Enabling ControlNet to follow Localized Descriptions using Cross-Attention Control

Denis Lukovnikov (✉) and Asja Fischer

Ruhr University Bochum, Germany {denis.lukovnikov,asja.fischer}@rub.de

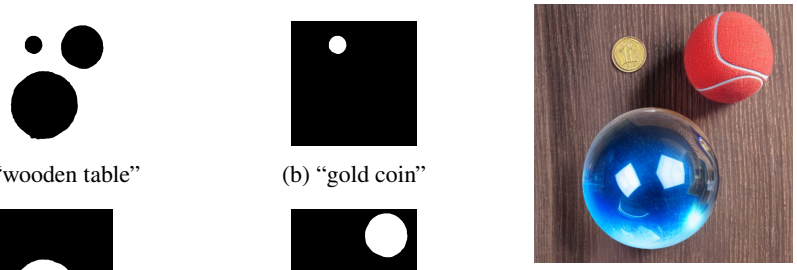
Abstract. ControlNet enables fine-grained control over image layout in prominent generators like Stable Diffusion. However, it lacks the ability to take into account localized textual descriptions that indicate which image region is described by which phrase in the prompt. In this work, we enable ControlNet to use localized descriptions using a training-free approach that modifies the cross-attention scores during generation. For doing so, we adapt and investigate several existing cross-attention control methods and identify shortcomings that cause failure or image degradation under some conditions. To address these shortcomings, we develop a novel cross-attention manipulation method. Qualitative and quantitative experimental studies demonstrate the effectiveness of the proposed augmented ControlNet.

Keywords: localized descriptions · layout-to-image · diffusion models

1 Introduction

Diffusion-based text-to-image models like Stable Diffusion [23] can generate high-quality images of various types of subjects from textual description. However, they lack fine-grained control over the composition of the generated image, which would increase their usefulness in various applications. The default training method does not address generation scenario's where additional control inputs can be used to describe the desired composition of the image (e.g., using line art or segmentation maps). Recent work has explored fine-tuning adapters (e.g. ControlNet [34], GLIGEN [15], T2I-Adapters [20]) to the diffusion model's U-Net that enable precise control over the layout of the generated images. Arguably the most popular and effective among these is ControlNet. However, it lacks the ability to use localized descriptions that specify which objects should be generated in the different parts of the image.

Therefore, in this work, we extend the pre-trained segmentation-based ControlNet for controlling image layout with cross-attention control as a means to improve the assignment of objects and reduce missing or misplaced objects and concept bleeding that frequently occur when ControlNet is used for more complicated scenes with multiple similar objects. Cross-attention control methods have the advantages that they neither require significant computational overhead nor additional training and that they can be easily plugged into existing models. Specifically, the contributions of this work are threefold. Firstly, we investigate several training-free attention-based extensions of ControlNet to improve its grounding with a given localized textual description and identify



(a) "wooden table" (b) "gold coin" (c) "blue crystal ball" (d) "red tennis ball"

(e) Desired output image for prompt "*A photo of a {blue crystal ball}$_c$, a {red tennis ball}$_d$ and a {gold coin}$_b$ on a {wooden table}$_a$*"

Fig. 1: An example of the task. The input consists of masks (a)-(d) and the annotated prompt in the caption of (e). The desired output is shown in (e). See Sec. 2.

important characteristics of such methods. To the best of our knowledge, this investigation presents the first in-depth comparison of this family of methods in the context of ControlNet. Secondly, we develop a novel cross-attention (CA) control method that facilitates better object alignment between the text prompt and the generated image, while minimizing image quality degradation. Lastly, we conduct an extensive empirical study using the newly developed SIMPLESCENES dataset as well as COCO2017 [17]. Qualitative and quantitative experiments demonstrate the effectiveness of CA control in conjunction with ControlNet, as well as improvements compared to other CA control methods and other baselines that do not rely on ControlNet.

2 Task: Layout-to-Image with Localized Descriptions

The focus of this work lies in improving the faithfulness of a generated image of height H and width W to a localized description. The input for this task consists of (1) a prompt X consisting of N tokens, (2) a collection of R region masks $\{\mathbf{B}_r\}_{r=0,\dots,R}$, where $\mathbf{B}_r \in \{0,1\}^{H \times W}$ [1], and (3) region-token alignments $f_{\text{RT}} : [1..N] \rightarrow [0..R]$ [2] that specify which region each token in the text prompt belongs to. As an example, consider the prompt "*A photo of a {blue crystal ball}$_1$, a {red tennis ball}$_2$, and a {gold coin}$_3$ on a {wooden table}$_4$*", where each colored sub-sequence is associated with the corresponding mask shown in Figs. 1a to 1d.

The goal is to generate an image where (1) object boundaries follow mask boundaries and (2) where the objects described by the region-specific parts of the prompts (i.e, the region descriptions) are generated in the parts of the output image associated with their region description. The desired output for our example is given in Fig. 1e.

[1] The region mask $\mathbf{B}_r$ contains the value 1 for pixels where the object should be present.

[2] Region 0 is the entire image, so a token assigned to region 0 is relevant everywhere in image.

3 Background

In this section, we briefly review the diffusion-based text-to-image generation process, ControlNet, and some existing cross-attention control methods designed for using localized descriptions with diffusion models (DMs).

3.1 Text-to-Image with Denoising Diffusion

In text-to-image DMs, first, a textual description X of N input tokens is encoded by the text encoder, producing the text embeddings $\mathbf{X} = \{\vec{x}_n\}_{n=0..N}$, where each $\vec{x}_n \in \mathbb{R}^{d_x}$. Here, d_x is the dimensionality of the embedding. Stable Diffusion uses CLIP [22] as encoder, which is a transformer pre-trained on a text-image similarity task.

Then, a denoising model is used to iteratively denoise an initial $z_T \sim \mathcal{N}(0, I)$ into an image z_0 using some solver, such as DDIM [27]. At every iteration, the solver computes $z_{t-\delta} = s(u(z_t, t, \mathbf{X}), t, \delta)$, where t is the denoising step, $u(\cdot)$ is the denoising model, $s(\cdot)$ is the solver algorithm, and δ is the step size. The denoising model can be implemented as a U-Net [24], which is conditioned on both the input text X, and the noisy image z_t.

Conditioning on the text can be accomplished by using a cross-attention mechanism between the token embeddings $\mathbf{X} \in \mathbb{R}^{N \times d_x}$ and pixel-wise features $\mathbf{H} \in \mathbb{R}^{H \times W \times d_h}$:

$$\mathbf{A} = \mathrm{softmax}\Big(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\Big) \quad \text{and} \quad \mathbf{C} = \mathbf{A}\mathbf{V} \ . \tag{1}$$

Here, $\mathbf{Q} = f_Q(\mathbf{H}) \in \mathbb{R}^{(H \cdot W) \times d}$ are the query vectors computed by projecting the pixel-wise feature maps $\mathbf{H}$. $\mathbf{K} = f_K(\mathbf{X}) \in \mathbb{R}^{N \times d}$ and $\mathbf{V} = f_V(\mathbf{X}) \in \mathbb{R}^{N \times d}$ are the key and value projections of the token embeddings. $\mathbf{A} \in \mathbb{R}^{(H \cdot W) \times N}$ are the cross-attention scores. Note that we omit layer and head indexes for clarity and that the dimensions H and W as well as the size of feature vectors d and d_h vary depending on the layer.

3.2 ControlNet

ControlNet [34] was recently proposed to improve control over the image composition. In addition to the prompt X, ControlNet expects an image c_{img} as part of the input for the generation process. In order to incorporate conditioning based on c_{img}, first a control model is defined that copies the down-sampling and middle blocks of the latent diffusion model's U-Net. The control model also contains an additional block of convolutional layers that encodes the control signal c_{img} and is trained from scratch. The features computed by the control model are added to the features computed by its sibling in the main U-Net before feeding them into the up-sampling blocks of the main U-Net. We refer the reader to Supplement C and to the original work [34] for a more detailed explanation. ControlNet supports different types of conditioning input, such as segmentation maps, depth maps or human pose. Each type requires the training of a separate control model dedicated to that type of conditioning. Combining several control signals [35] is an active research area.

Note that while ControlNet allows us to control the image layout using segmentation maps, it lacks a mechanism to precisely control what object is generated inside each

region. As a consequence, when faced with ambiguous layouts or improbable region assignments, plain ControlNet can not correctly process the prompt, as illustrated in our qualitative study in Sec. 5.1.

3.3 Cross-attention control

Modifying cross-attention [2] scores in the transformer [29] blocks of the U-Net can provide a degree of spatial control and attribute assignment. Here we give a brief introduction to previously proposed methods aimed at implementing the ability to follow localized descriptions via cross-attention control. In addition to the token embeddings $\vec{x}_n$, these methods expect the region masks $\mathbf{B}_r$, as well as the region-token alignments f_{RT} as inputs. In general, CA control mechanisms stimulate cross-attention from the specified region to the corresponding set of tokens and/or prevent from attending to the descriptions of other regions.

eDiff-I (community edition) The first cross-attention control method we consider is a re-implementation [26] of the approach proposed by Balaji et al. [3]. It takes the region-annotated prompt and the region masks, and forces cross-attention to attend to certain words from the corresponding regions by modifying the cross-attention scores to[3]

$$\mathbf{A} = \text{softmax}\Big(\mathbf{W} + \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\Big) \ , \quad \text{with} \tag{2}$$

$$\mathbf{W} = W' \cdot \log(1+\sigma^2) \cdot \text{std}(\mathbf{Q}\mathbf{K}^T) \cdot \mathbf{B}_{f_{\text{RT}}} \ . \tag{3}$$

Here, $\mathbf{W}$ is scheduled to decrease as the denoising process progresses, due to the dependence on σ, which is a scalar specifying the current noise level. W' is a hyper-parameter controlling the overall degree of attention change and $\mathbf{B}_{f_{\text{RT}}} \in \{0,1\}^{(H\cdot W)\times N}$ are the masks $\mathbf{B}_r$ stacked according to f_{RT}.

CAC Instead of boosting cross-attention scores between regions and their descriptions in the prompt, CAC [9] applies a binary mask that eliminates attention between regions and non-matching region descriptions. The binary mask is applied *after* softmax normalization, that is[4]

$$\mathbf{A} = \text{softmax}\Big(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\Big) \odot \max(1 - \mathbf{B}_R, \mathbf{B}_{f_{\text{RT}}}) \ , \tag{4}$$

and thus attention weights are no longer normalized after applying the mask. $\mathbf{B}_R \in \{0,1\}^{(H\cdot W)\times N}$ is a mask that is set to one for all tokens that belong to *any* region description.

[3] Note that this formulation by the community slightly differs from the one proposed by [3]. We use this formulation since in our early experiments, we found it to perform slightly better.

[4] Since the source code for [9] has not been made available at the time of this writing, we had to rely on the descriptions given in the paper for our implementation.

DenseDiffusion In this variant of cross-attention control [12], attention scores for the tokens describing a region are increased while attention scores to other tokens are decreased. In addition, the method also proposes to scale the degree of change by the region size fraction S and uses a schedule that decreases polynomially. Cross-attention scores are modified by redefining $\mathbf{W}$ from Eq. 2 as follows

$$\mathbf{W} = W' \cdot \left(\frac{t}{T}\right)^5 \cdot (1 - S) \cdot (\mathbf{B}_{f_{\mathrm{RT}}} \odot \mathbf{M}_+ - (1 - \mathbf{B}_{f_{\mathrm{RT}}}) \odot \mathbf{M}_-) \ , \tag{5}$$

where $\mathbf{M}_+$ and $\mathbf{M}_-$ specify the maximum increase and decrease for every token, i.e.

$$\mathbf{M}_+ = \max(\mathbf{Q}\mathbf{K}^T) - \mathbf{Q}\mathbf{K}^T \qquad \text{and} \qquad \mathbf{M}_- = \mathbf{Q}\mathbf{K}^T - \min(\mathbf{Q}\mathbf{K}^T) \ . \tag{6}$$

In addition to cross-attention control, DenseDiffusion [12] also includes self-attention control using a similar method.

4 Approach

In this work, we use the segmentation-based ControlNet on top of Stable Diffusion as it already provides us with a means to control image layout with high precision by specifying segmentation maps. Note that even though the image layout is controlled, it still leaves the model with freedom how to assign the objects mentioned in the text prompt to the regions. In this work we focus on investigating cross-attention control methods in conjunction with ControlNet to enable it to solve the task described in Section 2.

4.1 Cross-Attention Control in ControlNet

In a first attempt to enable ControlNet to solve the task defined in Section 2, we adapt and integrate several representative cross-attention control methods into ControlNet. More precisely, we implement the methods described in Sec. 3.3, and apply cross-attention control in both the control network as well as the main diffusion U-Net. Note that different layers of the U-Net work at different resolutions as the network consists of a stack of down-scaling layers, followed by up-scaling layers. Therefore, we down-scale the mask $\mathbf{B}_r$ as necessary.

4.2 Cross-Attention Control Design Considerations

The previously discussed cross-attention control methods have certain shortcomings. Firstly, most methods are sensitive to the selection of the time steps during which attention manipulation is performed and to what degree it is performed. Most methods studied here rely on the assumption that the image layout is determined in the initial denoising steps and do not modify attention in later generation stages. Such methods (e.g., DenseDiffusion and eDiff-I) therefore place a high degree of control in the initial stages of decoding and quickly drop it to near-zero values by roughly $t = 750$ (if generation starts with $t = T = 1000$). However, this procedure can still lead to concept bleeding in highly ambiguous cases, for example when generating objects of similar shapes and

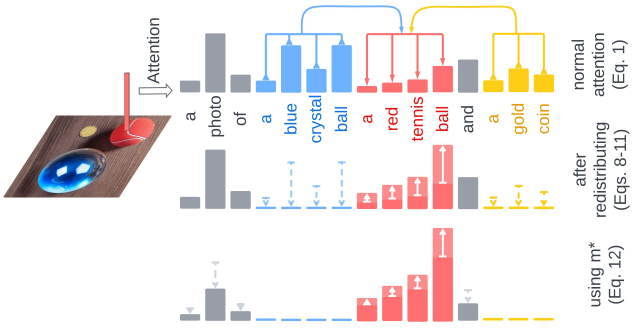


Fig. 2: A diagram illustrating attention redistribution and attention boosting.

color. After the initial heavily controlled stage, the model becomes uncontrolled and fine details such as object texture can no longer be clearly assigned when multiple similar objects are present. Thus, it is desirable to have an attention control method that remains active throughout the denoising process while still minimizing image quality degradation.

A second consideration is that at different heads of different layers and at different generation stages, the attention weights behave differently and indiscriminate boosting of attention can lead to a decrease in image quality and a higher sensitivity to the attention control schedule. The exception is CAC, since it only disables attention to the descriptions of irrelevant regions throughout the entire generation process.

Thirdly, when simply disabling attention to irrelevant tokens, like CAC, the attention "mass" is either mostly transfered to the most probable tokens or is lost. This can be problematic when the initial random image x_T leads the model to mostly attend to the wrong region descriptions, in this case, attention to the wrong regions is dropped but the attention weights to the correct region remain at their initial (possibly low) values. In addition, in CAC-style control, the attention weights no longer sum to one.

4.3 Attention Redistribution

To address these shortcomings, we propose a cross-attention manipulation method that we refer to as **cross-attention redistribution (CA-Redist)** and that redistributes attention from irrelevant region descriptions to the relevant one. Concretely, this is accomplished by (1) computing the total amount of region-specific attention m, which can vary across heads and layers, (2) separately normalizing region-specific and region-agnostic attention weights to obtain $\mathbf{A}_{\text{local}}$ and $\mathbf{A}_{\text{global}}$, respectively, and (3) mixing the two resulting attention distributions using $m \in [0, 1]^{H \cdot W}$. Note that this is done separately for every pixel, and that it is assumed that every pixel in the image is assigned to exactly one region description. Fig. 2 provides an illustration of CA-Redist. This method is defined as follows:

$$\mathbf{A} = m \odot \mathbf{A}_{\text{local}} + (1 - m) \odot \mathbf{A}_{\text{global}} \quad , \quad \text{with} \tag{7}$$

$$\mathbf{A}_{\text{local}} = \text{softmax}\Big(\log(\mathbf{B}_{f_{\text{RT}}}) + \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}} \Big) \ , \tag{8}$$

$$\mathbf{A}_{\text{global}} = \text{softmax}\Big(\log(1 - \mathbf{B}_R) + \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}} \Big) \quad \text{and} \tag{9}$$

$$m = \sum_{n=0}^{N} \mathbf{A}[\cdot, n] \cdot \mathbf{B}_R[\cdot, n] \ , \tag{10}$$

where $\mathbf{A}$ is defined as in Eq. 1, $\mathbf{A}[\cdot, n]$ is its n-th column, and $\mathbf{B}_R$ and $\mathbf{B}_{f_{\text{RT}}}$ are as defined earlier in Section 3.3. Thus, $\mathbf{A}_{\text{global}}$ computes an attention distribution over all tokens except those in any region description and $\mathbf{A}_{\text{local}}$ is zero everywhere except the correct region description. The mixture between the two makes sure to retain the same attention weights for the non-region tokens (in other words, keeping $\mathbf{A}_{\text{global}} \approx \mathbf{A}$ for tokens where $\mathbf{B}_R$ is zero).

The attention to relevant region-specific parts of the prompt can further be increased by replacing m with m^* as defined below, where m can be modified in two ways, using hyper-parameters $W_m \geq 0$ and $W_a \geq 0$ that boost the attention to relevant parts of the prompt multiplicatively or additively, respectively.

$$m^* = \min\Big(1, \max\big(0, m \cdot (1 + W_m \cdot W'') + W_a \cdot W'' \cdot (1 - S)\big)\Big) \ , \tag{11}$$

where S is the fraction of the surface area that a region occupies in the image (same as defined for DenseDiffusion) and W'' specifies the schedule of attention boost in CA-Redist and depends on the current denoising step t:

$$W'' = \begin{cases} 1 & \text{if } t \geq T_{\text{s}} \\ \frac{1}{2} + \frac{1}{2}\sin(\pi \cdot \frac{t - T_{\text{thr}}}{T_{\text{s}} - T_{\text{e}}}) & \text{if } T_{\text{s}} > t > T_{\text{e}} \\ 0 & \text{if } T_{\text{e}} \geq t \end{cases} \ , \tag{12}$$

$$\text{with } \ T_{\text{s}} = T_{\text{thr}} + RT/2 \quad \text{and} \quad T_{\text{e}} = T_{\text{thr}} - RT/2 \ . \tag{13}$$

This schedule is controlled by the threshold step $T_{\text{thr}} \in [1..T]$ and the threshold softness $R \in [0, 1]$. Unless otherwise specified, in our experiments, we set $T_{\text{thr}} = T$. This simplifies the schedule to the following:

$$W'' = \begin{cases} \frac{1}{2} + \frac{1}{2}\sin(\pi \cdot \frac{t - T}{R \cdot T}) & \text{if } t > T \cdot (1 - R/2) \\ 0 & \text{otherwise} \end{cases} \ . \tag{14}$$

In Eq. 12, t starts from T so attention boost is active more in the initial stages of denoising with a value between zero and one and gradually decays to zero as denoising progresses. This schedule for $R = 0.4$ is illustrated in Fig. 3.

Multiplicative manipulation using W_m is stronger for heads with higher attention weights to region descriptions and remains low for those that attended to tokens outside of any region description. Additive manipulation using W_a forces attention to increase to

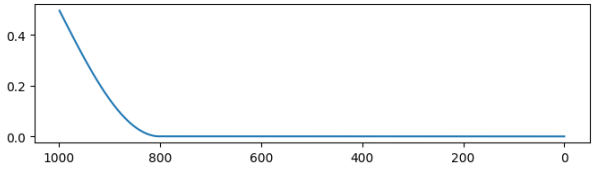


Fig. 3: CA-Redist schedule W'' if $T_{\text{thr}} = T$ and $R = 0.4$.

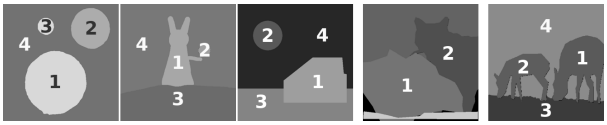


Fig. 4: Layouts used for qualitative comparison throughout this paper (Fig. 5 and Fig. 6).

region-specific tokens in all heads. Note that the computational overhead this attention manipulation introduces is similar to other attention control methods, and is negligible compared to the computational requirements of running the diffusion model.

5 Experiments

We compare the following methods of cross-attention control on top of the (lightly fine-tuned) ControlNet: (1) eDiff-I, (2) CAC, (3) DenseDiffusion (DD), and (4) CA-Redist. We also compare against the original implementations of GLIGEN [5] and DenseDiffusion [6] as points of reference of related work that does not rely on ControlNet. Note that we used the main variant of GLIGEN that takes as input bounding boxes and localized descriptions.[7] Comparison with SceneComposer [33] and SpaText [1] was not possible because the code and data have not been publicly released at the time of this writing. It must also be noted that both these approaches have been extensively trained, whereas our proposed method also works training-free with segmentation- and scribble-based ControlNet. In our experiments, we did minimal fine-tuning of a small part of ControlNet on a readily available dataset to align color schemes since we found it slightly improved image quality in our early experiments. To adapt to the task, we fine-tuned the segmentation ControlNet on panoptic segmentation data from COCO2017 [17] using randomized colors. More details on the experimental setup can be found in Supplement A.2. Our code is available at https://github.com/lukovnikov/ca-redist.

5.1 Qualitative study

A qualitative comparison of the different attention control methods is presented in Fig. 5. The layouts used are specified by the first three images in Fig. 4, where the numbers cor-

[5] As provided in the Huggingface Diffusers library

[6] https://github.com/naver-ai/densediffusion

[7] GLIGEN also provides a variant that takes a segmentation map as conditioning but it does not use localized textual descriptions so it is unfit for comparison.

Fig. 5: A qualitative comparison of different cross-attention control methods in ControlNet-extended Stable Diffusion 1.5. See Fig. 4 for layout specification.

respond to the numbered phrases in the prompts in Fig. 5. For a qualitative comparison using ControlNet trained for sketch conditioning (scribbles), please see Supplement D.

Comparison of baselines: The baselines that don't rely on ControlNet appear to fail at the task with challenging inputs. Note that **GLIGEN** [15] only allows to use bounding boxes for conditioning in image generation with localized descriptions, so the exact layout is not expected to match. But despite enabling its adapter throughout the entire generation process ($\tau = 1$), GLIGEN mostly failed to assign the right textures and colors to objects for the challenging prompts involving multiple round objects. This can be speculated to be attributable to the fact that during training, GLIGEN is insufficiently exposed to such challenging examples and doesn't learn to take into account localized descriptions in later generation steps.

In comparison, the original implementation of **DenseDiffusion** [12] is better at assigning objects to regions. However, we see it fail in examples where the shape and colors in the early generation stages are ambiguous, as illustrated in the last two columns of Fig. 5. Also, it seems to ignore smaller object masks and does not adhere to mask boundaries as precisely as ControlNet (however, it must be noted that DenseDiffusion is completely training-free).

Finally, **plain ControlNet**, pre-trained with semantic segmentation labels, and fine-tuned on COCO2017 data for panoptic segmentation with randomized colors (referred to as ControlNet* in the figures and tables), can already assign the correct description to the correct region if the mask shapes are distinctive enough. This is illustrated in the fifth column, where all shapes can be unambiguously matched with a region description (e.g. "fire ball" is a circular shape, "doll house" is a trapezoid shape). However, when faced with ambiguity in layout specification (e.g. three circles), plain ControlNet randomly assigns objects and colors and suffers from concept bleeding (for example, assigning "*gold*" to a ball whereas we described a gold coin). Additionally, it can struggle when faced with improbable descriptions, such as a rabbit mage standing on clouds.

Comparison of attention control methods: For **CAC**-style control, we observe that it does help resolve ambiguity and improve grounding behavior, but not very consistently across different seeds. In the first column, for example, the assignment completely failed. We also see that it does not resolve the issue of improbable assignments, leaving the images largely the same as plain ControlNet for the rabbit example. We also observed that small objects are sometimes not generated.

The other methods appear to provide satisfactory degree of control over object assignment in most cases. However, as we can see in the last two columns of Fig. 5, for the prompt "*an apricot, a pumpkin, and an orange*", **DenseDiffusion** and **eDiff-I** suffer from the aforementioned control scheduling problems, where objects of similar color and shape are not assigned correctly. Both generate two or three pumpkins, ignoring other described objects. We observed similar behavior with other test cases.

In contrast, CA-Redist adheres to localized descriptions better than DenseDiffusion and eDiff-I in more challenging control scenarios (objects of similar shape and color) while maintaining image quality.

Ablation: The bottom three rows of Fig. 5 show an ablation of CA-Redist, which shows it is still effective when only W_m is non-zero (CA-Redist (m)) or only W_a is non-zero (CA-Redist (a)). When both are zero (CA-Redist (none)), the images don't always satisfy all region descriptions. This shows that some form of attention boosting is still necessary.

Qualitative analysis with objects from COCO-2017: A comparison of different methods using object shapes from COCO-2017 examples is shown in Fig. 6. Note that baselines (+CAC, +DD, +eDiff-I) fail to always assign the correct descriptions to the right locations and properly separate features. For example, for cats and dogs, +DD generates two cats and eDiff-I, while largely performing quite well still assigns a mixture of cat- and dog-like features to the region annotated as a "grey dog".

5.2 Quantitative study

Automatic evaluation for this task is challenging because of the inherent difficulty of evaluating image quality and faithfulness to a localized description. Moreover, there is a lack of standardized open-sourced datasets and evaluation methodology. Nevertheless, to be able to perform a quantitative analysis, we constructed a challenging dataset of simple scenes with multiple objects (SIMPLESCENES) that allows to estimate image faithfulness of localized descriptions as well as image quality, and additionally investigate image quality on COCO2017.

SIMPLESCENES Our goals are two-fold: measuring (1) image quality to detect image degradation and (2) faithfulness to localized descriptions.[8] Since COCO images frequently contain objects with overlapping bounding boxes, these could pollute metrics for measuring conformity to localized descriptions.

[8] ControlNet follows segmentation map conditioning very well so we do not evaluate this.

Table 1: Image quality and localized prompt faithfullness using our SIMPLESCENES dataset. Arrows indicate if higher (↑) or lower (↓) is better. The format is MEAN$^{\pm\text{STD}}$(BEST), over five seeds.

	BRISQUE ↓	LAION Aest ↑	LocalCLIP Logits ↑	LocalCLIP Prob. ↑
GLIGEN ($\tau = 1$)	30.56 $^{\pm 1.78}$ (10.23)	5.51 $^{\pm 0.04}$ (6.08)	21.60 $^{\pm 0.17}$ (23.19)	0.33 $^{\pm 0.01}$ (0.50)
DD (original)	32.69 $^{\pm 1.14}$ (12.96)	5.76 $^{\pm 0.03}$ (6.33)	21.52 $^{\pm 0.06}$ (22.83)	0.45 $^{\pm 0.01}$ (0.58)
ControlNet*	25.23 $^{\pm 1.46}$ (7.65)	5.70 $^{\pm 0.02}$ (6.25)	20.99 $^{\pm 0.12}$ (22.60)	0.25 $^{\pm 0.01}$ (0.41)
+CAC	26.97$^{\pm 1.02}$ (8.99)	5.71$^{\pm 0.05}$ (6.29)	22.28$^{\pm 0.23}$ (23.84)	0.44$^{\pm 0.02}$ (0.61)
+DD (w=0.5)	24.50$^{\pm 1.94}$ (8.39)	5.74$^{\pm 0.05}$ (6.21)	22.93$^{\pm 0.11}$ (24.18)	0.48$^{\pm 0.02}$ (0.59)
+eDiff-I (w=0.5)	23.75$^{\pm 1.15}$ (10.86)	5.73$^{\pm 0.02}$ (6.22)	23.36$^{\pm 0.13}$ (24.43)	0.58$^{\pm 0.02}$ (0.68)
+CA-Redist (m+a)	25.40$^{\pm 2.29}$ (10.10)	5.74$^{\pm 0.02}$ (6.22)	23.77$^{\pm 0.11}$ (24.89)	0.62$^{\pm 0.01}$ (0.73)
+CA-Redist (m)	27.37$^{\pm 2.00}$ (10.54)	5.68$^{\pm 0.05}$ (6.18)	23.70$^{\pm 0.13}$ (24.80)	0.62$^{\pm 0.01}$ (0.72)
+CA-Redist (a)	24.86$^{\pm 1.33}$ (8.41)	5.69$^{\pm 0.01}$ (6.20)	23.52$^{\pm 0.07}$ (24.74)	0.58$^{\pm 0.01}$ (0.70)
+CA-Redist (none)	25.13$^{\pm 1.54}$ (8.14)	5.65$^{\pm 0.04}$ (6.20)	23.26$^{\pm 0.04}$ (24.60)	0.56$^{\pm 0.01}$ (0.70)

Dataset: For this reason, and to focus on more challenging cases, we create the SIMPLESCENES dataset. The dataset consists of 124 examples, each containing 3-4 objects and randomized descriptions, which proved to be challenging for the tested methods. The dataset is described in more detail in Supplement A.1.

Metrics: We use the following evaluation methodology for this dataset. For measuring general image quality, we use reference-free (since we don't have reference images) im-

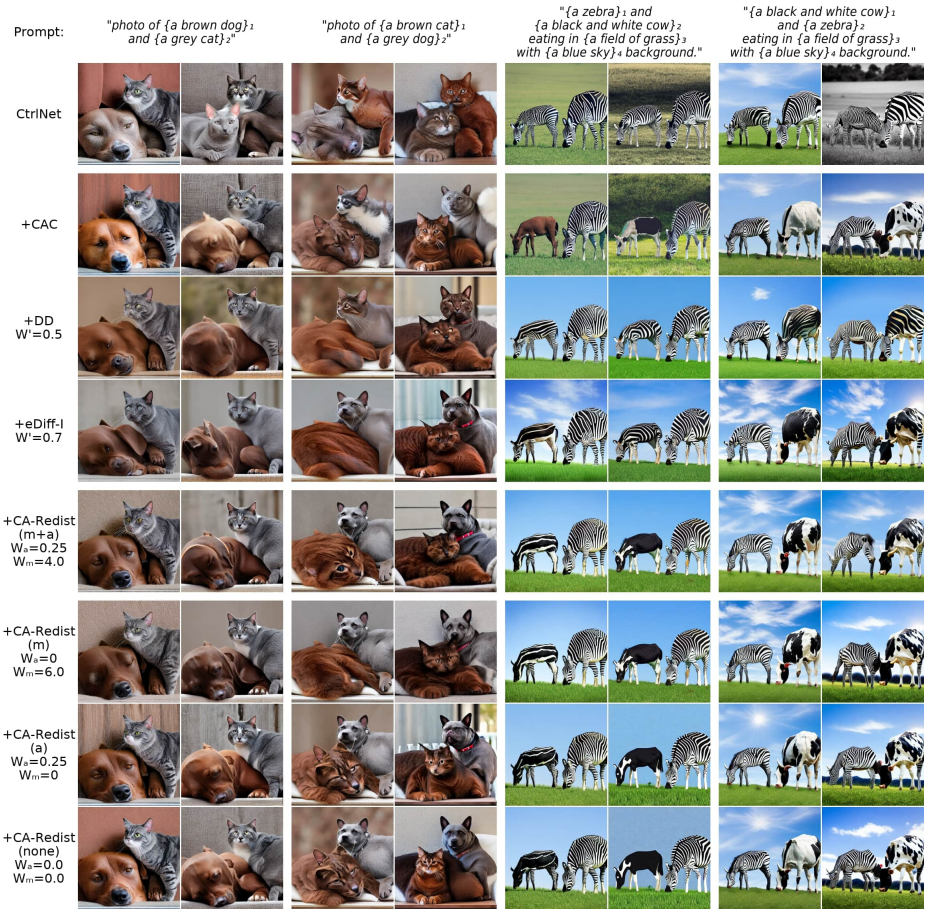


Fig. 6: A qualitative comparison with complex shapes. In the example for cats and dogs, region 1 corresponds to the bottom left region and region 2 to the region on the right. In the example for cows and zebras, region 1 corresponds to the animal shape on the right and region 2 to the animal shape on the left. For exact layout specifications, see Fig. 4. We can observe some concept bleeding for different methods, especially evident in the cows and zebras examples, even for CA-Redist. Concept bleeding is more severe for the DenseDiffusion and eDiff-I baselines while the CAC baseline generated a brown animal instead of a black and white cow.

age quality assessment methods, such as BRISQUE [19] and MANIQA [32], as well as the LAION Aesthetics Score predictor [25]. For measuring conformity to localized description, we use the localized CLIP (LocalCLIP) Logits and Probabilities, which are computed as follows: For every object mask, we crop the image to contain only the masked region of the generated image, and use CLIP to compute text-image similarities between all the localized phrases (e.g. "*blue crystal ball*") and all the cropped object images. The reported CLIP Logits correlate linearly with the similarities while the reported CLIP Probabilities result from normalizing the logits over all objects in the image.
Results: The numbers reported in Table 1 demonstrate that CA-Redist does not suffer from image quality loss, with all image quality metrics being on par with plain ControlNet. BRISQUE values are significantly lower than that for GLIGEN and DenseDiffusion. Regarding MANIQA numbers, all tested methods achieved MANIQA scores of 0.665 ± 0.015, thus not showing any measurable image quality differences among the tested methods. We did not report these numbers in Table 1 because of space constraints. On the other hand, the LocalCLIP metrics indicate that CA-Redist is superior when it comes to conformity to localized descriptions: while plain ControlNet achieves lower values than GLIGEN and DenseDiffusion, ControlNet with CA-Redist leads to a significant improvement in LocalCLIP scores.
Ablation: From the three ablation settings (CA-Redist variants (a), (m) and (none)), it appears that moderately boosting attention does not result in measurable image quality decrease. However, no attention boost (CA-Redist (none)) results in slightly lower faithfulness to the localized prompt (as indicated by the localized CLIP metrics), which confirms our qualitative observations.

COCO2017 To further investigate image quality, we report FID [10] and KID [5] scores between 5000 samples generated using the segmentation maps from the COCO2017 validation set and the corresponding 5000 real images in Table 2. Even though ControlNet has worse FID and KID than GLIGEN, the addition of CA control does not result in quality loss measurable by these metrics. Note that GLIGEN and DenseDiffusion don't always adhere to the masks as closely as ControlNet since GLIGEN only uses bounding boxes instead of segmentation maps as input, and DenseDiffusion uses self-attention control. Thus, both don't follow the input segmentation maps as closely as ControlNet.

Table 2: FID and KID w.r.t. COCO2017 validation images.

	FID ↓	KID ($\times 10^3$) ↓
GLIGEN (τ = 1.0)	23.84	4.289
DD (original)	37.68	7.923
ControlNet*	28.84	5.150
+CAC	27.42	4.916
+DD (w=0.5)	28.30	5.422
+eDiff-I (w=0.5)	28.72	6.074
+CA-Redist (m+a)	27.11	5.276
+CA-Redist (m)	27.78	5.547
+CA-Redist (a)	26.15	4.618

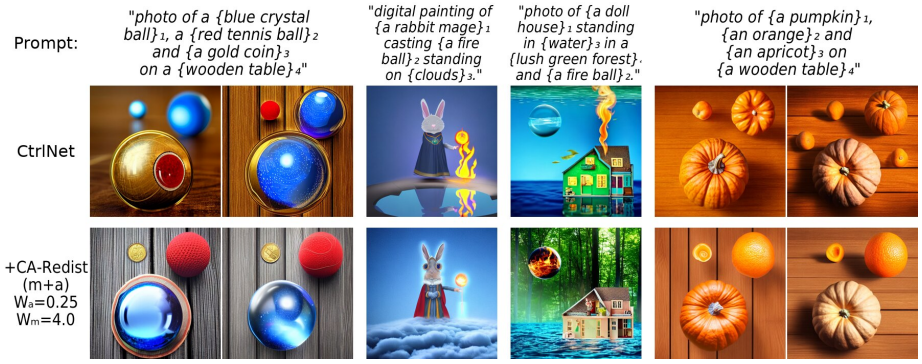


Fig. 7: A qualitative comparison between sketch-based ControlNet with and without CA-Redist. See Supplement D for a full comparison.

Fig. 8: Image quality with increasing control strength for eDiff-I and DenseDiffusion cross-attention control with ControlNet.

5.3 Image quality vs control strength

In Fig. 8 are shown some images generated using eDiff-I-based and DenseDiffusion-based attention control. We can observe that at lower attention control strengths (controlled by the inference hyper-parameter W'), eDiff-I does not correctly follow localized descriptions. Fidelity increases with increasing W', however, it comes at a cost to image quality, which is particularly noticeably for higher values. For DenseDiffusion, increasing W' to its maximum value of 1.0 does not improve fidelity to localized descriptions but also introduces more subtle image quality changes.

5.4 CA-Redist for sketch-based control

We performed additional analysis into using sketch (scribble) conditioning instead of segmentation maps, as well as using more complex shapes. In Fig. 7, we show that CA-Redist also performs well with ControlNet trained for sketch conditioning (scribbles). A full qualitative comparison is provided in Supplement D.

6 Related Work

Since the early works on controllable image synthesis [11, 7], the emergence of neural-network based generative models opened new frontiers for this task, especially with methods like ControlNet [34] and GLIGEN [15]. Several works have addressed tasks similar to ours with GANs [28, 39, 16]. However, these works either do not use descriptions or are limited to a restricted set of object classes.

More recently, several works have proposed methods [36, 33, 1, 4, 12, 9, 3, 6, 21, 31, 8, 18] for image synthesis with localized descriptions and regions specified by masks or bounding boxes. Several works [33, 1, 38, 30, 14] modify and train the diffusion model or adapters to condition on localized prompts (and localized reference images). Several training-free methods for this task have also been proposed [4, 12, 9, 3, 6, 21, 31, 8, 18], many of which rely on manipulating cross-attention (and self-attention) in some way.

7 Conclusion

The analysis performed in this work indicates that ControlNet is able to interpret region descriptions when mask shapes are un-ambiguous. However, when faced with similarly shaped masks, it no longer has sufficient information to correctly interpret the prompt. In such cases, additional input from the user can be used to specify which objects should be generated where. However, ControlNet is not able to process such inputs. As we demonstrate, this can be solved by integrating cross-attention control. We found, however, that some design choices are crucial in more ambiguous conditions, such as when generating multiple different objects of similar shape and color. To cover these cases better, it is important to prevent cross-attention from attending to tokens from irrelevant region descriptions throughout the *entire* generation process. Taking these considerations into account, we develop a novel cross-attention control method that shows superior generation results in our qualitative and quantitative analysis. For the latter we created a small but challenging data set, which can serve as a testbed for future work.

An interesting avenue for future work would be looking into hierarchical regions with localized descriptions that specify the objects as well as their parts (e.g. specifying where a tiger should be drawn and where its eye). It is not clear how well the current attention control methods support overlapping regions. In addition, the segmentation-based ControlNet we used for most of our study does not support overlapping regions and tries to strictly follow the region outlines.

Acknowledgments. We would like to thank all the anonymous reviewers for their feedback, which helped improve this work. This work was funded by the Ministry of Culture and Science of Northrhine-Westphalia as part of the Lamarr Fellow Network.

References

1. Avrahami, O., Hayes, T., Gafni, O., Gupta, S., Taigman, Y., Parikh, D., Lischinski, D., Fried, O., Yin, X.: Spatext: Spatio-textual representation for controllable image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18370–18380 (2023)

2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. CoRR **abs/1409.0473** (2014), https://api.semanticscholar.org/CorpusID:11212020
3. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
4. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation (2023)
5. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
6. Chen, M., Laina, I., Vedaldi, A.: Training-free layout control with cross-attention guidance. arXiv preprint arXiv:2304.03373 (2023)
7. Chen, T., Cheng, M.M., Tan, P., Shamir, A., Hu, S.M.: Sketch2photo: Internet image montage. ACM transactions on graphics (TOG) **28**(5), 1–10 (2009)
8. Couairon, G., Careil, M., Cord, M., Lathuilière, S., Verbeek, J.: Zero-shot spatial layout conditioning for text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2174–2183 (2023)
9. He, Y., Salakhutdinov, R., Kolter, J.Z.: Localized text-to-image generation for free via cross attention control. arXiv preprint arXiv:2306.14636 (2023)
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
11. Johnson, M., Brostow, G.J., Shotton, J., Arandjelovic, O., Kwatra, V., Cipolla, R.: Semantic photo synthesis. In: Computer graphics forum. vol. 25, pp. 407–413. Wiley Online Library (2006)
12. Kim, Y., Lee, J., Kim, J.H., Ha, J.W., Zhu, J.Y.: Dense text-to-image generation with attention modulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7701–7711 (2023)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Li, L., Qiu, W., Yan, X., He, J., Zhou, K., Cai, Y., Lian, Q., Liu, B., Chen, Y.C.: Omnibooth: Learning latent control for image synthesis with multi-modal instruction. arXiv preprint arXiv:2410.04932 (2024)
15. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22511–22521 (2023)
16. Li, Z., Wu, J., Koh, I., Tang, Y., Sun, L.: Image synthesis from layout with locality-aware mask adaption. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13819–13828 (2021)
17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
18. Mao, J., Wang, X.: Training-free location-aware text-to-image synthesis. arXiv preprint arXiv:2304.13427 (2023)
19. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on image processing **21**(12), 4695–4708 (2012)
20. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
21. Phung, Q., Ge, S., Huang, J.B.: Grounded text-to-image synthesis with attention refocusing. arXiv preprint arXiv:2306.05427 (2023)

22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICLR) (2021)
23. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR) (2022)
24. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
25. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5B: An open large-scale dataset for training next generation image-text models. Conference on Neural Information Processing Systems (NeurIPS) (2022)
26. Simo, R.: Paint-with-words, implemented with stable diffusion. https://github.com/cloneofsimo/paint-with-words-sd (2023)
27. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (ICLR) (2022)
28. Sushko, V., Schönfeld, E., Zhang, D., Gall, J., Schiele, B., Khoreva, A.: You only need adversarial supervision for semantic image synthesis. arXiv preprint arXiv:2012.04781 (2020)
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
30. Wang, X., Darrell, T., Rambhatla, S.S., Girdhar, R., Misra, I.: Instancediffusion: Instance-level control for image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6232–6242 (2024)
31. Xiao, J., Li, L., Lv, H., Wang, S., Huang, Q.: R&b: Region and boundary aware zero-shot grounded text-to-image generation. arXiv preprint arXiv:2310.08872 (2023)
32. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1191–1200 (2022)
33. Zeng, Y., Lin, Z., Zhang, J., Liu, Q., Collomosse, J., Kuen, J., Patel, V.M.: Scenecomposer: Any-level semantic image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22468–22478 (2023)
34. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
35. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Uni-controlnet: All-in-one control to text-to-image diffusion models. arXiv preprint arXiv:2305.16322 (2023)
36. Zheng, G., Zhou, X., Li, X., Qi, Z., Shan, Y., Li, X.: Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22490–22499 (2023)
37. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision **127**(3), 302–321 (2019)
38. Zhou, D., Li, Y., Ma, F., Yang, Z., Yang, Y.: Migc++: Advanced multi-instance generation controller for image synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)

39. Zhu, P., Abdal, R., Qin, Y., Wonka, P.: Sean: Image synthesis with semantic region-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5104–5113 (2020)