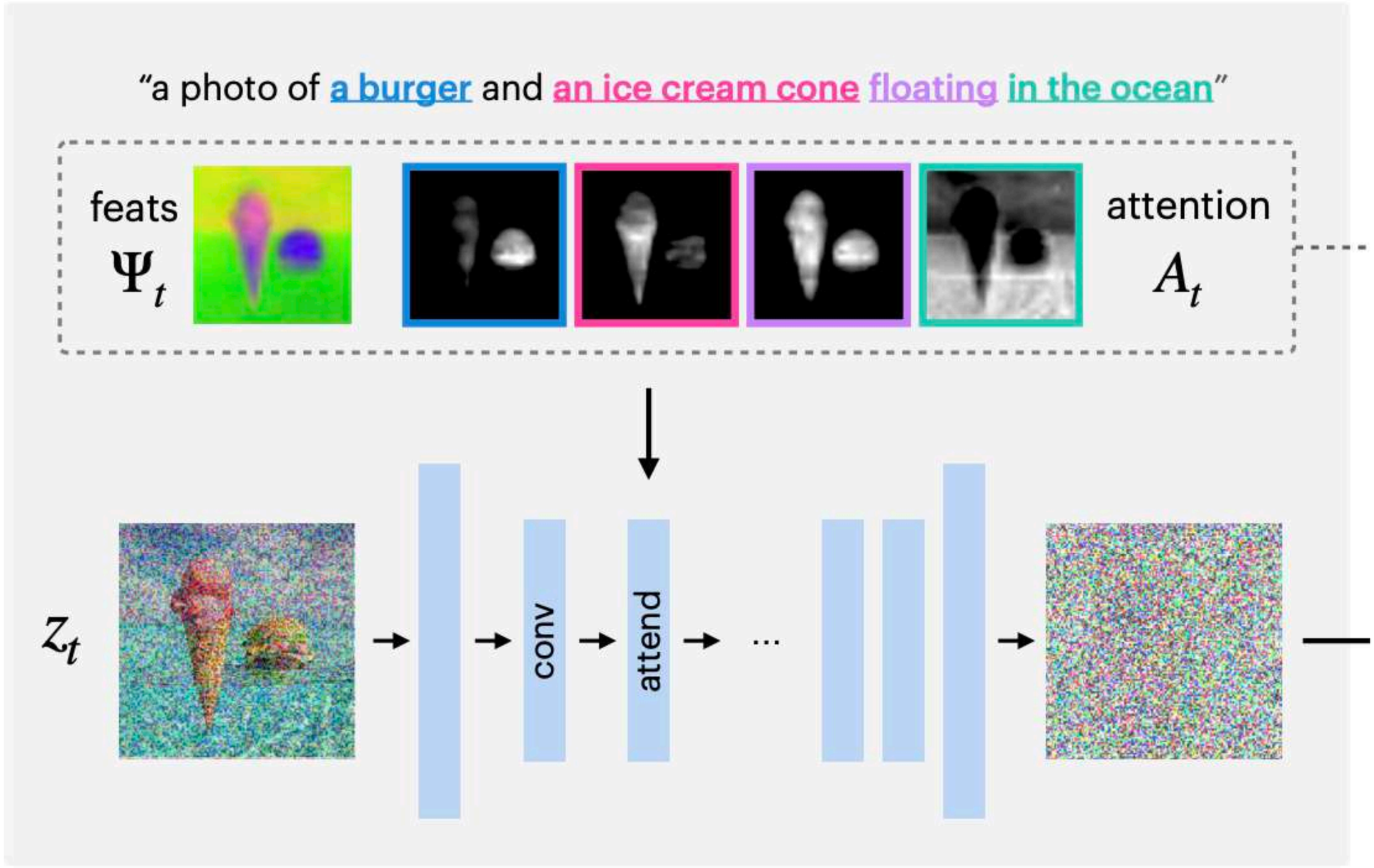
Diffusion Self-Guidance for Controllable Image Generation

Кокоева Мария

Диффузионные модели




"a photo of a burger and an ice cream cone floating in the ocean"
feats
Ψ_t
attention
A_t
z_t
conv
attend
...

Guidance

$$\hat{\epsilon}_t = \epsilon_\theta(z_t; t, y) - s\sigma_t \nabla_{z_t} \log p(y|z_t)$$

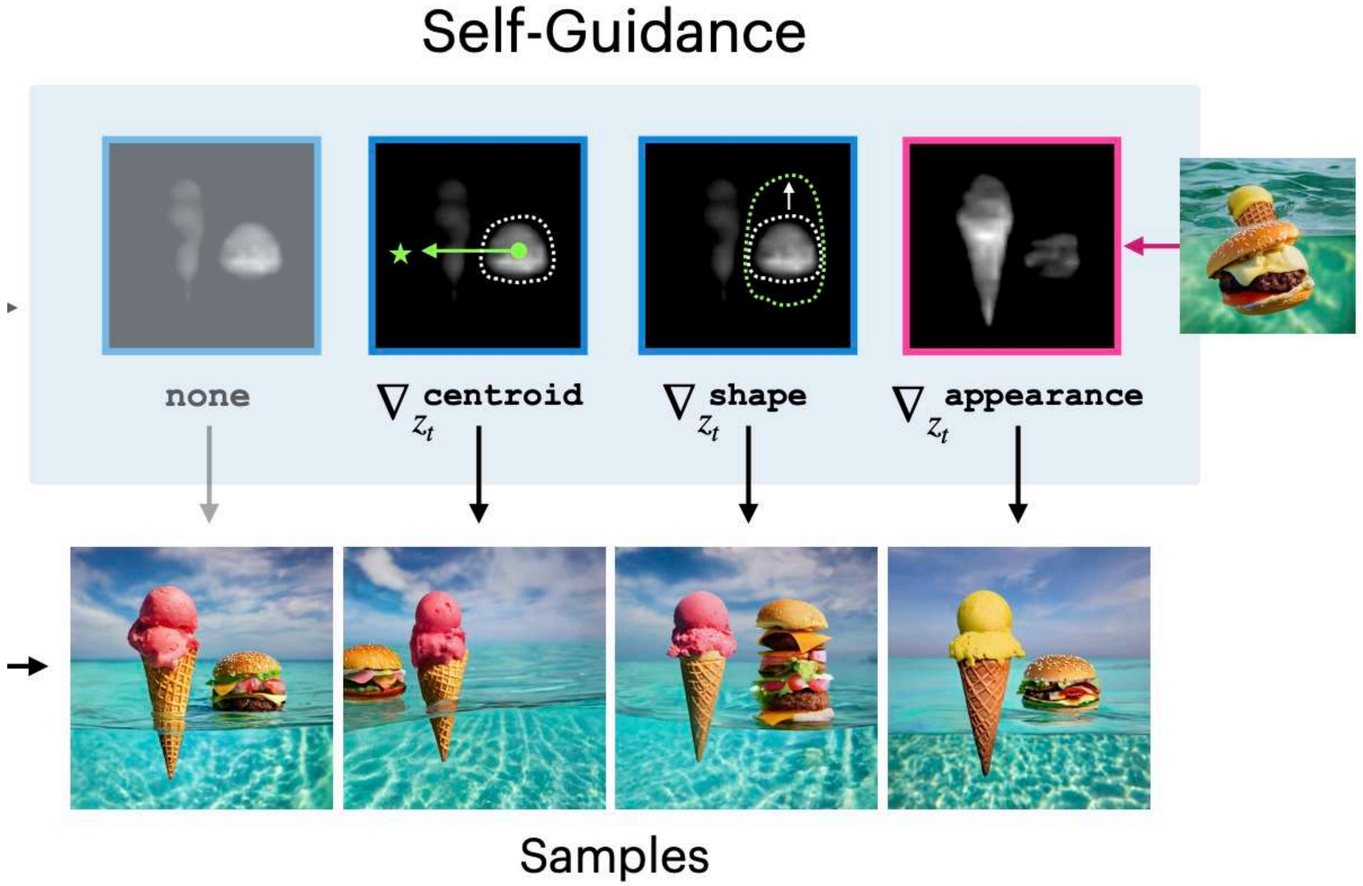
$$\hat{\epsilon}_t = (1+s)\epsilon_\theta(z_t; t, y) - s\epsilon_\theta(z_t; t, \emptyset) + v\sigma_t \nabla_{z_t} g(z_t; t, y)$$

$$\texttt{centroid}\,(k) = \frac{1}{\sum_{h,w} \mathcal{A}_{h,w,k}} \begin{bmatrix} \sum_{h,w} w \cdot \mathcal{A}_{h,w,k} \\ \sum_{h,w} h \cdot \mathcal{A}_{h,w,k} \end{bmatrix}$$

$$\texttt{size}\,(k) = \frac{1}{HW} \sum_{h,w} \mathcal{A}_{h,w,k}$$

$$\texttt{shape}(k) = \mathcal{A}_k^{\text{thresh}}$$

$$\texttt{appearance}(k) = \frac{\sum_{h,w} \texttt{shape}(k) \odot \Psi}{\sum_{h,w} \texttt{shape}(k)}$$



$$g = w_0 \overbrace{\frac{1}{|O|-1} \sum_{o \neq o_k \in O} \frac{1}{|\mathcal{A}|} \sum_{i=0}^{|\mathcal{A}|} \|\mathtt{shape}_{i,t,\mathrm{orig}}(o) - \mathtt{shape}_{i,t}(o)\|_1}^{\text{Fix all other object shapes}}$$

$$+ w_1 \overbrace{\frac{1}{|O|} \sum_{o \in O} \|\mathtt{appearance}_{t,\mathrm{orig}}(o) - \mathtt{appearance}_{t}(o)\|_1}^{\text{Fix all appearances}}$$

$$+ w_2 \overbrace{\frac{1}{|\mathcal{A}|} \sum_{i=0}^{|\mathcal{A}|} \|\mathcal{T}\left(\mathtt{shape}_{i,t,\mathrm{orig}}(o_k)\right) - \mathtt{shape}_{i,t}(o_k)\|_1}^{\text{Guide } o_k\text{'s shape to translated original shape}}$$

(a) Original (b) Move up (c) Move down (d) Move left (e) Move right (f) Shrink (g) Enlarge

$$g = w_0 \overbrace{\frac{1}{|O|} \sum_{o \in O} \frac{1}{|\mathcal{A}|} \sum_{i=0}^{|\mathcal{A}|}}^{\text{Fix all object shapes}} \|\mathbf{shape}_{i,t,\text{orig}}(o) - \mathbf{shape}_{i,t}(o)\|_1$$

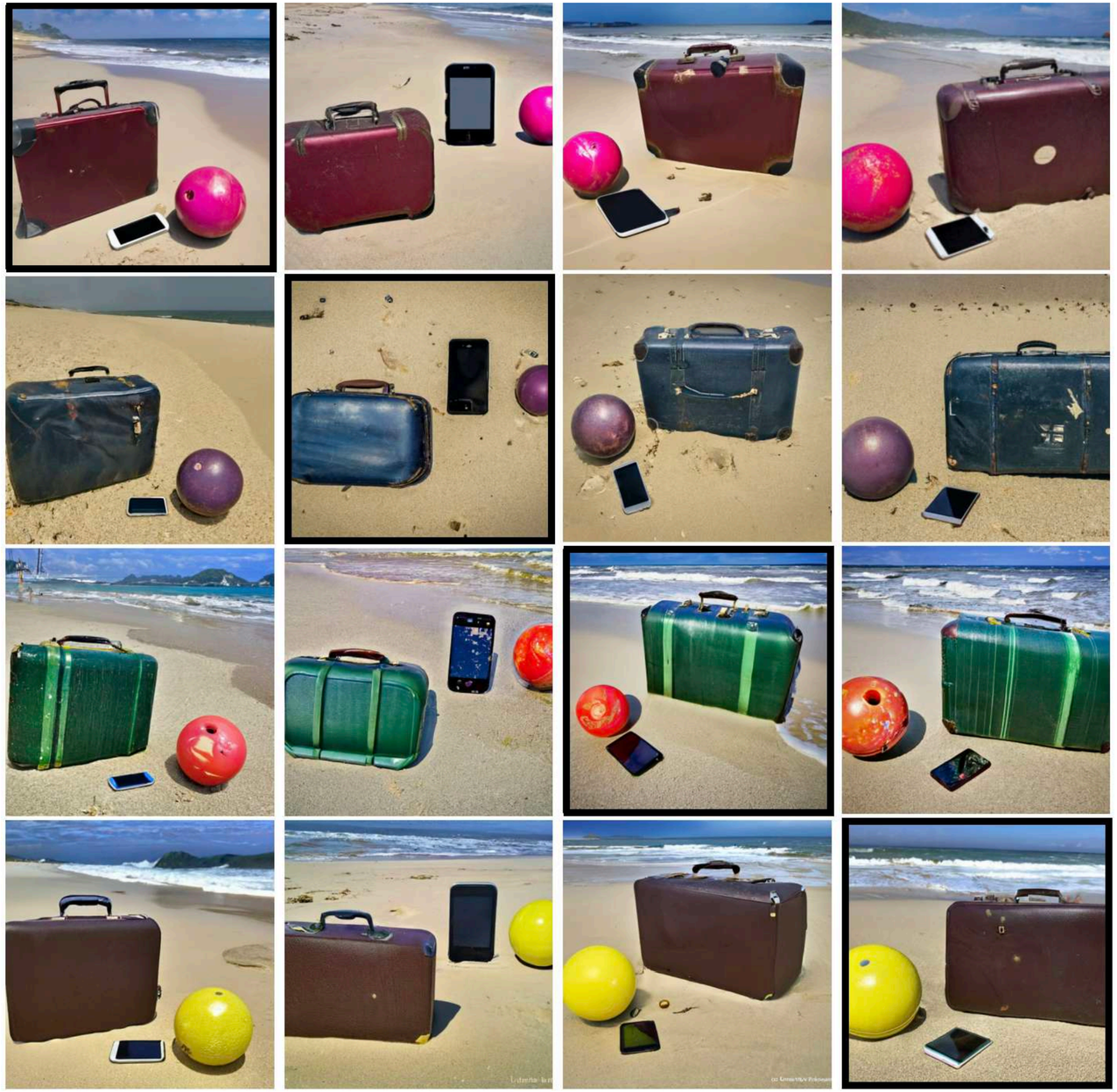
(a) Original **(b) New appearances** (c) ControlNet [39] (d) PtP [11]

Copy object shapes from A

$$g = w_0 \overbrace{\frac{1}{|O|} \sum_{o \in O} \frac{1}{|\mathcal{A}|} \sum_{i=0}^{|\mathcal{A}|} \|\texttt{shape}_{i,t,A}(o) - \texttt{shape}_{i,t}(o)\|_1}$$

Copy object appearance from B

$$+ w_1 \overbrace{\frac{1}{|O|} \sum_{o \in O} \|\texttt{appearance}_{t,B}(o) - \texttt{appearance}_{t}(o)\|_1}$$



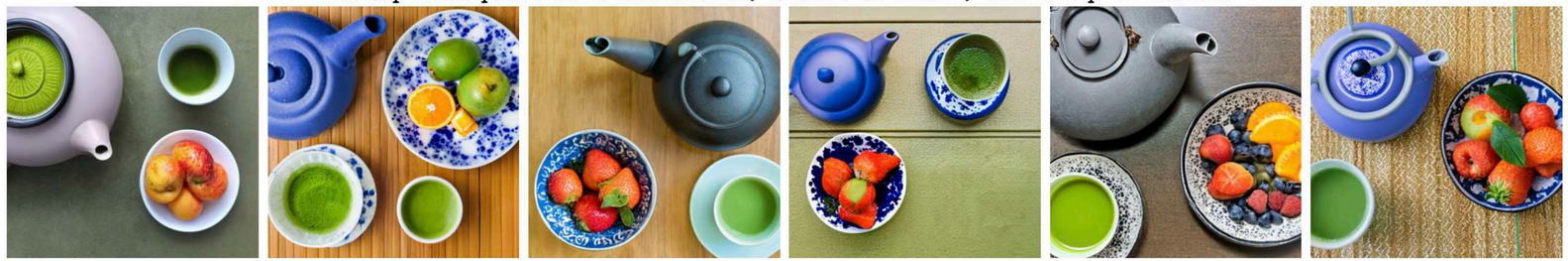
$$g = w_0 \overbrace{\frac{1}{J}\sum_j \frac{1}{|\mathcal{A}|}\sum_{i=0}^{|\mathcal{A}|} \|\texttt{shape}_{i,t,j}(o_{k_j}) - \texttt{shape}_{i,t}(o_k)\|_1}^{\text{Copy each object's shape, position, and size}}$$

$$+ w_1 \overbrace{\frac{1}{J}\sum_j \|\texttt{appearance}_{t,j}(o_{k_j}) - \texttt{appearance}_t(o_k)\|_1}^{\text{Copy each object's appearance}}$$

"a photo of a picnic blanket, a fruit tree, and a car by the lake"

(a) Take `blanket` (b) Take `tree` (c) Take `car` (d) **Result** (e) + Target layout (f) **Final result**

"a top-down photo of a tea kettle, a bowl of fruit, and a cup of matcha"



(a) Take `matcha` (b) Take `kettle` (c) Take `fruit` (d) **Result** (e) + Target layout (f) **Final result**

"a photo of a dog wearing a knit sweater and a baseball cap drinking a cocktail"

(a) Take `sweater` (b) Take `cocktail` (c) Take `cap` (d) **Result*** (e) + Target layout (f) **Final result**

"a photo of a hot dog, fries, and a soda on a solid background"

(a) Real image (b) Reconstruct (c) Swap w. fries (d) Width ↓ (e) Width ↓, height ↑ (f) Restyle

"a photo of an eclair and a shot of espresso"

(g) Real image (h) Reconstruct (i) Move (j) Width ↓ (k) Width, height ↑ (l) Restyle

(a) Original (b) Ours (b) Random samples without self-guidance