Algorithm 1 Conditional EBM Training Algorithm

Input: data dist $p_D(\boldsymbol{x})$, relational scene descriptions $R_D(\boldsymbol{r})$, step size λ, number of steps K, data augmentation $D(\cdot)$, stop gradient operator $\Omega(\cdot)$, EBM $E_\theta(\cdot)$, Encoder $\mathrm{Enc}(\cdot)$, Parser $P(\cdot)$
$\mathcal{B} \leftarrow \varnothing$
while not converged **do**
  $\boldsymbol{x}_i^+ \sim p_D$
  $R_i \sim R_D$
  $\tilde{\boldsymbol{x}}_i^0 \sim \mathcal{B}$ with 99% probability and $\mathcal{U}$ otherwise
  $X \sim \mathcal{B}$ for nearest neighbor entropy calculation

  ▷ *Parse a relational scene description:*
  $\{\boldsymbol{r}_1, \dots \boldsymbol{r}_m\} \leftarrow P(R_i)$

  ▷ *Apply data augmentation to sample:*
  $\tilde{\boldsymbol{x}}_i^0 = D(\tilde{\boldsymbol{x}}_i^0)$

  ▷ *Generate sample using Langevin dynamics:*
  for sample step $k = 1$ to K **do**
    $\tilde{\boldsymbol{x}}_i^{k-1} = \Omega(\tilde{\boldsymbol{x}}_i^{k-1})$
    $\tilde{\boldsymbol{x}}^k \leftarrow \tilde{\boldsymbol{x}}^{k-1} - \nabla_{\boldsymbol{x}} \sum_{j=1}^m E_\theta(\tilde{\boldsymbol{x}}^{k-1} \mid \mathrm{Enc}(\boldsymbol{r}_j)) + \omega, \;\; \omega \sim \mathcal{N}(0, \sigma)$
  end for

  ▷ *Generate two variants of $\boldsymbol{x}^-$ with and without gradient propagation:*
  $\boldsymbol{x}_i^- = \Omega(\tilde{\boldsymbol{x}}_i^k)$
  $\hat{\boldsymbol{x}}_i^- = \tilde{\boldsymbol{x}}_i^k$

  ▷ *Optimize objective $\mathcal{L}_{CD} + \mathcal{L}_{KL}$ wrt θ:*
  $\mathcal{L}_{\mathrm{CD}} = \frac{1}{N} \sum_i \sum_{j=1}^m (E_\theta(\boldsymbol{x}_i^+ \mid \mathrm{Enc}(\boldsymbol{r}_j) - E_\theta(\boldsymbol{x}_i^- \mid \mathrm{Enc}(\boldsymbol{r}_j))$
  $\mathcal{L}_{\mathrm{KL}} = \sum_{j=1}^m E_{\Omega(\theta)}(\hat{\boldsymbol{x}}_i^- \mid \mathrm{Enc}(\boldsymbol{r}_j)) - \log(NN(\hat{\boldsymbol{x}}_i^-, X)$

  ▷ *Optimize objective $\mathcal{L}_{CD} + \mathcal{L}_{KL}$ wrt θ:*
  $\Delta\theta \leftarrow \nabla_\theta(\mathcal{L}_{\mathrm{CD}} + \mathcal{L}_{\mathrm{KL}})$
  Update θ based on $\Delta\theta$ using Adam optimizer

  ▷ *Update replay buffer $\mathcal{B}$*
  $\mathcal{B} \leftarrow \mathcal{B} \cup \tilde{\boldsymbol{x}}_i^-$
end while

Algorithm 2 Image-to-text Retrieval

Input: input image $\boldsymbol{x}$, relational scene descriptions $\{R_1, \dots, R_n\}$, EBM $E_\theta(\cdot)$, Parser $P(\cdot)$, Encoder $\mathrm{Enc}(\cdot)$, output energy list $\mathcal{O}$, caption prediction $\mathcal{C}$
$\mathcal{O} \leftarrow []$

▷ *Generate image-caption matching energies iteratively*
for number of scene relations descriptions $i = 1$ to n **do**
  ▷ *Parse a relational scene description:*
  $\{\boldsymbol{r}_1, \dots \boldsymbol{r}_m\} \leftarrow P(R_i)$
  $\boldsymbol{e}_i = \sum_{j=1}^m E_\theta(\boldsymbol{x} \mid \mathrm{Enc}(\boldsymbol{r}_j))$

  ▷ *output energy list $\mathcal{O}$*
  $\mathcal{O}.append(\boldsymbol{e}_i)$
end for

▷ *Final output:*
$\mathcal{C} = \arg\min \mathcal{O}$

Algorithm 3 Image Generation Algorithm

Input: Relational scene description R, number of data augmentation applications N, step size λ, number of steps K, data augmentation $D(\cdot)$, EBM $E_\theta(\cdot)$, Parser $P(\cdot)$, Encoder $\mathrm{Enc}(\cdot)$
$\tilde{\boldsymbol{x}}^0 \sim \mathcal{U}$

▷ *Parse a relational scene description:*
$\{\boldsymbol{r}_1, \dots \boldsymbol{r}_m\} \leftarrow P(R)$

▷ *Generate samples through N iterative steps of data augmentation/Langevin dynamics:*
for sample step $n = 1$ to N **do**
  ▷ *Apply data augmentation to samples:*
  $\tilde{\boldsymbol{x}}^0 = D(\tilde{\boldsymbol{x}}_i^0)$

  ▷ *Run K steps of Langevin dynamics:*
  for sample step $k = 1$ to K **do**
    $\tilde{\boldsymbol{x}}^k \leftarrow \tilde{\boldsymbol{x}}^{k-1} - \sum_{i=1}^n \nabla_{\boldsymbol{x}} E_\theta(\tilde{\boldsymbol{x}}^{k-1} \mid \mathrm{Enc}(\boldsymbol{r}_i)) + \omega, \;\; \omega \sim \mathcal{N}(0, \sigma)$
  end for

  ▷ *Iteratively refine samples:*
  $\tilde{\boldsymbol{x}}^0 = \tilde{\boldsymbol{x}}^k$
end for

▷ *Final output:*
$\boldsymbol{x} = \tilde{\boldsymbol{x}}^0$

Algorithm 4 Image Editing Algorithm

Input: input image $\tilde{\boldsymbol{x}}^0$, relational scene description R, number of data augmentation applications N, step size λ, number of steps K, data augmentation $D(\cdot)$, EBM $E_\theta(\cdot)$, Parser $P(\cdot)$, Encoder $\mathrm{Enc}(\cdot)$
▷ *Parse a relational scene description:*
$\{\boldsymbol{r}_1, \dots \boldsymbol{r}_m\} \leftarrow P(R)$

▷ *Generate samples through N iterative steps of data augmentation/Langevin dynamics:*
for sample step $n = 1$ to N **do**
  ▷ *Apply data augmentation to samples:*
  $\tilde{\boldsymbol{x}}^0 = D(\tilde{\boldsymbol{x}}_i^0)$

  ▷ *Run K steps of Langevin dynamics:*
  for sample step $k = 1$ to K **do**
    $\tilde{\boldsymbol{x}}^k \leftarrow \tilde{\boldsymbol{x}}^{k-1} - \sum_{i=1}^n \nabla_{\boldsymbol{x}} E_\theta(\tilde{\boldsymbol{x}}^{k-1} \mid \mathrm{Enc}(\boldsymbol{r}_i)) + \omega, \;\; \omega \sim \mathcal{N}(0, \sigma)$
  end for

  ▷ *Iteratively refine samples:*
  $\tilde{\boldsymbol{x}}^0 = \tilde{\boldsymbol{x}}^k$
end for

▷ *Final output:*
$\boldsymbol{x} = \tilde{\boldsymbol{x}}^0$