

CLIPascene: Scene Sketching with Different Types and Levels of Abstraction

Yael Vinker
Tel Aviv University

Yuval Alaluf
Tel Aviv University

Daniel Cohen-Or
Tel Aviv University

Ariel Shamir
Reichman University

<https://clipascene.github.io/CLIPascene/>

Various Scenes, Types, and Levels of Abstraction

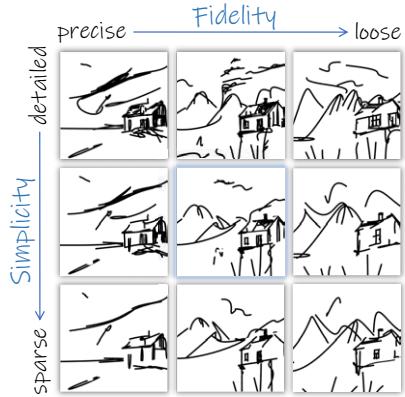


Figure 1. Our method converts a *scene* image into a sketch with different types and levels of abstraction by disentangling abstraction into two axes of control: *fidelity* and *simplicity*. The sketches on the left were selected from a complete *matrix* generated by our method (an example is shown on the right), encompassing a broad range of possible sketch abstractions for a given image. Our sketches are generated in vector form, which can be easily used by designers for further editing.

Abstract

In this paper, we present a method for converting a given scene image into a sketch using different types and multiple levels of abstraction. We distinguish between two types of abstraction. The first considers the fidelity of the sketch, varying its representation from a more precise portrayal of the input to a looser depiction. The second is defined by the visual simplicity of the sketch, moving from a detailed depiction to a sparse sketch. Using an explicit disentanglement into two abstraction axes — and multiple levels for each one — provides users additional control over selecting the desired sketch based on their personal goals and preferences. To form a sketch at a given level of fidelity and simplification, we train two MLP networks. The first network learns the desired placement of strokes, while the second network learns to gradually remove strokes from the sketch without harming its recognizability and semantics. Our approach is able to generate sketches of complex scenes including those with complex backgrounds (e.g. natural and urban settings) and subjects (e.g. animals and people) while depicting gradual abstractions of the input scene in terms of fidelity and simplicity.

1. Introduction

Several studies have demonstrated that abstract, minimal representations are not only visually pleasing but also helpful in conveying an idea more effectively by emphasizing the essence of the subject [4, 16]. In this paper, we concentrate on converting photographs of natural scenes to sketches as a prominent minimal representation.

Converting a photograph to a sketch involves abstraction, which requires the ability to understand, analyze, and interpret the complexity of the visual scene. A scene consists of multiple objects of varying complexity, as well as relationships between the foreground and background (see Figure 2). Therefore, when sketching a scene, the artist has many options regarding how to express the various components and the relations between them (see Figure 3).

In a similar manner, computational sketching methods must deal with scene complexity and consider a variety of abstraction levels. Our work focuses on the challenging task of *scene* sketching while doing so using different *types* and multiple *levels* of abstraction. Only a few previous works attempted to produce sketches with multiple levels of abstraction. However, these works focus specifically on the task of

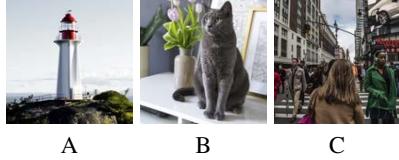


Figure 2. Scene complexity. (A) contains a single, central object with a simple background, (B) contains multiple objects (the cat and vase) with a slightly more complicated background, and (C) contains both foreground and background that include many details. Our work tackles all types of scenes.

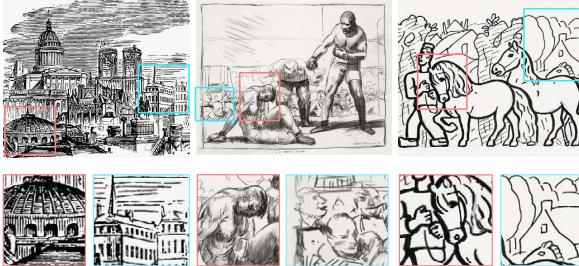


Figure 3. Drawings of different scenes by different artists. Notice the significant differences in style and level of abstraction between the drawings — moving from more detailed and precise (left) to more abstract (right). The second row shows how the level of abstraction not only varies *between* drawings, but also *within the same* drawing. Where each drawing contains areas that are relatively more detailed (red) and more abstract (blue).

object sketching [29, 43] or *portrait* sketching [2], and often simply use the number of strokes to define the level of abstraction. We are not aware of any previous work that attempts to separate different *types* of abstractions. Moreover, existing works for *scene* sketching often focus on producing sketches based on certain styles, without taking into account the abstraction level, which is an essential concept in sketching. Lastly, most existing methods for scene sketching do not produce sketches in vector format. Providing vector-based sketches is a natural choice for sketches as it allows further editing by designers (such as in Fig. 6).

We define two axes representing two types of abstractions and produce sketches by gradually moving along these axes. The first axis governs the *fidelity* of the sketch. This axis moves from more precise sketches, where the sketch composition follows the geometry and structure of the photograph to more loose sketches, where the composition relies more on the semantics of the scene. An example is shown in Figure 4, where the leftmost sketch follows the contours of the mountains on the horizon, and as we move right, the mountains and the flowers in the front gradually deviate from the edges present in the input, but still convey the correct semantics of the scene. The second axis governs the level of details of the sketch and moves from detailed to sparse depictions, which appear more abstract. Hence, we refer to this axis as the *simplicity* axis. An example can be



Figure 4. The *fidelity* axis. From left to right, using the same number of strokes the sketches gradually depart from the geometry of the input image, but still convey the semantics of the scene.

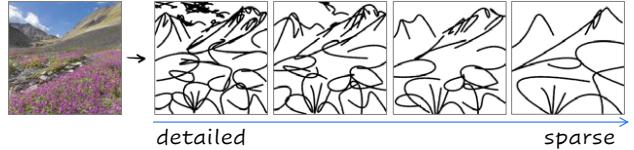


Figure 5. The *simplicity* axis. On the left, we start with a more detailed sketch and as we move to the right the sketch is gradually simplified while still remaining consistent with the overall appearance of the initial sketch.

seen in Figure 5, where the same general characteristics of the scene (*e.g.* the mountains and flowers) are captured in all sketches, but with gradually fewer details.

To deal with scene complexity, we separate the foreground and background elements and sketch each of them separately. This explicit separation and the disentanglement into two abstraction axes provide a more flexible framework for computational sketching, where users can choose the desired sketch from a range of possibilities, according to their goals and personal taste.

We define a sketch as a set of Bézier curves, and train a simple multi-layer perceptron (MLP) network to learn the stroke parameters. Training is performed per image (*e.g.* without an external dataset) and is guided by a pre-trained CLIP-ViT model [9, 34], leveraging its powerful ability to capture the semantics and global context of the entire scene.

To realize the *fidelity* axis, we utilize different intermediate layers of CLIP-ViT to guide the training process, where shallow layers preserve the geometry of the image and deeper layers encourage the creation of looser sketches that emphasize the scene’s semantics.

To realize the *simplicity* axis, we jointly train an additional MLP network that learns how to best discard strokes gradually and smoothly, without harming the recognizability of the sketch. As shall be discussed, the use of the networks over a direct optimization-based approach allows us to define the level of details *implicitly* in a learnable fashion, as opposed to explicitly determining the number of strokes.

The resulting sketches demonstrate our ability to cope with various scenes and to capture their core characteristics while providing gradual abstraction along both the fidelity and simplicity axes, as shown in Figure 1. We compare our results with existing methods for scene sketching. We addi-



Figure 6. Artistic stylization of the strokes using Adobe Illustrator.

tionally evaluate our results quantitatively and demonstrate that the generated sketches, although abstract, successfully preserve the geometry and semantics of the input scene.

2. Related Work

Free-hand sketch generation differs from edge-map extraction [5, 46] in that it attempts to produce sketches that are representative of the style of human drawings. Yet, there are significant differences in drawing styles among individuals depending on their goals, skill levels, and more (see Figure 3). As such, computational sketching methods must consider a wide range of sketch representations.

This ranges from methods aiming to produce sketches that are grounded in the edge map of the input image [20, 23, 42, 47], to those that aim to produce sketches that are more abstract [3, 11, 13, 14, 27, 31, 33, 43, 52]. Several works have attempted to develop a unified algorithm that can output sketches with a variety of styles [6, 25, 50]. There are, however, only a few works that attempt to provide various levels of abstraction [2, 29, 43]. In the following, we focus on scene-sketching approaches, and we refer the reader to [48] for a comprehensive survey on computational sketching techniques.

Photo-Sketch Synthesis Various works formulate this task as an image-to-image translation task using paired data of corresponding images and sketches [20, 22, 26, 49]. Others approach the translation task via unpaired data, often relying on a cycle consistency constraint [6, 40, 50]. Li *et al.* [20] introduce a GAN-based contour generation algorithm and utilize multiple ground truth sketches to guide the training process. Yi *et al.* [50] generate portrait drawings with unpaired data by employing a cycle-consistency objective and a discriminator trained to learn a specific style.

Recently, Chan *et al.* [6] propose an unpaired GAN-based approach. They train a generator to map a given image into a sketch with multiple styles defined explicitly from four existing sketch datasets with a dedicated model trained for each desired style. They utilize a CLIP-based loss to achieve semantically-aware sketches. As these works rely on curated datasets, they require training a new model for each desired style while supporting a single level of sketch

abstraction. In contrast, our approach does not rely on any explicit dataset and is not limited to a pre-defined set of styles. Instead, we leverage the powerful semantics captured by a pre-trained CLIP model [34]. Additionally, our work is the only one among the alternative scene sketching approaches that provides sketches with multiple levels of abstraction and in vector form, which allows for a wider range of editing and manipulation.

Sketch Abstraction While abstractions are fundamental to sketches, only a few works have attempted to create sketches at multiple levels of abstraction, while no previous works have done so over an entire scene. Berger *et al.* [2] collected portrait sketches at different levels of abstraction from seven artists to learn a mapping from a face photograph to a portrait sketch. Their method is limited to faces only and requires a new dataset for each desired level of abstraction. Muhammad *et al.* [29] train a reinforcement learning agent to remove strokes from a given sketch without harming the sketch’s recognizability. The recognition signal is given by a sketch classifier trained on nine classes from the QuickDraw dataset [14]. Their method is therefore limited only to objects from the classes seen during training and requires extensive training.

CLIPasso Most similar to our work is CLIPasso [43] which was designed for *object* sketching at multiple levels of abstraction. They define a sketch as a set of Bézier curves and optimize the stroke parameters with respect to a CLIP-based [34] similarity loss between the input image and generated sketch. Multiple levels of abstraction are realized by reducing the number of strokes used to compose the sketch. In contrast to CLIPasso, our method is not restricted to objects and can handle the challenging task of *scene* sketching. Additionally, while Vinker *et al.* examine only a single form of abstraction, we disentangle abstraction into two distinct axes controlling both the simplicity and the fidelity of the sketch. Moreover, in CLIPasso, the user is required to *explicitly* define the number of strokes needed to obtain the desired abstraction. However, different images require a different number of strokes, which is difficult to determine in advance. In contrast, we *implicitly* learn the desired number of strokes by training two MLP networks to achieve a desired trade-off between simplicity and fidelity with respect to the input image.

3. Method

Given an input image \mathcal{I} of a scene, our goal is to produce a set of corresponding sketches at n levels of *fidelity* and m levels of *simplicity*, forming a sketch abstraction matrix of size $m \times n$. We begin by producing a set of sketches along the *fidelity* axis (Sections 3.1 and 3.2) with no simplification, thus forming the top row in the abstraction matrix.

Next, for each sketch at a given level of fidelity, we perform an iterative visual *simplification* by learning how to best remove select strokes and adjust the locations of the remaining strokes (Section 3.3). For clarity, in the following we describe our method taking into account the entire scene as a whole. However, to allow for greater control over the appearance of the output sketches, and to tackle the high complexity presented in a whole scene, our final scheme splits the image into two regions – the salient foreground object(s), and the background. We apply our 2-axes abstraction method to each region separately, and then combine them to form the matrix of sketches (details in Section 3.4).

3.1. Training Scheme

We define a sketch as a set of n strokes placed over a white background, where each stroke is a two-dimensional Bézier curve with four control points. We mark the i -th stroke by its set of control points $z_i = \{(x_i, y_i)^j\}_{j=1}^4$, and denote the set of the n strokes by $Z = \{z_i\}_{i=1}^n$. Our goal is to find the set of stroke parameters that produces a sketch adequately depicting the input scene image.

An overview of our training scheme used to produce a single sketch image is presented in the gray area of Figure 7. We train an MLP network, denoted by MLP_{loc} , that receives an initial set of control points $Z_{init} \in R^{n \times 4 \times 2}$ (marked in blue) and returns a vector of offsets $MLP_{loc}(Z_{init}) = \Delta Z \in R^{n \times 4 \times 2}$ with respect to the initial stroke locations. The final set of control points are then given by $Z = Z_{init} + \Delta Z$, which are then passed to a differentiable rasterizer \mathcal{R} [21] that outputs the rasterized sketch,

$$\mathcal{S} = \mathcal{R}(Z_{init} + \Delta Z). \quad (1)$$

For initializing the locations of the n strokes, we follow the saliency-based initialization introduced in Vinker *et al.* [43], in which, strokes are initialized in salient regions based on a relevancy map extracted automatically [7].

To guide the training process, we leverage a pre-trained CLIP model due to its capabilities of encoding shared information from both sketches and natural images. As opposed to Vinker *et al.* [43] that use the ResNet-based [15] CLIP model for the sketching process (and struggles with depicting a scene image), we find that the ViT-based [9] CLIP model is able to capture the global context required for generating a coherent sketch of a whole scene, including both foreground and background. This also follows the observation of Raghu *et al.* [35] that ViT models better capture more global information at lower layers compared to ResNet-based models. We further analyze this design choice in the supplementary material.

The loss function is then defined as the L2 distance between the activations of CLIP on the image \mathcal{I} and sketch \mathcal{S} at a layer ℓ_k :

$$\mathcal{L}_{CLIP}(\mathcal{S}, \mathcal{I}, \ell_k) = \| CLIP_{\ell_k}(\mathcal{S}) - CLIP_{\ell_k}(\mathcal{I}) \|_2^2. \quad (2)$$

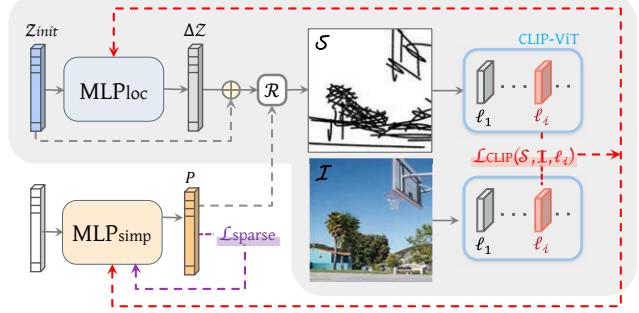


Figure 7. Single sketch generation scheme. In gray, we show our training scheme for producing a single sketch image at a single level of fidelity. In the bottom left we show the additional components used to generate a single sketch at a single level of simplicity.

At each step during training, we back-propagate the loss through the CLIP model and the differentiable rasterizer \mathcal{R} whose weights are frozen, and only update the weights of MLP_{loc} . This process is repeated iteratively until convergence. Observe that no external dataset is needed for guiding the training process, as we rely solely on the expressiveness and semantics captured by the pre-trained CLIP model. This training scheme produces a *single* sketch image at a *single* level of fidelity and simplicity. Below, we describe how to control these two axes of abstraction.

3.2. Fidelity Axis

To achieve different levels of fidelity, as illustrated by a single row in our abstraction matrix, we select different activation layers of the CLIP-ViT model for computing the loss defined in Equation (2). Optimizing via deeper layers leads to sketches that are more semantic in nature and do not necessarily confine to the precise geometry of the input. Specifically, in all our examples we train a separate MLP_{loc} using layers $\{\ell_2, \ell_7, \ell_8, \ell_{11}\}$ of CLIP-ViT and set the number of strokes to $n = 64$. Note that it is possible to use the remaining layers to achieve additional fidelity levels (see the supplementary material).

3.3. Simplicity Axis

Given a sketch \mathcal{S}_k at fidelity level k , our goal is to find a set of sketches $\{\mathcal{S}_k^1, \dots, \mathcal{S}_k^m\}$ that are visually and conceptually similar to \mathcal{S}_k but have a gradually simplified appearance. In practice, we would like to learn how to best remove select strokes from a given sketch and refine the locations of the remaining strokes without harming the overall recognizability of the sketch.

We illustrate our sketch simplification scheme for generating a single simplified sketch \mathcal{S}_k^j in the bottom left region of Figure 7. We train an additional network, denoted as MLP_{simp} (marked in orange), that receives a random-valued vector and is tasked with learning an n -dimensional vector $P = \{p_i\}_{i=1}^n$, where $p_i \in [0, 1]$ represents the prob-

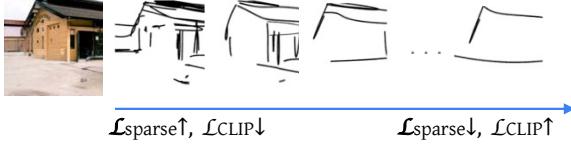


Figure 8. Trade-off between $\mathcal{L}_{\text{sparse}}$ and $\mathcal{L}_{\text{CLIP}}$. As the sketch becomes sparser, $\mathcal{L}_{\text{sparse}}$ obtains lower score. However, the sketch also becomes less recognizable with respect to the input image, resulting in a higher penalty for $\mathcal{L}_{\text{CLIP}}$.

ability of the i -th stroke appearing in the rendered sketch. P is passed as an additional input to \mathcal{R} which outputs the simplified sketch \mathcal{S}_k^j in accordance.

To implement the probabilistic-based removal or addition of strokes (which are discrete operations) into our learning framework, we multiply the width of each stroke z_i by p_i . When rendering the sketch, strokes with a very low probability will be “hidden” due to their small width.

Similar to Mo *et al.* [28], to encourage a sparse representation of the sketch (*i.e.* one with fewer strokes) we minimize the normalized L1 norm of P :

$$\mathcal{L}_{\text{sparse}}(P) = \frac{\|P\|_1}{n}. \quad (3)$$

To ensure that the resulting sketch still resembles the original input image, we additionally minimize the $\mathcal{L}_{\text{CLIP}}$ loss presented in Equation (2), and continue to fine-tune MLP_{loc} during the training of MLP_{simp} . Formally, we minimize the sum:

$$\mathcal{L}_{\text{CLIP}}(\mathcal{S}_k^j, \mathcal{I}, \ell_k) + \mathcal{L}_{\text{sparse}}(P). \quad (4)$$

We back-propagate the gradients from $\mathcal{L}_{\text{CLIP}}$ to both MLP_{loc} and MLP_{simp} while $\mathcal{L}_{\text{sparse}}$ is used only for training MLP_{simp} (as indicated by the red and purple dashed arrows in Figure 7).

Note that using the MLP network rather than performing a direct optimization over the stroke parameters (as is done in Vinker *et al.*) is crucial as it allows the optimization to restore strokes that may have been previously removed. If we were to use direct optimization, the gradients of deleted strokes would remain removed since they were multiplied by a probability of 0.

Here, both MLP networks are simple 3-layer networks with SeLU [18] activations. For MLP_{simp} we append a Sigmoid activation to convert the outputs to probabilities.

Balancing the Losses. Naturally, there is a trade-off between $\mathcal{L}_{\text{CLIP}}$ and $\mathcal{L}_{\text{sparse}}$, which affects the appearance of the simplified sketch (see Figure 8). We utilize this trade-off to gradually alter the level of simplicity.

Finding a balance between $\mathcal{L}_{\text{sparse}}$ and $\mathcal{L}_{\text{CLIP}}$ is essential for achieving recognizable sketches with varying de-

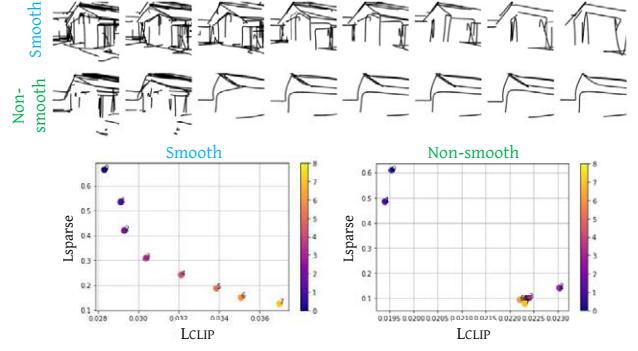


Figure 9. Smooth v.s. non-smooth simplification. In the first row, the simplification appears perceptually smooth, where a consistent change in the degree of abstraction is performed. The second row demonstrates a non-smooth simplification, as there is a visible “jump” between the second and third sketches. These visual patterns are illustrated quantitatively in the corresponding graphs, where each dot in the graph represents a single sketch.

gress of abstraction. Thus, we define the following loss:

$$\mathcal{L}_{\text{ratio}} = \left\| \frac{\mathcal{L}_{\text{sparse}}}{\mathcal{L}_{\text{CLIP}}} - r \right\|_2^2, \quad (5)$$

where the scalar factor r (denoting the *ratio* of the two losses) controls the strength of simplification. As we decrease r , we encourage the network to output a sparser sketch and vice-versa. The final objective for generating a single simplified sketch \mathcal{S}_k^j , is then given by:

$$\mathcal{L}_{\text{simp}} = \mathcal{L}_{\text{CLIP}} + \mathcal{L}_{\text{sparse}} + \mathcal{L}_{\text{ratio}}. \quad (6)$$

To achieve the set of gradually simplified sketches $\{\mathcal{S}_k^1, \dots, \mathcal{S}_k^m\}$, we define a set of corresponding factors $\{r_k^1, \dots, r_k^m\}$ to be applied in Equation (5). The first factor r_k^1 , is derived directly from Equation (5), aiming to reproduce the strength of simplification present in \mathcal{S}_k :

$$r_k^1 = \frac{1}{\mathcal{L}_{\text{CLIP}}(\mathcal{S}_k, \mathcal{I}, \ell_k)}, \quad (7)$$

where $\mathcal{L}_{\text{sparse}}$ equal to 1 means that no simplification is performed. The derivation of the remaining factors r_k^j is described next.

Perceptually Smooth Simplification As introduced above, the set of factors determines the strength of the visual simplification. When defining the set of factors r_k^j , we aim to achieve a *smooth* simplification. By *smooth* we mean that there is no large change perceptually between two consecutive steps. This is illustrated in Figure 9, where the first row provides an example of *smooth* transitions, and the second row demonstrates a non-smooth transition, where there is a large perceptual “jump” in the abstraction level between the second and third sketches, and almost no perceptual change in the following levels.

We find that the simplification appears more smooth

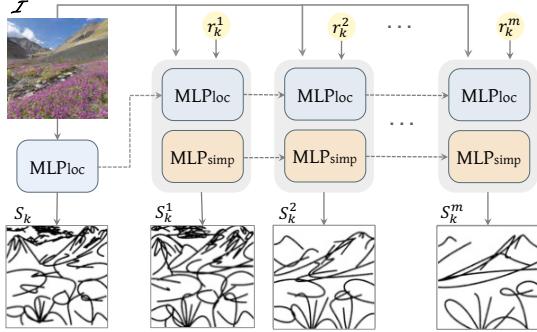


Figure 10. Iterative simplification of the sketch S_k . To produce a simplified sketch S_k^j we iteratively fine-tune MLP_{loc} (blue) and MLP_{simp} (orange) w.r.t \mathcal{L}_{ratio} loss defined by each r_k^j .

when \mathcal{L}_{sparse} is exponential with respect to \mathcal{L}_{CLIP} . The two graphs at the bottom of Figure 9 describe this observation quantitatively, illustrating the trade-off between \mathcal{L}_{sparse} and \mathcal{L}_{CLIP} for each sketch. The smooth transition in the first row forms an exponential relation between \mathcal{L}_{sparse} and \mathcal{L}_{CLIP} , while the large ‘‘jump’’ in the second row is clearly shown in the right graph.

Given this, we define an exponential function recursively by $f(j) = f(j - 1)/2$. The initial value of the function is defined differently for each fidelity level k as $f_k(1) = r_k^1$. To define the following set of factors $\{r_k^2, \dots, r_k^m\}$ we sample the function f_k , where for each k , the sampling step size is set proportional to the strength of the \mathcal{L}_{CLIP} loss at level k . Hence, layers that incur a large \mathcal{L}_{CLIP} value are sampled with a larger step size. We found this procedure achieves simplifications that are perceptually smooth. This observation aligns well with the Weber-Fechner law [10, 45] which states that human perception is linear with respect to an exponentially-changing signal. An analysis of the factors and additional details regarding our design choices are provided in the supplementary material.

Generating the Simplified Sketches To generate the set of simplified sketches $\{S_k^1, \dots, S_k^m\}$, we apply the training procedure iteratively, as illustrated in Figure 10. We begin with generating S_k^1 w.r.t r_k^1 by fine-tuning MLP_{loc} and training MLP_{simp} from scratch. After generating S_k^1 , we sequentially generate each S_k^j for $2 \leq j \leq m$ by continuing training both networks for 500 steps and applying \mathcal{L}_{ratio} with the corresponding factor r_k^j .

3.4. Decomposing the Scene

The process described above takes the entire scene as a whole. However, in practice, we separate the scene’s foreground subject from the background and sketch each of them independently. We use a pretrained U²-Net [32] to extract the salient object(s), and then apply a pretrained LaMa [41] inpainting model to recover the missing regions (see Figure 11, top right).

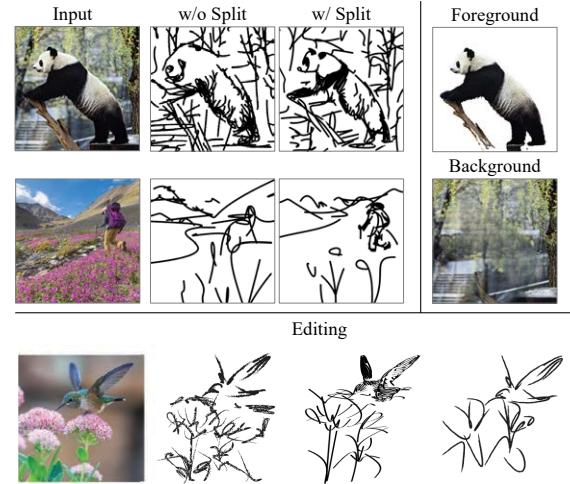


Figure 11. Scene decomposition. Top right – an example of the separation technique. Left – scene sketching results obtained with and without decomposing the scene. Bottom – examples of sketch editing by modifying the style of strokes.

We find that this separation helps in producing more visually pleasing and stable results. When performing object sketching, we additionally compute \mathcal{L}_{CLIP} over layer l_4 . This helps in preserving the object’s geometry and finer details. On the left part of Figure 11 we demonstrate the artifacts that may occur when scene separation is not applied. For example, over-exaggeration of features of the subject at a low fidelity level, such as the panda’s face, or, the object might ‘‘blend’’ into the background. Additionally, this explicit separation provides users with more control over the appearance of the final sketches (Figure 11, bottom right). For example, users can easily edit the brush’s style or combine the foreground and background sketches at different levels of abstraction.

4. Results

In the following, we demonstrate the performances of our scene sketching technique qualitatively and quantitatively, and provide comparisons to state-of-the-art sketching methods. Further analysis, results, and a user study are provided in the supplementary material.

4.1. Qualitative Evaluation

In Figures 1 and 13 we show sketches at different levels of abstraction on various scenes generated by our method. Notice how it is easy to recognize that the sketches are depicting the same scene even though they vary significantly in their abstraction level.

In Figures 1, 4 and 12 (top) we show sketch abstractions along the *fidility* axis, where the sketches become less precise as we move from left to right, while still conveying the semantics of the images (for example the mountains in

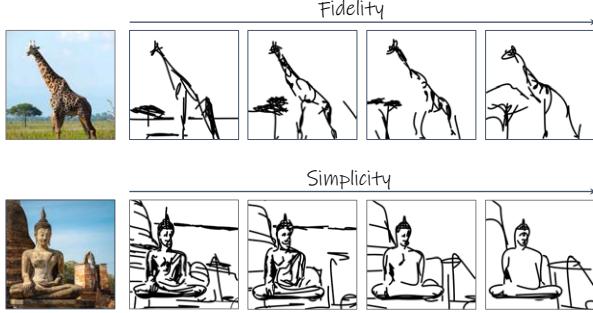


Figure 12. Sketches along the two abstraction axes.

the background in Figure 1 and the tree and giraffe’s body in Figure 12). In Figures 1, 5 and 12 (bottom) we show sketch abstractions along the *simplicity* axis. Our method successfully simplifies the sketches in a smooth manner, while still capturing the core characteristics of the scene. For example, notice how the shape of the Buddha sculpture is preserved across all levels. Observe that these simplifications are achieved *implicitly* using our iterative sketch simplification technique. Please refer to the supplemental file for many more results.

4.2. Comparison with Existing Methods

In Figure 13 we present a comparison to CLIPasso [43]. For a fair comparison, we use our scene decomposition technique to separate the input images into foreground and background and use CLIPasso to sketch each part separately before combining them. Also, since CLIPasso requires a predefined number of strokes as input, we set the number of strokes in CLIPasso to be the same as that learned implicitly by our method. For each image we show two sketches with two different levels of abstraction. As expected, CLIPasso is able to portray objects accurately, as it was designed for this purpose. However, in some cases, such as the sofa, CLIPasso fails to depict the object at a higher abstraction level. This drawback may result from the abstraction being learned from scratch per a given number of strokes, rather than gradually. Additionally, in most cases CLIPasso completely fails to capture the background, even when using many strokes (*e.g.* in the first and fourth rows). Our method captures both the foreground and the background in a visually pleasing manner, thanks to our *learned* simplification approach. For example, our method is able to convey the notion of the buildings in the first row or mountains in the second row with only a small number of simple scribbles. Similarly, our approach successfully depicts the subjects across all scenes.

In Figure 14 we present a comparison with three state-of-the-art methods for scene sketching [6, 20, 50]. On the left, as a simple baseline, we present the edge maps of the input images obtained using XDoG [46]. On the right, we present three sketches produced by our method depicting



Figure 13. Comparison to CLIPasso [43]. Note how CLIPasso fails to capture the background in most cases, especially at higher abstraction levels, despite having the same stroke budget.

three representative levels of abstraction.

The sketches produced by UPDG [50] and Chan *et al.* [6] are detailed, closely following the edge maps of the input images (such as the buildings in row 2). These sketches are most similar to the sketches shown in the leftmost column of our set of results, which also align well with the input scene structure. The sketches produced by Photo-Sketching [20] are less detailed and may lack the semantic meaning of the input scene. For example, in the first row, it is difficult to identify the sketch as being that of a person. Importantly, none of the alternative *scene* sketching approaches can produce sketches with varying abstraction levels, nor can they produce sketches in vector format. We note that in contrast to the methods considered in Figure 14, our method operates per image and requires no training data. However, this comes with the disadvantage of longer running time, taking 6 minutes to produce one sketch on a commercial GPU.

4.3. Quantitative Evaluation

In this section, we provide a quantitative evaluation of our method’s ability to produce sketch abstractions along both the simplicity and fidelity axes. To this end, we collected a variety of images spanning five classes of scene imagery: people, urban, nature, indoor, and animals, with seven images for each class. For each image, we created the 4×4 sketch abstraction matrix – resulting in a total of 560 sketches, and created sketches using the different methods presented in Section 4.2. To make a fair comparison with CLIPasso we generated sketches with four levels of abstraction, using the average number of strokes obtained by our method at the four simplicity levels. For UPDG and Chan *et al.*, we obtained sketches with three different styles, and averaged the quantitative scores across the three styles, as they represent the same abstraction level. For Photo-Sketching

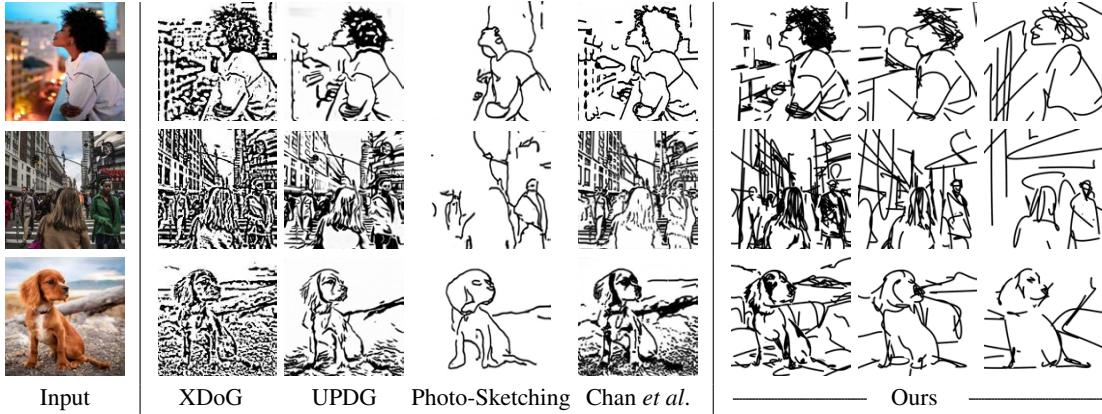


Figure 14. Scene sketching results and comparisons. From left to right are the sketches obtained using XDoG [46], UPDG [50], Photo-Sketching [20], and Chan *et al.* [6]. On the right, are three representative sketches produced by our method depicting three levels of abstraction. Note that UPDG and Chan *et al.* can produce sketches with three different styles, however all the sketches represent a similar level of abstraction. We choose one representative style but provide more style comparisons in the supplementary material.

Table 1. Comparison of the average MS-SSIM score, computed between the edge map of the input images and generated sketches.

	Ours				CLIPasso	UPDG	Chan et al.	Photo- Sketch
	Fidelity							
Simplicity	0.39	0.23	0.22	0.17	0.21			
	0.37	0.23	0.21	0.19	0.21			
	0.36	0.22	0.20	0.18	0.15	0.57	0.55	0.27
	0.34	0.22	0.18	0.14	0.13			

only one level of abstraction and one style is supported.

Fidelity Changes To measure the fidelity level of the generated sketches, we compute the MS-SSIM [44] score between the edge map of each input image (extracted using XDoG) and the corresponding sketch. In Table 1 we show the average resulting scores among all categories, where a higher score indicates a higher similarity. Examining the results matrix of our method, as we move right along the fidelity axis, the scores gradually decrease. This indicates that the sketches become “looser” with respect to the input geometry, as desired. The sketches by UPDG and Chan *et al.* obtained high scores, which is consistent with our observation that their method produces sketches that follow the edges of the input image. The scores for CLIPasso show that the fidelity level of their sketches does not change much across simplification levels and is similar to the fidelity of sketches of our method at the last two levels (the two rightmost columns). This suggests that CLIPasso is not capable of producing large variations of fidelity abstractions.

Sketch Recognizability A key requirement for successful abstraction is that the input scene will remain recognizable in the sketches across different levels of abstraction. To evaluate this, we devise the following recognition ex-

periment on the set of images described above. Using a pre-trained ViT-B/16 CLIP model (different than the one used for training), we performed zero-shot image classification over each input image and the corresponding resulting sketches from the different methods. We use a set of 200 class names taken from commonly used image classification and object detection datasets [19, 24] and compute the percent of sketches where at least 2 of the top 5 classes predicted for the input image were also present in the sketch’s top 5 classes. We consider the top 5 predicted classes since a scene image naturally contains multiple objects.

Table 2 shows the average recognition rates across all images for each of the described methods. The recognizability of the sketches produced by our method remains relatively consistent across different simplicity and fidelity levels, with a naturally slight decrease as we increase the simplicity level. We do observe a large decrease in the recognition score in the first column. This discrepancy can be attributed to the first fidelity level following the image structure closely, which makes it more difficult to depict the scene with fewer strokes. CLIPasso’s fidelity level is most similar to our two rightmost columns (as shown in Table 1). When comparing our recognition rates along these columns to the results of CLIPasso, one can observe that at higher simplicity levels, their method loses the scene’s semantics.

5. Conclusions

We presented a method for performing scene sketching with different types and multiple levels of abstraction. We disentangled the concept of sketch abstraction into two axes: *fidelity* and *simplicity*. We demonstrated the ability to cover a wide range of abstractions across various challenging scene images and the advantage of using vector representation and scene decomposition to allow for greater artis-

Table 2. Recognizability scores, using a CLIP ViT-B/16 model for zero-shot classification on the input image and generated sketches.

	Ours				CLIPasso	UPDG	Chan et al.	Photo-Sketch
	Fidelity							
Simplicity	0.92	1.00	0.95	0.97	0.92			
	0.54	0.97	1.00	0.91	0.83			
	0.54	0.93	0.94	0.89	0.70	0.87	0.91	0.62
	0.44	0.79	0.91	0.85	0.43			

tic control. It is our hope that our work will open the door for further research in the emerging area of computational generation of visual abstractions. Future research could focus on further extending these axes and formulate innovative ideas for controlling visual abstractions.

References

- [1] AUTOMATIC1111. Stable diffusion webui. <https://github.com/AUTOMATIC1111/stable-diffusion-webui>, 2022. 32
- [2] Itamar Berger, Ariel Shamir, Moshe Mahler, Elizabeth Carter, and Jessica Hodgins. Style and abstraction in portrait sketching. *ACM Trans. Graph.*, 32(4), jul 2013. 2, 3
- [3] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan, Jorma Laaksonen, and Michael Felsberg. Doodleformer: Creative sketch drawing with transformers. *ECCV*, 2022. 3
- [4] Irving Biederman and Ginny Ju. Surface versus edge-based determinants of visual recognition. *Cognitive psychology*, 20(1):38–64, 1988. 1
- [5] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, pages 679–698, 1986. 3
- [6] Caroline Chan, Frédéric Durand, and Phillip Isola. Learning to generate line drawings that convey geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7915–7925, 2022. 3, 7, 8, 13, 14, 33, 34, 35, 39
- [7] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 387–396, 2021. 4
- [8] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. *CoRR*, abs/1711.02257, 2017. 13
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 4, 14
- [10] Gustav Theodor Fechner. Elemente der psychophysik. 1860. 6
- [11] Kevin Frans, Lisa B. Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *CoRR*, abs/2106.14843, 2021. 3
- [12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 32
- [13] Songwei Ge, Vedanuj Goswami, Larry Zitnick, and Devi Parikh. Creative sketch generation. In *International Conference on Learning Representations*, 2021. 3
- [14] David Ha and Douglas Eck. A neural representation of sketch drawings. In *International Conference on Learning Representations*, 2018. 3
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4
- [16] Aaron Hertzmann. Why do line drawings work? a realism hypothesis. *Perception*, 49(4):439–451, mar 2020. 1
- [17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 31
- [18] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *Advances in neural information processing systems*, 30, 2017. 5
- [19] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 8
- [20] Mengtian Li, Zhe Lin, Radomir Mech, Ersin Yumer, and Deva Ramanan. Photo-sketching: Inferring contour drawings from images. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1403–1412. IEEE, 2019. 3, 7, 8, 13, 14, 39
- [21] Tzu-Mao Li, Michal Lukáč, Gharbi Michaël, and Jonathan Ragan-Kelley. Differentiable vector graphics rasterization for editing and learning. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 39(6):193:1–193:15, 2020. 4
- [22] Yijun Li, Chen Fang, Aaron Hertzmann, Eli Shechtman, and Ming-Hsuan Yang. Im2pencil: Controllable pencil illustration from photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1525–1534, 2019. 3
- [23] Yi Li, Yi-Zhe Song, Timothy M. Hospedales, and Shaogang Gong. Free-hand sketch synthesis with deformable stroke models. *CoRR*, abs/1510.02644, 2015. 3
- [24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 8
- [25] Difan Liu, Matthew Fisher, Aaron Hertzmann, and Evangelos Kalogerakis. Neural strokes: Stylized line drawing of 3d shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14204–14213, 2021. 3

- [26] Vijish Madhavan. Artline. <https://github.com/vijishmadhavan/ArtLineTechnical-Details>, 2020. 3
- [27] Daniela Mihai and Jonathon Hare. Learning to draw: Emergent communication through sketching. *Advances in Neural Information Processing Systems*, 34:7153–7166, 2021. 3
- [28] Haoran Mo, Edgar Simo-Serra, Chengying Gao, Changqing Zou, and Ruomei Wang. General virtual sketching framework for vector line art. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2021)*, 40(4):51:1–51:14, 2021. 5
- [29] Umar Riaz Muhammad, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. Learning deep sketch abstraction. *CoRR*, abs/1804.04804, 2018. 2, 3
- [30] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 31
- [31] Yonggang Qi, Guoyao Su, Pinaki Nath Chowdhury, Mingkang Li, and Yi-Zhe Song. Sketchlattice: Latticed representation for sketch manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 953–961, 2021. 3
- [32] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition*, 106:107404, 2020. 6, 12
- [33] Shuwen Qiu, Sirui Xie, Lifeng Fan, Tao Gao, Song-Chun Zhu, and Yixin Zhu. Emergent graphical conventions in a visual communication game. *arXiv preprint arXiv:2111.14210*, 2021. 3
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 2, 3, 14
- [35] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021. 4
- [36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 31
- [37] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 31
- [38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 31, 32
- [39] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 31
- [40] Jifei Song, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. Learning to sketch with shortcut cycle consistency, 2018. 3
- [41] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2149–2159, 2022. 6, 12
- [42] Zhengyan Tong, Xuanhong Chen, Bingbing Ni, and Xiaohang Wang. Sketch generation with drawing process guided by vector flow and grayscale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 609–616, 2021. 3
- [43] Yael Vinker, Ehsan Pajouheshgar, Jessica Y. Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. Clipasso: Semantically-aware object sketching. *ACM Trans. Graph.*, 41(4), jul 2022. 2, 3, 4, 7, 13, 14, 17, 32
- [44] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multi-scale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003, volume 2, pages 1398–1402. Ieee, 2003. 8
- [45] E H Weber. De pulsu, resorptione, auditu et tactu. 1834. 6
- [46] Holger Winnemöller, Jan Eric Kyprianidis, and Sven C. Olsen. Xdog: An extended difference-of-gaussians compendium including advanced image stylization. *Comput. Graph.*, 36:740–753, 2012. 3, 7, 8
- [47] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 3
- [48] Peng Xu, Timothy M Hospedales, Qiyue Yin, Yi-Zhe Song, Tao Xiang, and Liang Wang. Deep learning for free-hand sketch: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3
- [49] Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2019. 3
- [50] Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. Unpaired portrait drawing generation via asymmetric cycle mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8217–8225, 2020. 3, 7, 8, 33, 35, 39
- [51] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 31

- [52] Tao Zhou, Chen Fang, Zhaowen Wang, Jimei Yang, Byung-moon Kim, Zhili Chen, Jonathan Brandt, and Demetri Terzopoulos. Learning to sketch with deep q networks and demonstrated strokes. *ArXiv*, abs/1810.05977, 2018. 3

Appendix

Table of Contents

A Implementation Details	12
A.1 Image Preprocessing	12
A.2 MLP Training	12
A.3 Matrix Composition	13
B User Study	13
C Additional Quantitative Analysis	14
C.1. Sketch Recognizability	14
C.2. Number of Strokes by Simplicity Level	15
D Ablation Study: General Design Choices	17
D.1. ViT vs. ResNet	17
D.2. Foreground-Background Separation	17
E Ablation Study: Simplicity Axis	19
E.1. Explicitly Defining the Number of Strokes	20
E.2. Replace the Ratio Loss With a Target Number of Strokes	20
E.3. Fine-tuning MLP-loc During Simplification	21
E.4. Defining the Function f_k as an Exponential	21
E.5. Defining a Different Set of Factors for Each Layer	22
E.6. Defining a Different Sampling Step for Each f_k	23
F Ablation Study: Fidelity Axis	24
F.1. Using I-4 for Object Sketching	24
F.2. Using Other ViT Layers for Training	24
G Additional Results	24
H Additional Comparisons	31
H.1. Diffusion Models	31
H.2 Scene Sketching Approaches	32

A. Implementation Details

In this section, we provide specific details about the implementation of our method. We will further release all code and image sets used for evaluations to facilitate further research and comparisons.

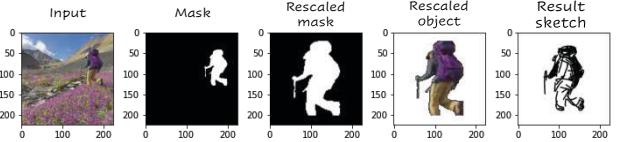


Figure 15. Object re-scaling procedure.

A.1. Image Preprocessing

As stated in the paper, we use a pre-trained U²-Net salient object detector [32] to extract the scene’s salient object(s). To receive a binary map, we threshold the resulting map from U²-Net such that pixels with a value smaller than 0.5 are classified as background, and the remaining pixels are classified as salient objects. We then use this mask as an input to a pre-trained LaMa [41] inpainting model to recover the missing regions in the background image.

Object Scaling In the case where the saliency detection process detected only one object, and this object fills less than 70% of the image size, we perform an additional preprocessing step to increase the size of the object before sketching. This assists in sketching key features of the object when using a large number of strokes. Specifically, we first take the masked object and compute its bounding box. We then shift the masked object to the center of the image and resize the object such that it covers $\approx 70\%$ of the image. We apply the sketching procedure on the scaled object and then resize and shift the resulting sketch back to the original location in the input image. Note that since our sketches are given in vector representation, it is possible to re-scale and shift them without changing their resolution. This process is illustrated in Figure 15.

A.2. MLP Training

Hyper-parameters In all experiments, we set the number of strokes to $n = 64$ in the first phase of sketching and train MLP_{loc} for 2,000 iterations. For generating the series of simplified sketches (Section 3.4 in the main paper), we perform 8 iterative steps. As discussed in Section 3.4, for each fidelity level k , we define a separate function f_k for defining the set of ratios used in \mathcal{L}_{ratio} . Along with this function, we define a separate step size for sampling the function f_k . For simplifying the background sketches, we set this step size to be $\{0.35, 0.45, 0.5, 0.9\}$ for layers $\{2, 7, 8, 11\}$, respectively. For simplifying the object sketches, we set the

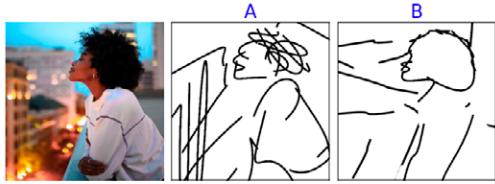


Figure 16. Example of how images were presented to participants in the user study.

step sizes to be $\{0.45, 0.4, 0.5, 0.9\}$. Each simplification step is obtained by training MLP_{simp} and MLP_{loc} for 500 iterations. We employ the Adam optimizer with a constant learning rate of $1e-4$ for training both MLP networks. The input to MLP_{simp} is set to be a random-valued vector of dimension n .

Augmentations As also done in Vinker *et al.* [43], we apply random affine augmentations (*i.e.* random perspective and random cropping transformations) to both the input image and generated sketch before passing them as inputs to the CLIP model for computing the loss.

GradNorm We train MLP_{simp} and MLP_{loc} with three different losses simultaneously in order to achieve our visual simplifications. As these losses compete with each other, training has the potential to be highly unstable. For example, when training with multiple losses, the gradients of one loss may be stronger than the other, resulting in the need to weigh the losses accordingly. To help achieve a more stable training process and to ensure that each loss contributes equally to the optimization process, we use GradNorm [8], which automatically balances the training process by dynamically adjusting the gradient magnitudes. This balancing is achieved by weighing the losses inversely proportional to their contribution to the overall gradient.

A.3. Matrix Composition

As stated in the main paper, we separate the scene into two regions (based on their saliency map) and apply the sketching scheme to both independently, we then combine the resulting sketches to form the final matrix. To combine the foreground and background we simply aggregate the corresponding strokes at a given level of fidelity and simplicity. Note that we also export the mask used to separate them, if the user wish to locate it behind the object to avoid the collision of strokes. We also export the separate matrices, to allow users to combine sketches from different levels of abstraction as a post process.

Ours v.s. CLIPasso		Ours v.s. Chan <i>et al.</i>		Ours v.s. Photo-Sketching	
Ours	84.8%	Ours	52.5%	Ours	75.6%
CLIPasso	7.0%	Chan <i>et al.</i>	27.3%	Photo-Sketching	15.8%
Equal	8.2%	Equal	20.2%	Equal	8.6%

Table 3. Results of our user study. We compare 30 sketches produced by our method to three alternative methods: CLIPasso [43], Chan *et al.* [6], and Photo-Sketching [20]. For each method, we specify the percent of responses that preferred our sketch, the sketch of the alternative method, or found the sketches to be similar in their ability to capture the scene semantics.

B. User Study

In Section 4.3 of the main paper, we presented a quantitative evaluation measuring the fidelity and recognizability of our sketches based on automatic metrics. As opposed to the fidelity measure, which can be determined by measuring the distance from the edge map of the input scene, validating the recognizability is more challenging. To this end, we also conducted a user study to further validate the findings presented by the CLIP zero-shot classification approach.

The user study examines how well the sketches depict the input scene, considering both the foreground and background. Using 30 images from the set described in Section 4.3, we compared our sketches with three alternative methods: CLIPasso [43], Chan *et al.* [6], and Photo-Sketching [20]. The participants were presented with the input image along with two sketches, one produced by our method and the other by the alternative method (with the sketches presented in random order). An example is provided in Figure 16. The following question was posed to participants:

Which sketch better depicts the image content?

In your answer, please relate to:

- (1) *Preservation of both foreground and background.*
- (2) *Semantic preservation - *i.e.*, reflecting the meaning of the elements.*

Participants could choose between three options: “A”, “B”, and “A and B at a similar level”.

It is important to note that in order to make a fair comparison, we compared the methods which produce abstract sketches (CLIPasso [43] and Photo-Sketching [20]) with our abstract sketches (in row 4 of the matrix, *i.e.*, at the highest abstraction level). Conversely, we compared the sketches of Chan *et al.* [6], which are more detailed and have greater fidelity, to our sketches in their most detailed form (top left corner of the matrix). We applied CLIPasso

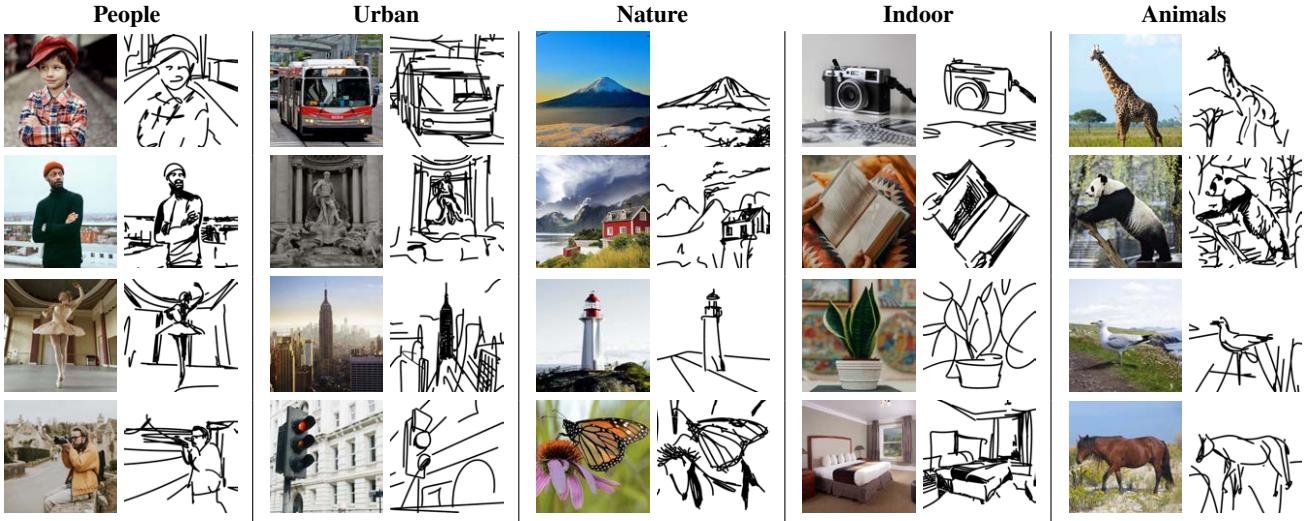


Figure 17. Example images and representative sketches used for our quantitative evaluations.

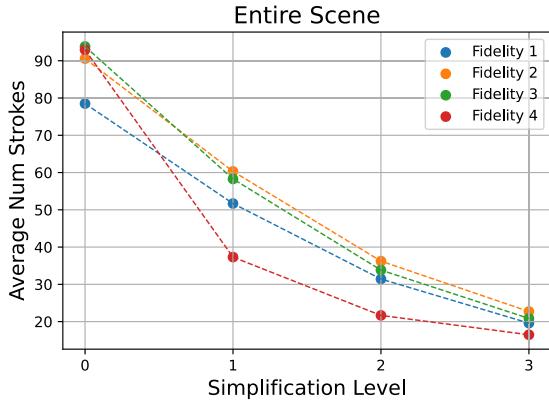


Figure 18. Examining the number of strokes used to compose the sketch across each fidelity and simplification level. Results are averaged across all images across all scene categories. In the supplementary materials, we additionally illustrate the number of strokes split between foreground and background and between the five scene categories.

using our scene decomposition technique, and the same number of strokes as learned by our method, and for Chan *et al.* we used the contour style sketches. We collected responses from 25 participants for the survey, which contained a total of 90 questions (i.e., 2,250 responses were collected in total).

The average resulting scores among all participants and images are shown in Table 3. In the first, second, and third columns, we show the scores obtained when comparing to CLIPasso [43], Chan *et al.* [6], and Photo-Sketching [20] respectively. Compared to CLIPasso and Photo-Sketching, our method achieved significantly higher rates (84.8% and

75.6% of responses favored our method over the respective alternatives). Conversely, only 7% and 15.8% of responses preferred the results of the alternative method, respectively. Although sketches produced by Chan *et al.* [6] are highly detailed, 52.5% of the responses preferred our sketches, while only 20% considered our sketches and Chan *et al.*'s sketches to be similar.

The results of the user study support the findings presented in the main paper, demonstrating that sketches produced by our method faithfully capture both the foreground and background elements in the scene among varying abstraction levels.

C. Additional Quantitative Analysis

In this section, we provide additional details, examples, and results regarding the quantitative evaluations presented in the paper. First, in Figure 17 we present example inputs and representative generated sketches for each of our five scene categories used for evaluations.

C.1. Sketch Recognizability

To compute our recognizability metrics, we perform zero-shot classification using a pre-trained ViT-B/16 [9] CLIP model. Observe this model is different than the ViT-B/32 model used to generate sketches, ensuring a more fair evaluation of our sketches. When performing the zero-shot classification, we follow the evaluation setup used in CLIP [34] and apply 80 prompt templates when defining our 200 classes to CLIP's text encoder. This includes prompts of the form: “a rendering of a {}”, “a drawing of a {}”, and “a sketch of a {}”. We then compute the cosine similarity between all text embeddings and the embedding corresponding to either our input image or generated sketches.

In Figure 19, we present example zero-shot classification results obtained on various input images and sketches across our five scene categories.

C.2. Number of Strokes by Simplicity Level

We examine our method’s ability to generate sketches at varying levels of simplicity. For that purpose, we measure the final number of strokes used to generate the sketches. We extract this information from the generated sketch SVGs across all 560 sketches. We present the results in Figure 18, split between the different fidelity levels (indicated by different colors) and simplicity levels (shown along the x-axis). As can be seen, the number of strokes decreases as we move along the simplicity axis, across all fidelity levels.

In Figure 20, we present the same results but split between the different scene categories and split between composing the foreground and background sketches. As can be seen, the resulting functions for the different fidelity levels follow an exponential relation as we strengthen the simplification level.

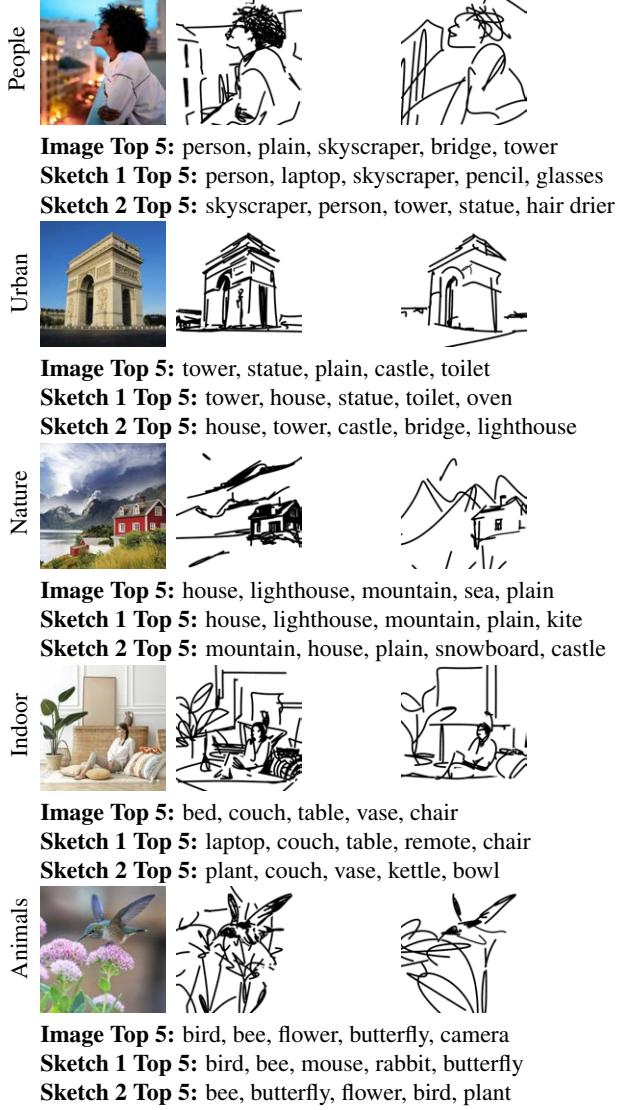


Figure 19. Examples of CLIP zero-shot class predictions on various input images and representative sketches of varying abstractions. These predictions are then used to compute a recognizability metric for each scene category across different levels of abstractions (see Section 4.3 in the main paper).

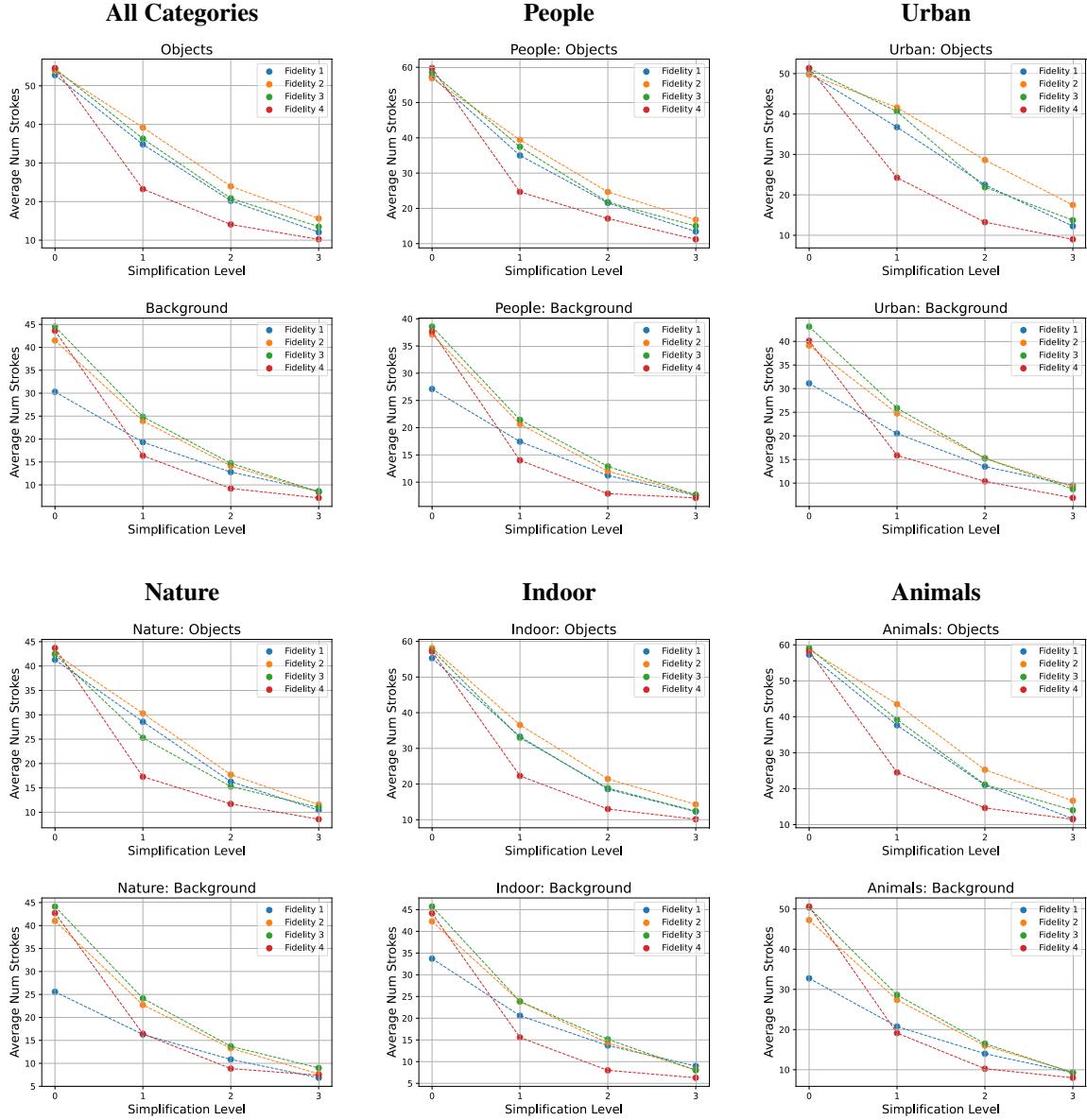


Figure 20. Examining the number of strokes used to compose the sketch across each fidelity and simplification level, split between the different scene categories.

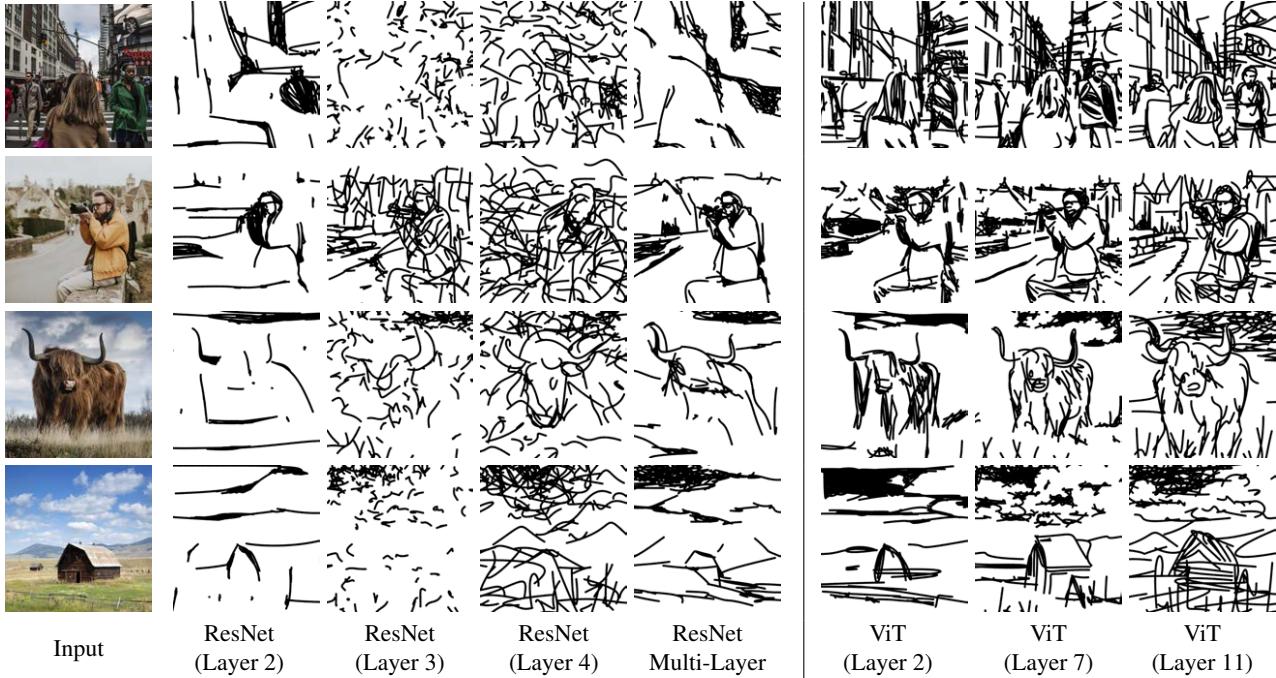


Figure 21. Ablation study on using ResNet-CLIP and ViT-CLIP for guiding the training process. For both variants, we generate the sketches for the entire scene together (*i.e.* no scene decomposition is performed) and set the number of strokes to 128. In addition to evaluating ResNet using a single layer for computing \mathcal{L}_{CLIP} , we additionally show results obtained when multiple ResNet layers are used to compute \mathcal{L}_{CLIP} (marked as “ResNet Multi-Layer”). For this variant, we follow CLIPasso [43] as set the layer weights to 0, 0, 1, 1, 0 for layers ℓ_1 to ℓ_5 , respectively. In addition, we use the output of the fully connected layer and set its weight to 0.1 in the loss computation.

D. Ablation Study: General Design Choices

D.1. ViT vs. ResNet

In Figure 21, we demonstrate the scene sketching results obtained with a ResNet101-based CLIP compared to those obtained with the ViT-based CLIP model employed in this work. When computing \mathcal{L}_{CLIP} using a single layer of ResNet (*i.e.* layer 2, 3, or 4), we are unable to capture the input scene, indicating that a combination of the layers must be used for capturing the more global details of a complete scene. However, even when computing \mathcal{L}_{CLIP} using multiple ResNet layers (as was done in CLIPasso), the network still struggles in capturing the details of the scene. For example, in row 3, although we are able to roughly capture the outline of the bull’s head and horns, the network is unable to capture the bull’s body and scene background. In contrast, when replacing the ResNet model with the more powerful ViT model, we are able to capture both the scene foreground and background, even when using a single layer for computing the loss. This naturally allows us to control the level of fidelity of the generated sketch by simply altering the single ViT layer that is used for computing \mathcal{L}_{CLIP} .

D.2. Foreground-Background Separation

In Section 3.4 of the main paper, we introduced our scene decomposition technique where the foreground and background components of the input are sketched separately and then merged. In Figure 22 we provide additional sketching results obtained with and without the scene decomposition for both abstraction axes. At the top, we present the resulting sketches along the fidelity axis. Observe how the house in the leftmost sketch in the first row appears to disappear within the entire scene. Furthermore, note the artifacts that appear in the face of the dog as the abstraction increases. In contrast, by explicitly separating the foreground and background, we can apply additional constraints over the foreground sketches to help mitigate unwanted artifacts. As a result, we are able to better maintain the correct structure of both the house and dog in the provided examples.

At the bottom of Figure 22 we show the resulting sketches along the simplification axis. Note how the house in the first row almost disappears completely, and that there are not enough strokes to depict the mountains in the background. By considering the entire scene as a whole, the model has no explicit control over how to balance the level of details placed between the object and the background. As a result, more strokes are typically used to sketch the

background, which consumes a larger portion of the entire image (and therefore leads to a larger reduction in \mathcal{L}_{CLIP}).

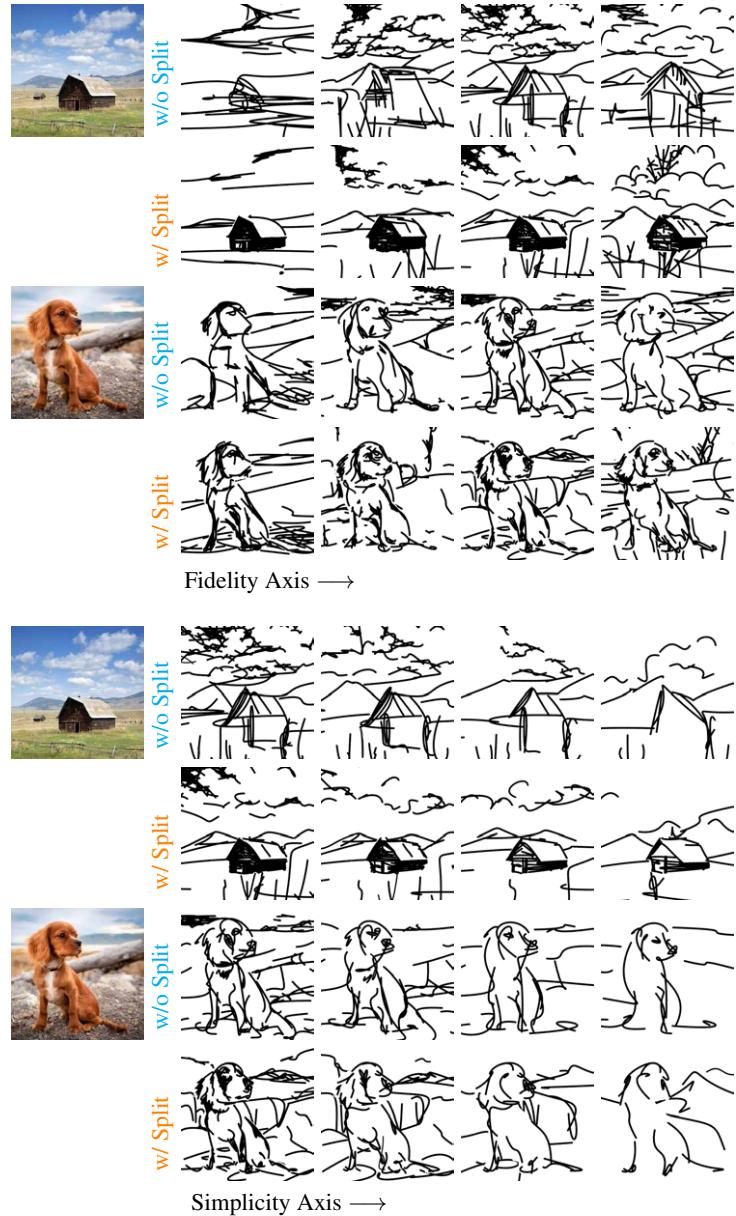


Figure 22. Scene sketching results obtained with and without decomposing the scene. We show sketches generated across both the fidelity axis (first four rows) and the simplicity axis (bottom four rows).

E. Ablation Study: Simplicity Axis

In this section, we analyze the design choices for the simplification training scheme and corresponding loss objectives.

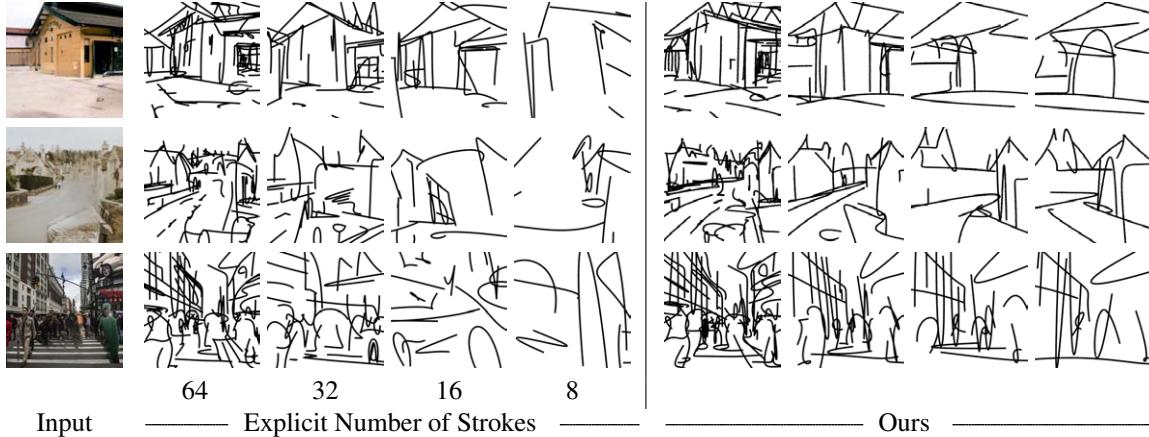


Figure 23. Ablation results of explicitly defining the number of strokes (left) compared to using our implicit simplification scheme (right). The presented sketches of all three input images are of the last fidelity level (obtained using layer 11 of CLIP-ViT).



Figure 24. Ablation results of replacing our \mathcal{L}_{ratio} loss with an alternative loss $\|\mathcal{L}_{sparse} - n_{target}\|$ guided by a desired number of strokes n_{target} .

E.1. Explicitly Defining the Number of Strokes

We begin by analyzing our simplification scheme as a whole. That is, are we able to achieve a smooth simplification of an input scene by simply varying the number of strokes used to sketch the scene? In Figure 23 we provide results comparing our implicit simplification scheme (on the right) with results obtained by sketching the scene using a varying number of strokes defined in advance (on the left). For the latter results, we use 64, 32, 16, and 8 strokes for sketching. For each input, we present the simplifications achieved at the last level of fidelity (*i.e.* using layer 11 of CLIP-ViT for training).

Knowing in advance the number of strokes needed to achieve a specific level of abstraction is often challenging and varies between different inputs. For example, consider the image in the first row, containing a simpler scene of a house. When using only 16 strokes for sketching this image, we are able to capture the components of the scene such as the existence of the house in the center, its roof, and its door. However, when sketching the more complex urban scene in the third row, using 16 strokes may struggle

to capture the general structure of the buildings or may not converge at all. By *learning* how to simplify each sketch, our simplification scheme is able to adjust the number of strokes needed to more faithfully sketch a given image at various levels of simplification, while adapting to the complexity of the input scene in the learning process.

Moreover, we observe that when defining the number of strokes explicitly, we may fail to get a smooth simplification of the initial sketch since each sketch is generated independently and may converge to a different local minimum. For example, in the second row on the left, the sketches do not appear to be simplified versions of each previous step (*e.g.* between the second and third steps), but rather new sketches of the input scene with an increased level of visual simplification. In contrast, each of our simplification results is initialized with the previous result, resulting in a smoother transition between each image.

E.2. Replace the Ratio Loss With a Target Number of Strokes

We note in the main paper that achieving gradual visual simplification requires balancing between \mathcal{L}_{sparse} and

\mathcal{L}_{CLIP} . In our method, we do so by defining a set of factors used to define a balance between the relative strengths of the losses. This section examines another possible approach: encouraging the training process to achieve a certain number of strokes during training and reducing this number at each level. As opposed to Appendix E.1 where we restrict the number of strokes completely, here we include the target number of strokes as another objective in the training process, thus allowing deviance from this number.

To implement this approach we simply redefine L_{ratio} as follows:

$$L_{ratio} = \|\mathcal{L}_{sparse} - n_{target}\|, \quad (8)$$

where n_{target} is the desired number of strokes. Specifically, we define four levels of abstraction using 64, 32, 16, and 8 strokes. We then normalize the number of strokes to be between 0 and 1 and define n_{target} as $\{1, 0.5, 0.25, 0.125\}$ for each level, respectively.

In Figure 24 we show the result of this experiment. As can be seen on the left, such an approach does not achieve the desired simplification. As can be seen, on the left, the levels of abstraction are less gradual than on the right and do not reach full abstraction. This approach also relies on an arbitrary fixed number of strokes per abstraction level for all images, as opposed to allowing the ratio itself to implicitly define it in a content-dependent manner.

E.3. Fine-tuning MLP-loc During Simplification

As described in Section 3.3 of the main paper, when performing the simplification of a given sketch by training MLP_{simp} , we continue fine-tuning MLP_{loc} . Doing so is important since by training MLP_{loc} , we allow to slightly adjust the locations of strokes in the canvas, which helps encourage the simplified sketch to resemble the original input image. In Figure 25, we show sketch simplifications across various inputs obtained with and without the fine-tuning of MLP_{loc} .

When MLP_{loc} is held fixed, the simplification process is equivalent to simply selecting a subset of strokes to remove at each step. This approach will result in the appearance of visual simplification, but may not be sufficient to maintain the semantics of the scene. For example, in the last simplification step of the first example, the mountains in the background have disappeared, as have the buildings in the third example. In addition, the house in the second image can no longer be identified. On the other hand, our results, obtained with the fine-tuning of MLP_{loc} , produce the desired visual simplification, while still preserving the same semantics of the input scenes.

E.4. Defining the Function f-k as an Exponential

In Section 3.3 of the main paper, we describe the process of selecting the set of factors used to achieve a gradual

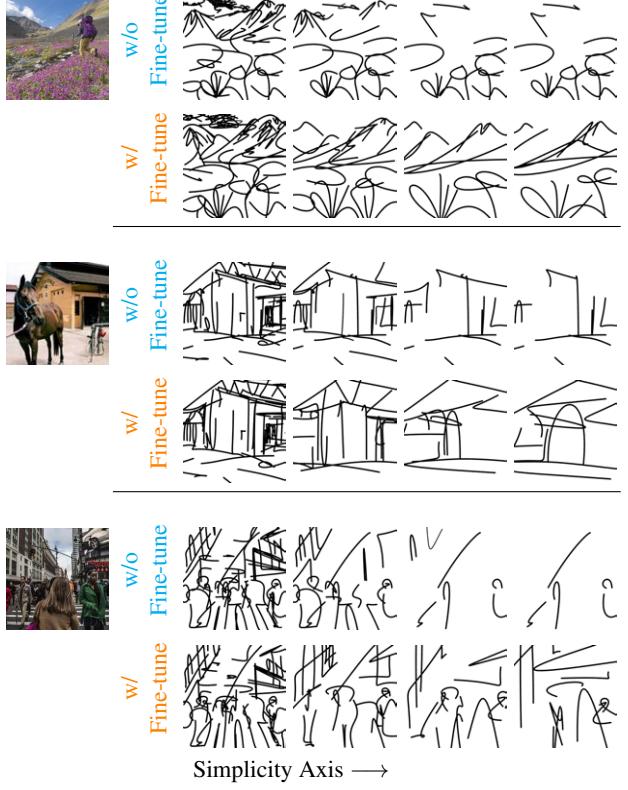


Figure 25. Ablation of fine-tuning MLP_{loc} when training MLP_{simp} during the simplification process. The sketches presented here are obtained using layer 11 of the CLIP-ViT model.

simplification of a given sketch. To do so, we defined a function f_k for each fidelity level k defining the balance between \mathcal{L}_{CLIP} and \mathcal{L}_{sparse} . We find that an f_k that models in an exponential relationship between \mathcal{L}_{CLIP} and \mathcal{L}_{sparse} achieves a simplification that is perceived smooth.

In this section, we demonstrate this effect by visually demonstrating the gradual simplification we achieve when choosing an f_k that gives a linear relation between \mathcal{L}_{CLIP} and \mathcal{L}_{sparse} . To do so, we define the linear f_k such that the sampled set of factors $\{r_k^1, \dots, r_k^m\}$ represent a constant step size by encouraging the removal of 8 strokes in each step.

In Figure 26 we present the results of this alternative setup. For each set of generated sketches, we additionally present two graphs: (1) the resulting L_{sparse} as a function of L_{CLIP} (left), and (2) the final number of strokes as a function of the simplification step (right). Each point in the graphs corresponds to a single sketch with the color of the points indicating the location of the corresponding sketch along the simplification axis. That is, 0 (or dark blue) indicates the leftmost, non-simplified sketch while 7 (or yellow) indicates the rightmost sketch with the highest level of simplification. Recall, as discussed in the main paper, the left

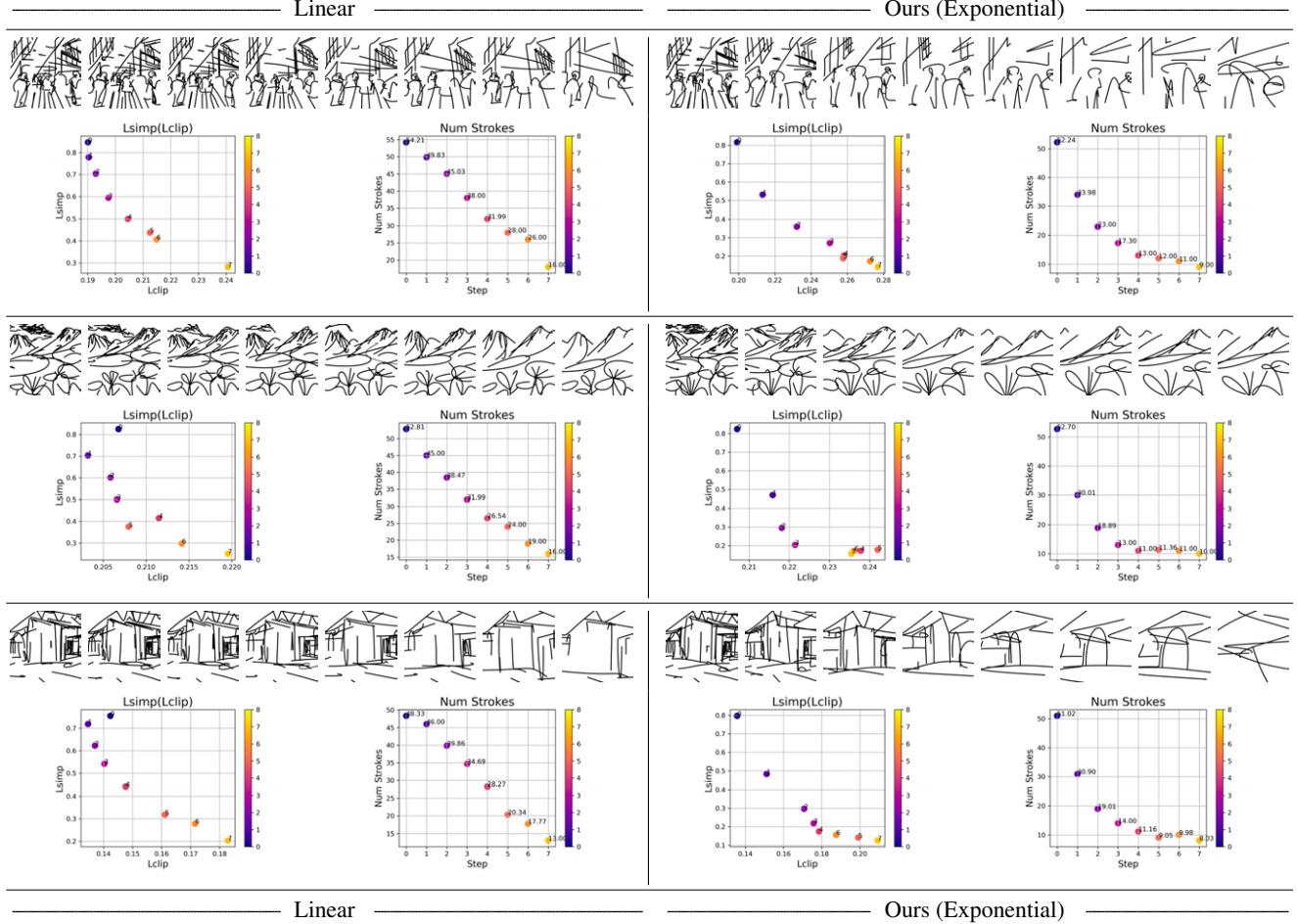


Figure 26. Ablation of defining an exponential f_k when performing an iterative simplification of the sketch. On the right-hand side, we define a linear relation between the two loss objectives. On the left side, we show sketch results obtained when defining an exponential relation between \mathcal{L}_{CLIP} and \mathcal{L}_{sparse} . For each set of simplified sketches, we additionally present two graphs depicting (1) the relation between the two loss objectives at each simplification step (left graph) and (2) the number of strokes used to compose the sketch (right graph).

graph should ideally depict an exponential relation between the two loss objectives in order for the simplification to appear smooth.

The results presented on the left side of Figure 26 show the sketches and corresponding graphs produced when using a linear f_k as defined above. The results on the right-hand side of the Figure show sketches obtained with our method when using the exponential f_k as described in the main paper. As can be seen, the sketches in the linear alternative (left) remain too detailed at the initial abstraction levels and do not convey the smooth and gradual change perceptually as is present with the exponential function (right).

E.5. Defining a Different Set of Factors for Each Layer

In Section 3.3 of the main paper, we describe the process of selecting the set of factors $\{r_k^1, \dots, r_k^m\}$ used to achieve a gradual simplification of a given sketch. In this section, we validate the use of different sets of factors for each fidelity level k . Note that, as stated in the main paper, the set of factors r_k^j determine the balance between \mathcal{L}_{CLIP} and \mathcal{L}_{sparse} , which directly determines the level of visual simplification.

We show on the left-hand side of Figure 27 the simplified sketches obtained for different levels k of fidelity when using the same set of factors. Specifically, we apply the set of factors used for layer ℓ_8 to the remaining ViT layers. As can be seen, the perceived level of visual simplification is not uniform between different layers: the simplification achieved for layer ℓ_{11} is too weak with very little perceived

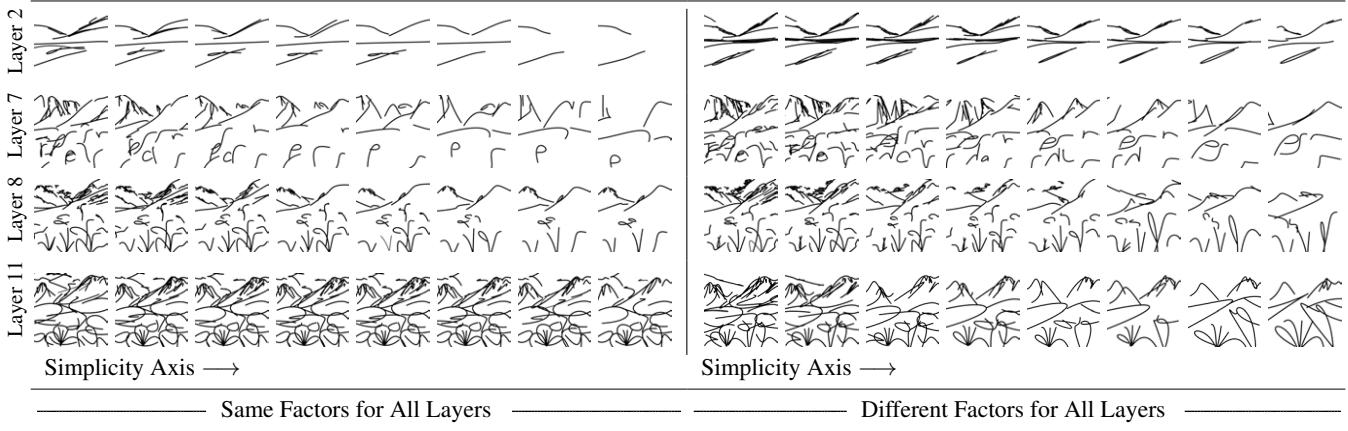


Figure 27. Ablation study on using a different set of factors for each fidelity level k when defining our \mathcal{L}_{ratio} loss. On the left, we show simplification results obtained when applying the same set of factors across all levels. On the right side, we show our simplification results obtained by adjusting the factors for each fidelity level.

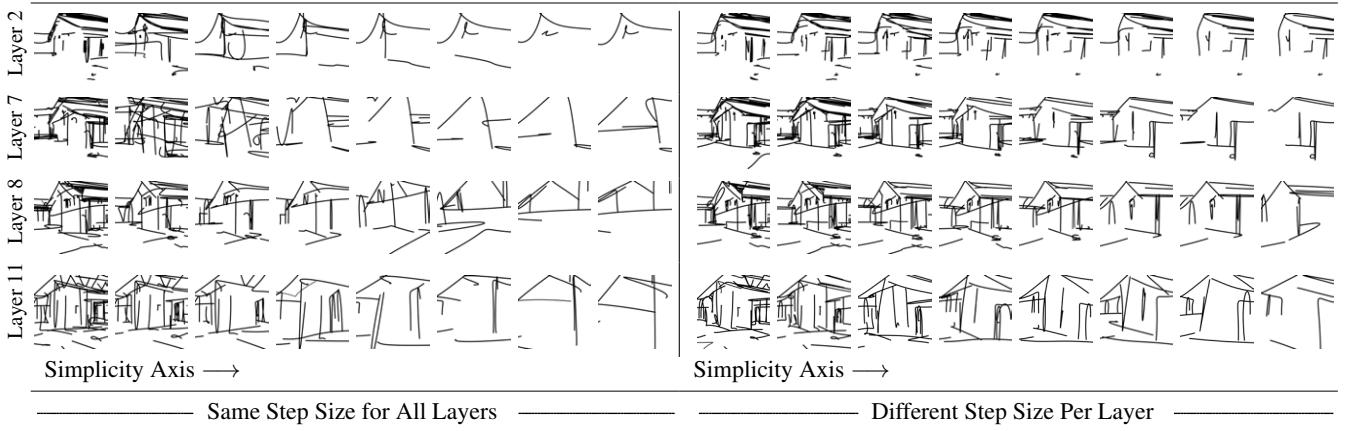


Figure 28. Ablation study on applying the same step size for sampling each function f_k during simplification. On the left, we apply the same step size across all fidelity levels, resulting in non-smooth and non-uniform simplifications between the different fidelity levels. On the right, we show our simplification results obtained by adjusting the step size for each level. As shown, we achieve smoother simplifications that are more consistent between the four different levels.

change realized across all steps. In contrast, for layer ℓ_2 the simplification is too strong with the background quickly “disappearing” as we move to the right. On the right-hand side of Figure 27, we present the results obtained with our method, where we fit a dedicated set of factors for each fidelity level k . As can be seen, the perceived level of abstraction among the different fidelity levels is more uniform and smooth.

E.6. Defining a Different Sampling Step for Each f_k

After defining the function f_k for each fidelity level k , we use a different step size for sampling this function for each k . We wish to achieve a similar appearance of simplification across different fidelity levels, and we find that in

order to do so, using a different step size for sampling f_k is crucial.

In Figure 28, on the left, we demonstrate results obtained when using the same step size for our four fidelity levels (*i.e.* ViT layers $\ell_2, \ell_7, \ell_8, \ell_{11}$). As can be seen, for layers ℓ_2 and ℓ_7 the gradual simplification is not smooth and a noticeable jump in the strength of the abstraction can be seen between the second and third sketches. Furthermore, for layer ℓ_7 , the change in step size caused the networks to converge to a noisy solution. Observe how the second sketch does not resemble a simplification of the previous one.

In contrast, the results on the right are obtained using the proposed approach of selecting different step sizes for each fidelity level k . As shown, the sketches do not suffer from the perceived artifacts present on the left. Moreover,

the simplification results are also smooth and uniform in appearance between the different layers.

F. Ablation Study: Fidelity Axis

Finally, in this section, we perform various ablation studies to validate the design choices made with respect to our fidelity abstraction axis.

F.1. Using ℓ_4 for Object Sketching

When decomposing the scene and sketching for the foreground image, we additionally compute L_{CLIP} over layer ℓ_4 of ViT. We found that doing so may help in preserving the geometry of more complex subjects, as illustrated in Figure 30. This is most noticeable in finer details such as in the facial details of the old man and the dog or the body shape of the panda, for example.

F.2. Using Other ViT Layers for Training

In order to obtain different levels of fidelity, we train MLP_{loc} guided by different layers of the CLIP-ViT model for computing \mathcal{L}_{CLIP} . Our model is based on the ViT-B/32 architecture that includes 11 intermediate layers. Our main paper presents the results of applying our training scheme to a subset of four layers: 2, 7, 8, and 11. This subset of layers represents a range of possible fidelity levels that can be achieved by our method. While we focus on presenting results using only these four layers, our method can naturally generate additional levels of fidelity by using the remaining intermediate layers. We present the results of using additional layers in Figure 29.

G. Additional Results

We begin with additional results generated by our method. In Figures 31 to 35 we provide 4×4 abstraction matrices for various scene images. In addition, we present additional examples of the added control provided by the separation technique in Figure 36. This includes: (1) editing the style of strokes using Adobe Illustrator and (2) combining the foreground and background sketches and varying levels of abstractions to achieve various artistic effects.

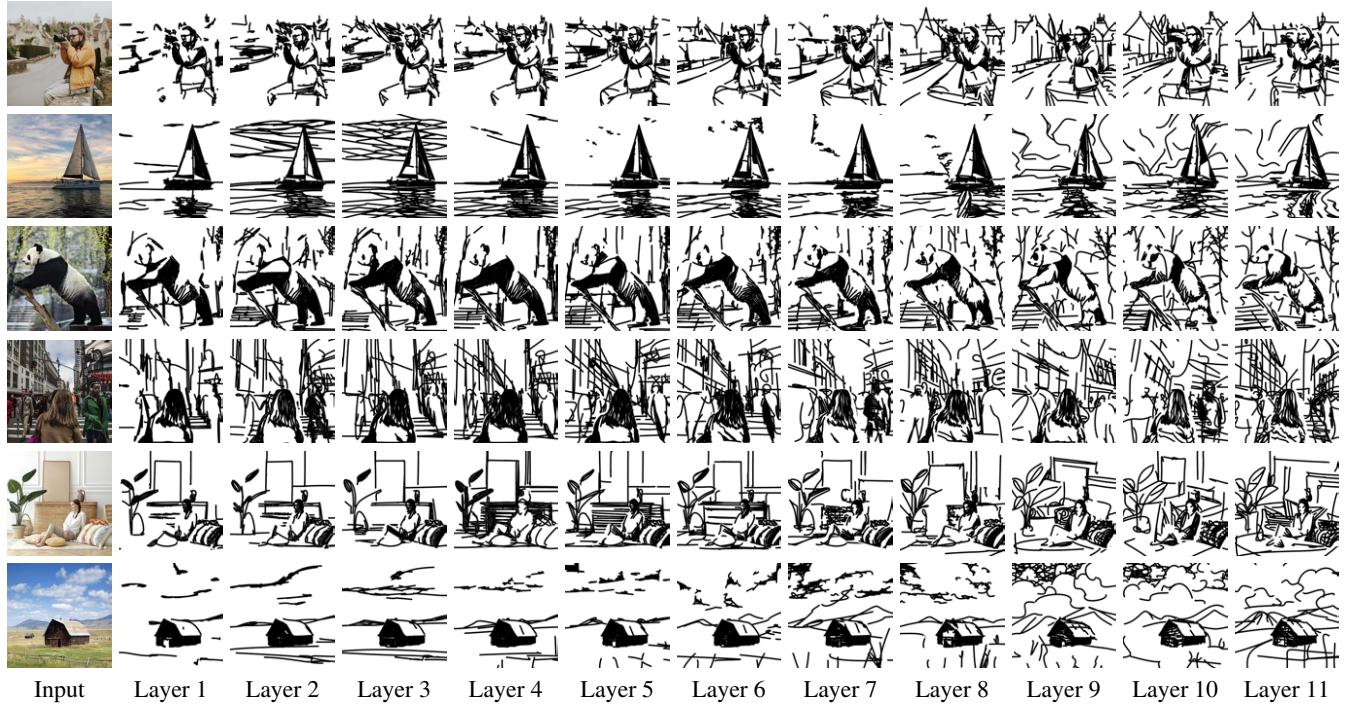


Figure 29. Ablation study on using different ViT layers for computing \mathcal{L}_{CLIP} for generating sketches at different levels of fidelity.



Figure 30. Ablation results for object sketching when additionally using layer ℓ_4 when computing \mathcal{L}_{CLIP} for object sketching.



Figure 31. The 4×4 matrix of sketches produced by our method. Columns from left to right illustrate the change in fidelity, from precise to loose, and rows from top to bottom illustrate the visual simplification.



Figure 32. The 4×4 matrix of sketches produced by our method. Columns from left to right illustrate the change in fidelity, from precise to loose, and rows from top to bottom illustrate the visual simplification.

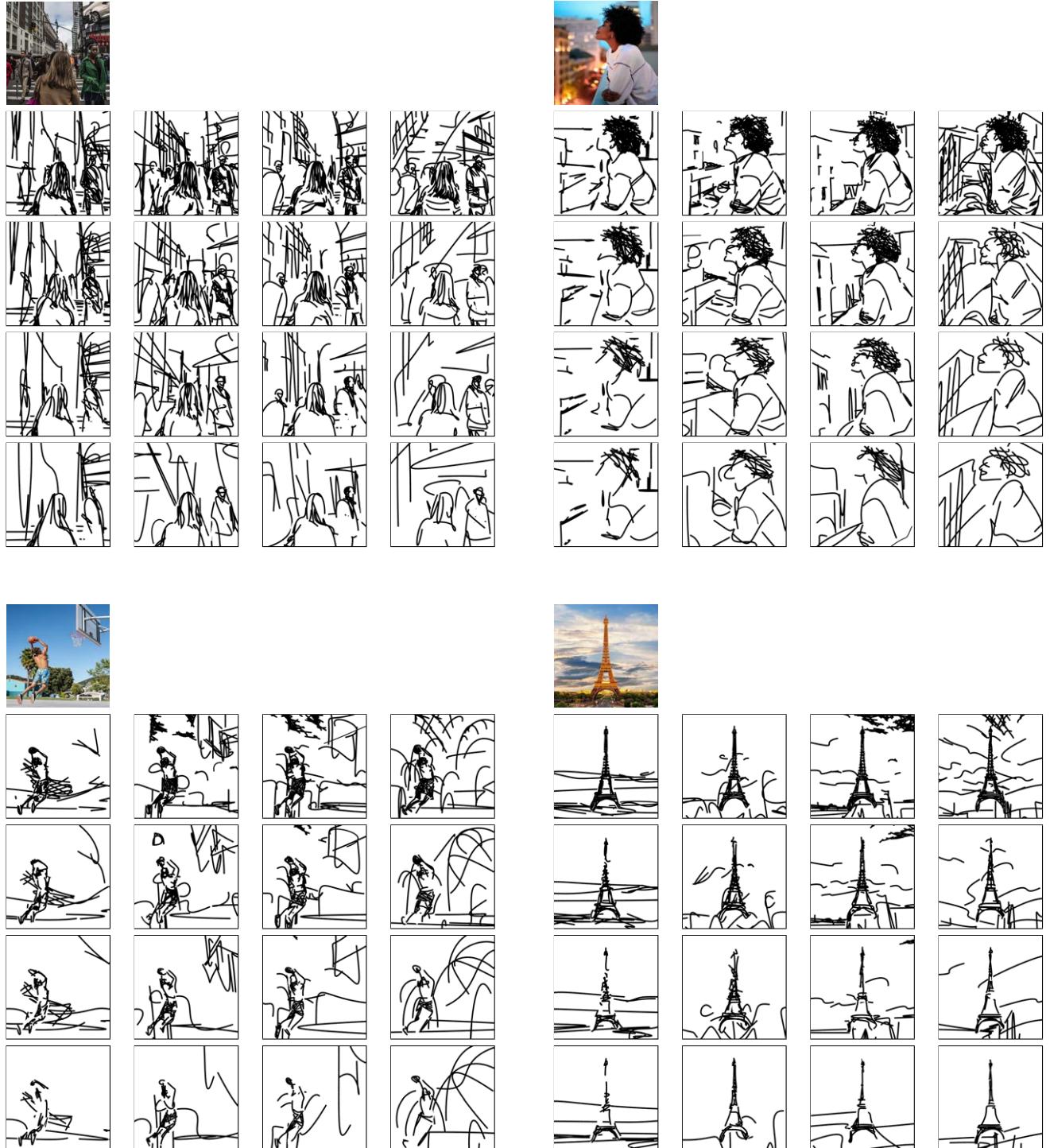


Figure 33. The 4×4 matrix of sketches produced by our method. Columns from left to right illustrate the change in fidelity, from precise to loose, and rows from top to bottom illustrate the visual simplification.

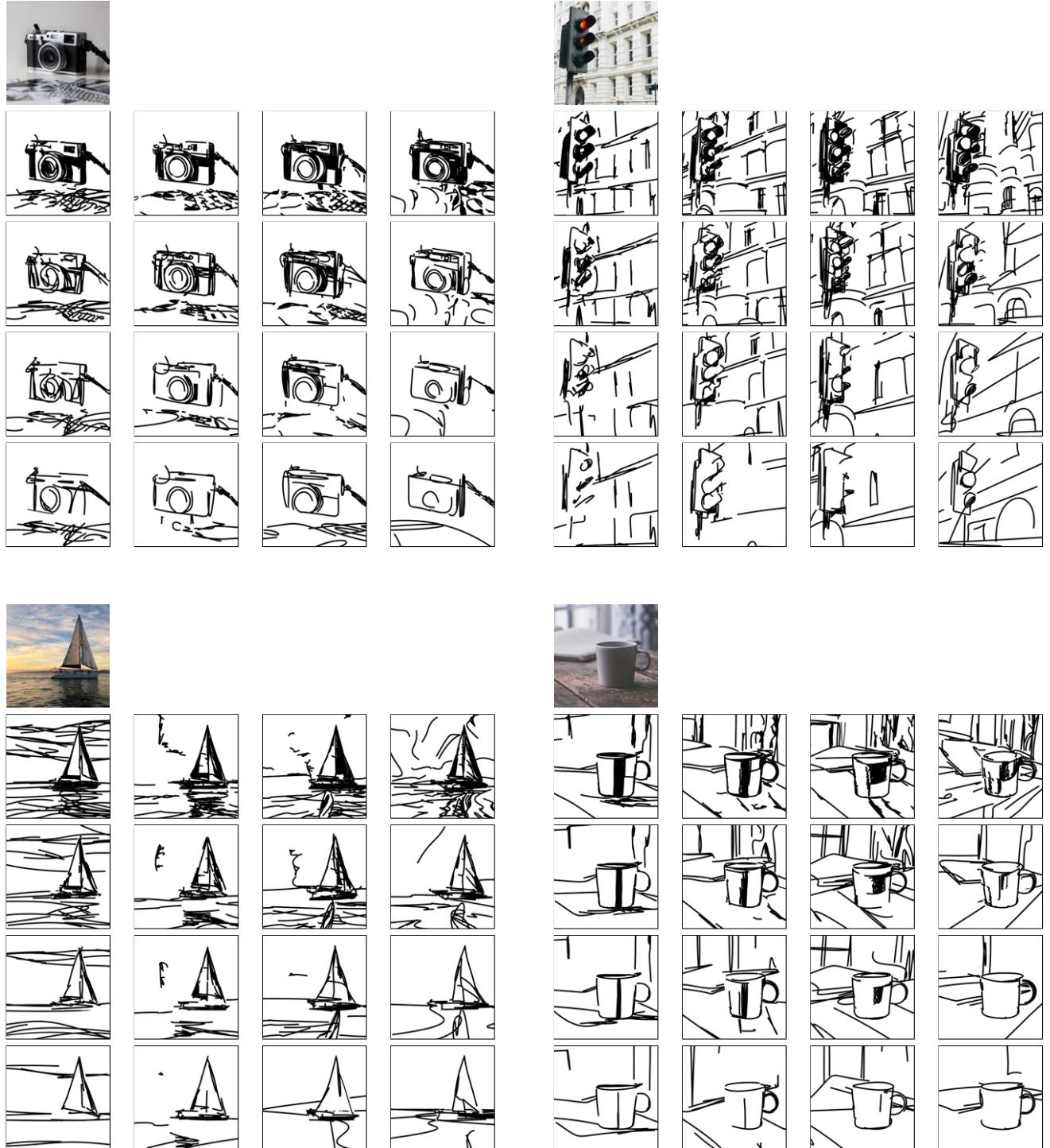


Figure 34. The 4×4 matrix of sketches produced by our method. Columns from left to right illustrate the change in fidelity, from precise to loose, and rows from top to bottom illustrate the visual simplification.



Figure 35. The 4×4 matrix of sketches produced by our method. Columns from left to right illustrate the change in fidelity, from precise to loose, and rows from top to bottom illustrate the visual simplification.

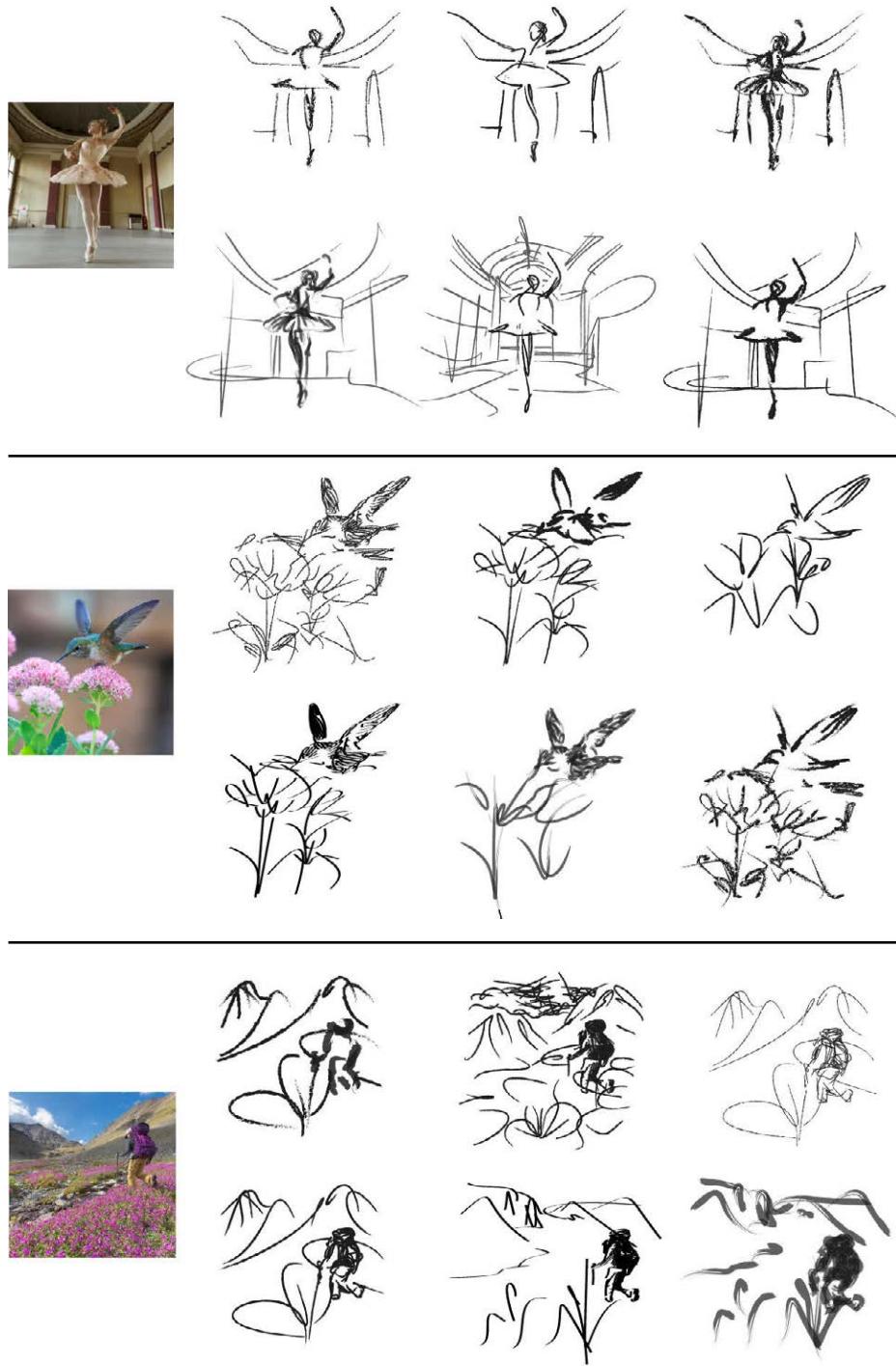


Figure 36. Additional control. For each image, we combine foreground and background sketches from different levels of abstraction, and edit the style of strokes using Adobe Illustrator. This illustrates the power of our method in providing various options for the user to edit the resulted sketches.

H. Additional Comparisons

H.1. Diffusion Models

Recent advancements in diffusion models [17] have demonstrated an unprecedented ability to generate amaz-

ing imagery guided by a target text prompt or image [30, 36–39, 51]. In this section, we explore whether such mod-

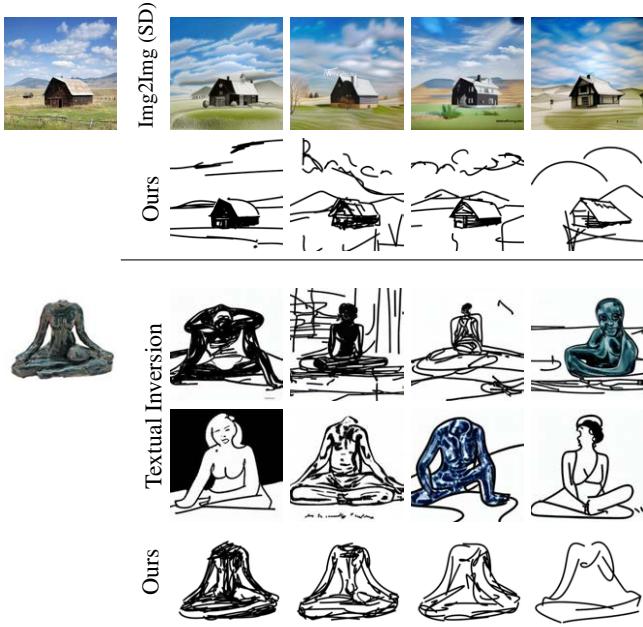


Figure 37. Comparison to various diffusion model-based techniques. At the top, we compare our sketch results to those obtained using Stable Diffusion [38] image-to-image technique guided by the text prompt “A black and white sketch image”. At the bottom, we compare with Textual Inversion [12] by learning new tokens representing a detailed sketch (top row) and abstract sketch (bottom) row. As shown, both approaches struggle in either capturing the desired sketch style or input subject.

els can be leveraged to generate abstract sketches of a given scene. We begin by exploring the recent Stable Diffusion model [38]. Given an input image, we perform a text-guided image-to-image translation of the input using text prompts such as: “A black and white sketch image” and “A black and white single line abstract sketch.” Results are illustrated in Figure 37. As can be seen in the top row, Stable Diffusion struggles in capturing the sketch style even when guiding the denoising process with keywords such as “A black and white sketch”. We do note that better results may be achieved with heavy prompt engineering or tricks such as prompt re-weighting methods [1]. However, doing so would require heavy manual overhead for each input image.

Another approach to assist in better capturing the desired sketch style would be to fine-tune the entire diffusion model on a collection of sketch images. However, this would require collecting a few hundred or thousands of images with matching captions and training a separate model for each desired style and level of abstraction.

As another diffusion-based approach, we consider the recent Textual Inversion (TI) technique from Gal *et al.* [12]. Given a few images (*e.g.* 5) of the desired style (*e.g.* sketch),

TI can be used to learn a new “word” representing the style. Users can then use a pre-trained text-to-image model such as the recent Latent Diffusion Model [38] to generate images of the learned style. For example, users can generate a sketch image of a house using the prompt “A photo of a house in the style of S_* ” where S_* represents our learned sketch style.

To evaluate TI’s ability to generate sketch images supported by our method, we collect 10 sketches generated by our method — 5 detailed and 5 abstract — and learn a new token representing each of the sketch styles. In a similar fashion, we can learn a new word representing a unique object of interest (*e.g.* the headless statue shown in Figure 37). We can then generate images of the learned object in our learned style using prompts of the form “A drawing of a S_{statue} in the style of S_{detailed} ” or “A drawing of a S_{statue} in the style of S_{abstract} ”. Example results are presented in the bottom half of Figure 37. As can be seen, TI struggles in composing both the learned style and subject in a single image. Specifically, TI either struggles in capturing the unique shape of the statue (*e.g.* its missing head) or struggles in capturing the learned sketch style (*e.g.* TI may generate images in color). In contrast, our method is able to generate a range of possible sketch abstractions that successfully capture the input subject.

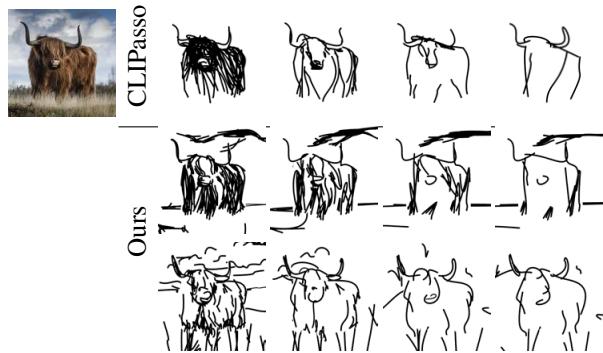


Figure 38. Comparisons with CLIPasso [43]. We show CLIPasso results obtained when applied over an entire scene image using 128 and 64 strokes respectively. We show our sketch results obtained using approximately the same number of strokes.

H.2. Scene Sketching Approaches

In Figure 38, we provide a comparison to CLIPasso. We applied CLIPasso on the masked object and obtained the abstraction by explicitly specifying the number of strokes (*i.e.* 64, 32, 16, and 8 strokes). In the second and third rows we show the simplification results obtained by our method, at two different fidelity levels. Since CLIPasso only offers a single axis of abstraction (mostly governed by simplification), the fidelity level of the sketch can not be explicitly controlled. Additionally, unlike CLIPasso, where the user

must manually determine the number of strokes required to achieve different levels of abstraction, our approach *learns* the desired number of strokes.

Lastly, observe that since each sketch of CLIPasso is generated independently, the resulting sketches may not portray a gradual, smooth simplification of the sketch since each optimization process may converge to a different local minimum. By training an MLP network to *learn* this gradual simplification, our resulting sketches depict a smoother simplification, where each sketch is a simplified version of the previous one.

In Figure 39 we provide additional scene sketching comparisons to alternative scene sketching methods. In Figure 40 we provide additional sketch comparisons to all styles supported by UPDG [50] and Chan *et al.* [6]. In Figure 41 we provide additional comparisons to CLIPasso. In Figures 42 to 44 we show the 35 sketches produced by the different sketch approaches used for the quantitative experiment. Note that in Figure 43 we show the results obtained by CLIPasso when using our scene decomposition technique, specifically, we separate the input images into foreground and background and use CLIPasso to sketch each image separately, and then combine the results.



Figure 39. Scene sketching results and comparisons.

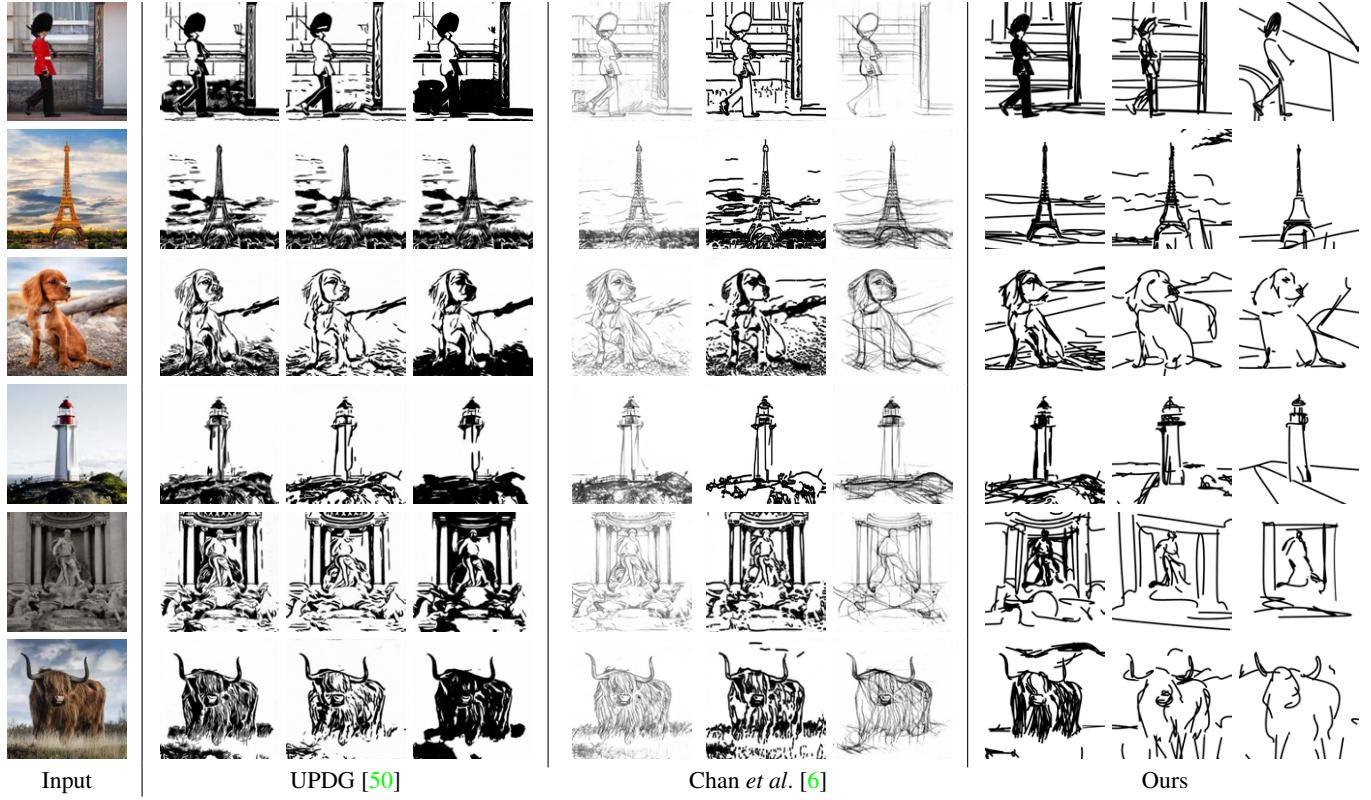


Figure 40. Scene sketching comparison to Chan *et al.* [6] and UPDG [50] across three different styles supported by each of their methods. For our results, we show three sketches illustrating the various levels of abstraction that our method is capable of achieving.

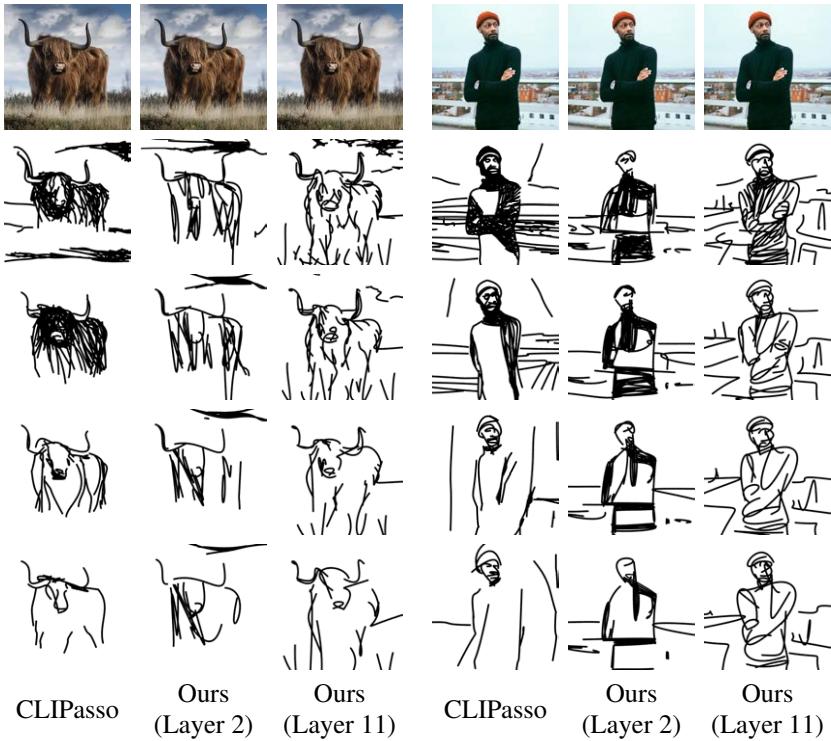
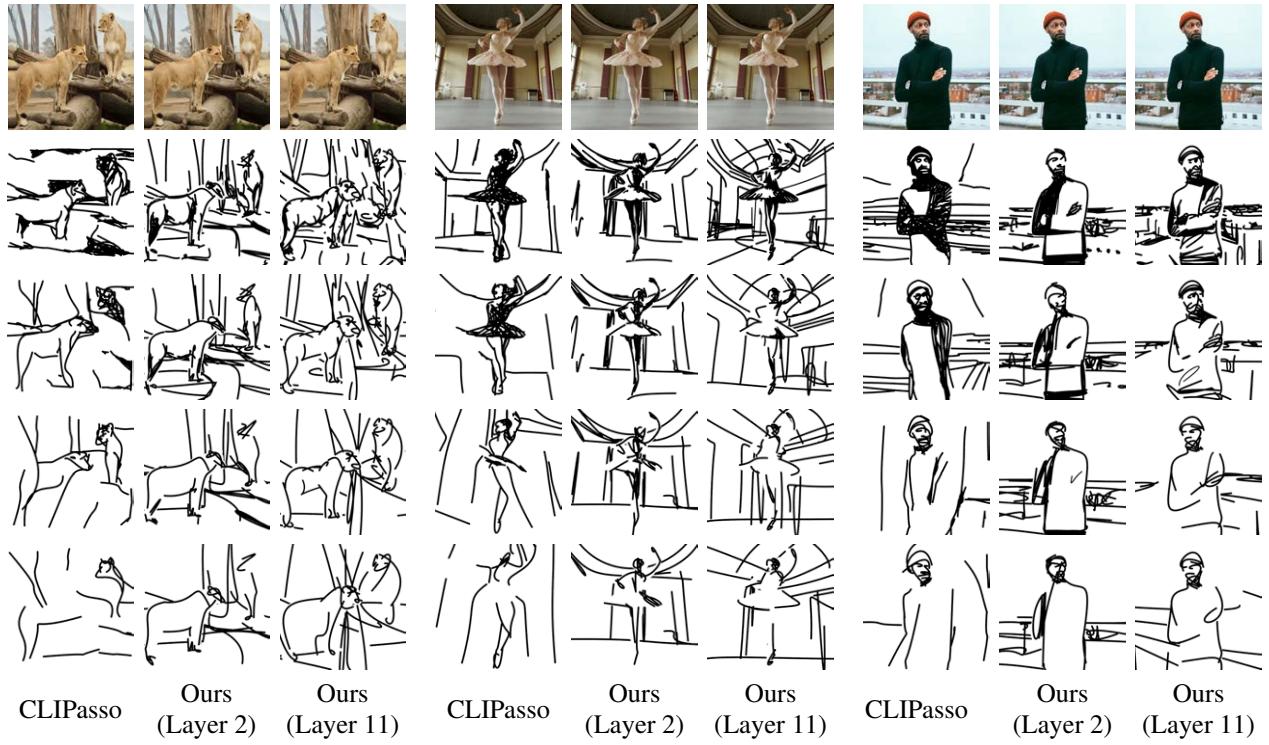


Figure 41. Comparison to CLIPasso. For CLIPasso, we generate the sketches using 128, 64, 32, and 16 strokes. In the top set of results, we use our scene decomposition technique and apply our implicit simplification starting with 64 strokes for the foreground and background sketches. For the bottom set of results, we do not use the scene decomposition approach and start with simplification using 128 strokes. We show our results for layers 2 and 11.

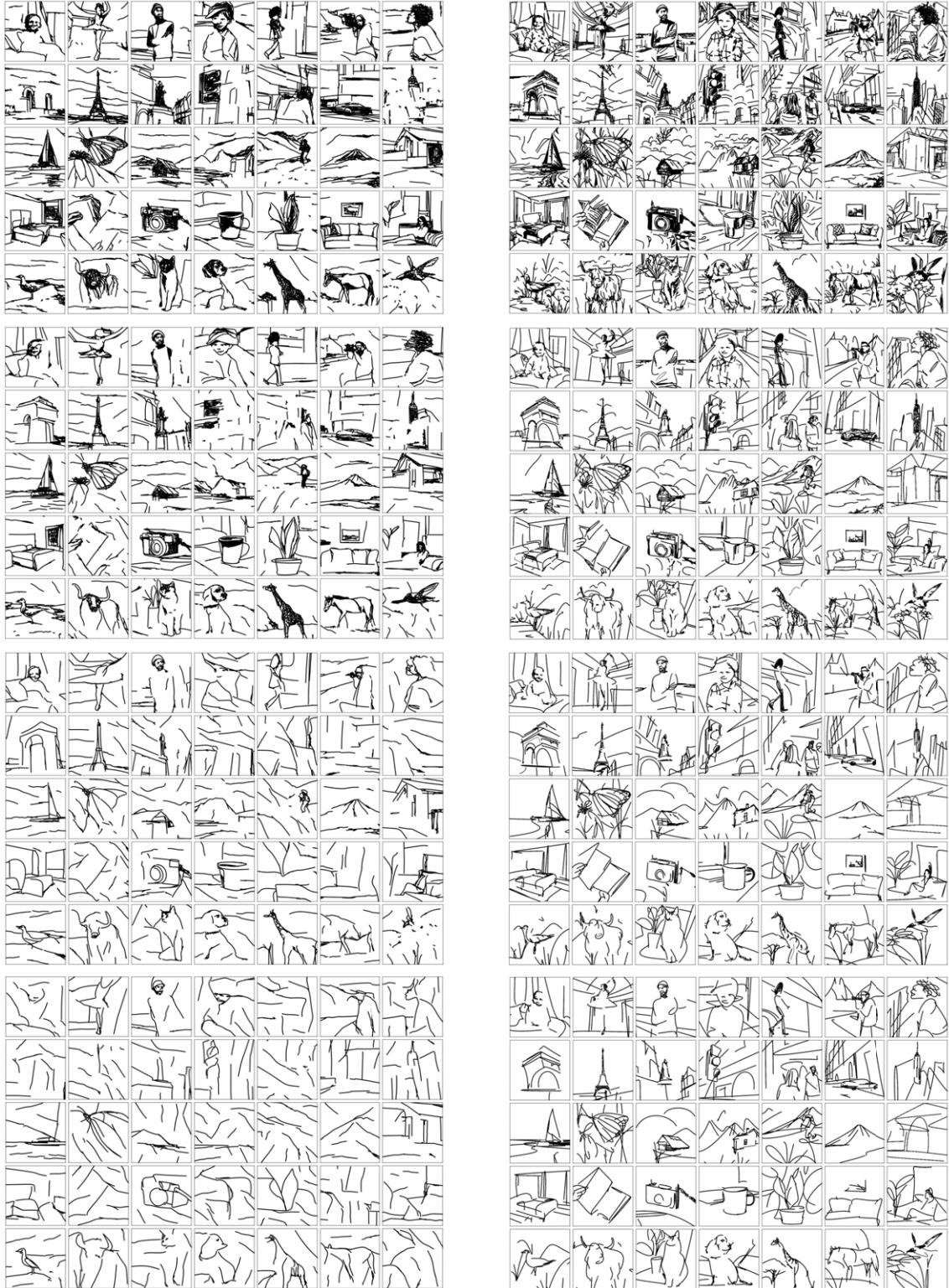


Figure 42. The 35 sketches produced for the quantitative experiment. On the left are the results of CLIPasso with four levels of abstraction, and on the right are our results with four levels of abstraction obtained using layer 11 of CLIP-ViT.



Figure 43. A comparison to CLIPasso with the scene separation technique on the 35 images used for the quantitative experiment. On the left are the results of CLIPasso with four levels of abstraction, when we separate the image into foreground and background, and sketch each of the separately. On the right are our results with four levels of abstraction obtained using layer 11 of CLIP-ViT.



Figure 44. The 35 sketches produced for the quantitative experiment. On the left are the results by Chan *et al.* [6] with the three provided styles, the last row on the left is by Photo-Sketch [20]. On the right are the results by UPDG [50] with the three provided styles.