Appendix

A Details of Implementation

A.1 MOE-Style Adapter

The MOE adapter is implemented as a set of parallel ConvNets composed of three consecutive convolution and non-linear activation layers. The entire model is comprised of nine individual MOE adapters, each of which consumes 70K parameters. Task keys are designated to each adapter, ensuring that they align with the corresponding visual conditions. Once the MOE adapter processes the input, the remaining model parameters become shared across all tasks. This architecture facilitates task adaptability while promoting parameter efficiency.

A.2 Task-aware HyperNet

The task-aware hypernet is applied to modulate the parameters of zero-conv layers in the ControlNet. Since the ControlNet can be considered as the hypernet of Stable Diffusion (fixed copy). Our idea can be concluded as the **control over control** or **meta-control** to let the task-aware hypernet learn the universe representation that is generalizable across different tasks. To implement it, we firstly map the task keys to instruction with a mapping function as: `{"hed": "hed edge to image", "canny": "canny edge to image", "seg": "segmentation map to image", "depth": "depth map to image", "normal": "normal surface map to image", "pose": "human pose skeleton to image", "hedsketch": "sketch to image", "bbox": "bounding box to image", "outpainting": "image outpainting"}`. Then, such instructions will be projected as text embeddings with the help of a language model (we adopt CLIPText in our implementation). The Task-aware HyperNet takes these task instruction embeddings, and projects them into different shapes to match the size of different zero-conv kernels, which will be modulated by these task embeddings accordingly. We would fix the parameters of task-aware hyperNet in the later stage of model training to ensure the stability of dynamics.

A.3 Data Collection

We have collected a large amount of training set (MultiGen-20M) including over 20M condition-image-prompt triplets across nine different tasks. We firstly download 3/4 of Laion-Aesthetics-V2 with score over six and filter out low-resolution (<512) images. As a result, 2.8M images are selected as source images. Then we apply the visual condition extractors as described in the main paper to collect `Canny`, `HED`, `Sketch`, `Depth`, `Normal Surface`, `Seg Map`, `Object Bounding Box`, `Human Skeleton` and `Outpainting`.

B Numerical Analysis of Task-Aware Modulated ControlNet

We show that our proposed task-aware modulated ControlNet preserves the properties of the original ControlNet structure. Specifically, we show **1)** The new task-aware modulated ControlNet preserves the zero-initialization property of ControlNet; **2)** The parameters of the task-aware modulated Controlnet can be updated once we start to train the model.

Denote the input feature map by $\boldsymbol{x}$, the frozen `SD Block` in Fig. 2 by $\mathcal{F}_{\mathrm{SD}}$, the extra condition by c, two zero convolution operators by $\mathcal{Z}^1_{\theta_1}(\cdot)$ and $\mathcal{Z}^2_{\theta_2}(\cdot)$, the trainable copy of `SD Block` by $\mathcal{G}^{\mathrm{SD}}_{\theta_{\mathrm{s}}}(\cdot)$, the task instruction by c_{task}, and the task-aware hyperNet by $\mathcal{H}_{\theta_{\mathcal{H}}}(\cdot)$. Then the output of the new task-aware modulated Controlnet can be expressed as

$$\boldsymbol{y}_c = \mathcal{F}_{\mathrm{SD}}(\boldsymbol{x}) + \mathcal{Z}^1_{\theta_1}(\mathcal{G}^{\mathrm{SD}}_{\theta_{\mathrm{s}}}(\boldsymbol{x} + \mathcal{Z}^2_{\theta_2}(c) \cdot \mathcal{H}_{\theta_{\mathcal{H}}}(c_{\mathrm{task}}))) \cdot \mathcal{H}_{\theta_{\mathcal{H}}}(c_{\mathrm{task}})\,. \tag{1}$$

Property of Zero Initialization. Similar to ControlNet [21], the weights and biases of the convolution layers are initialized as zeros. As a result, we have $\mathcal{Z}^1_{\theta_1}(\cdot) \equiv 0$ and $\boldsymbol{y}_c = \mathcal{F}_{\mathrm{SD}}(\boldsymbol{x})$, regardless of the initialization of $\mathcal{H}_{\theta_{\mathcal{H}}}(\cdot)$.

Gradient Analysis. We analyze the gradient of the modulated part

$$\nabla_\theta \left(Z^1_{\theta_1}(I) \cdot \mathcal{H}_{\theta_\mathcal{H}}(c_{\text{task}}) \right) = \mathcal{H}_{\theta_\mathcal{H}}(c_{\text{task}}) \cdot \nabla_\theta Z^1_{\theta_1}(I) + Z^1_{\theta_1}(I) \cdot \nabla_{\theta_\mathcal{H}} \mathcal{H}_{\theta_\mathcal{H}}(c_{\text{task}}), \tag{2}$$

where I is the input of the zero convolution layer.

When we start to train the network, the first part of the RHS of (2) follows similar analysis of ControlNet [21] since $\mathcal{H}_{\theta_\mathcal{H}}(c_{\text{task}})$ is constant when we analyze the gradient $\nabla_\theta Z^1_{\theta_1}(I)$. Since the parameters of $\mathcal{H}_{\theta_\mathcal{H}}(c_{\text{task}})$ are not initialized to zero, it is known that $\mathcal{H}_{\theta_\mathcal{H}}(c_{\text{task}}) \neq 0$. So the gradient dynamic follows the analysis of ControlNet. Therefore, we conclude that $Z^1_{\theta_1}(I) \neq 0$ after the first gradient update, and that the network can start to learn and update the following standard dynamics of stochastic gradient descent.

As for the second part of the RHS of (2), $Z^1_{\theta_1}(I) \equiv 0$ before the first gradient update, so the gradient is zero for $\theta_\mathcal{H}$. However, after the first gradient update of θ_1, we know $Z^1_{\theta_1}(I) \neq 0$, and $\theta_\mathcal{H}$ can be updated with non-zero gradients.

To conclude, the new task-aware Modulated ControlNet can still be efficiently updated and learned even if the convolution layers are initialized to zero.

C Zero-shot-task Results and Analysis

We show more zero-shot-task results in this section, where the tasks have not been trained on. In Fig. 11, we show zero-shot deblurring results guided by the keywords. Our deblurred images can successfully recover the fine-grained details of the images without training on such data. We note that some details are still missing, *e.g.,* the details in the painting in the first row are still not clear enough. In Fig. 12, we illustrate two zero-shot image colorization results. We believe that most parts of the generated images are acceptable, though the clothes of the second woman do not look the same to the input blurred image. In Fig. 13, we observe impressive zero-shot inpainting results. In the first row, the duck that is inputted in the text has been successfully generated in the inpainted image. The second row obtains acceptable results as well, though the faces do not look perfect. The overall zero-shot quality of UniControl is remarkable.

D Details of User Study

In the evaluation steps, we use Amazon Mechanical Turk (Mturk)[2] to perform user study. Specifically, we ask three Mturk master workers to select the best output result for each input condition. As shown in Fig. 8, we provide instructions on guidelines to select the best generated image. The annotators are provided the condition map and the text that describes the image, and are required to select the better output between the two generated images. Considering that images can both in good or bad qualities, we provide the tie option as well. We use the majority vote to determine the result of each image, which means that an image is considered as a better image if two or more annotators vote for it. We use 294 images for the tasks of Canny, HED, Surface Normal, Depth, Segmentation, User Sketch, and Outpainting. We adopt 100 images for the task of Human Skeleton and 187 images for the task of Bounding Box. In summary, we totally obtain 7,035 voting results for all nine tasks. 2/3 of source images in testing set are collected from MSCOCO with the remaining 1/3 from Laion. And it includes a very diverse range of topics including indoor scene, outdoor scene, oil painting, portrait, pencil sketch, animation, cartoon, etc.

E Failure Cases

We illustrate some failure cases in Fig. 10. In the first row, although our generated image successfully aligns the Bounding Box condition, the generated human has a distorted body. In the second row, our generated image looks similar to the ground truth; however, the human faces are blurred. We think that the reason is that UniControl inherits the data and model bias of Stable Diffusion, where the generated human commonly have issues. In the third row, the generated image does not look realistic. We believe that the training data can be improved both quantitatively and qualitatively.

[2] https://www.mturk.com

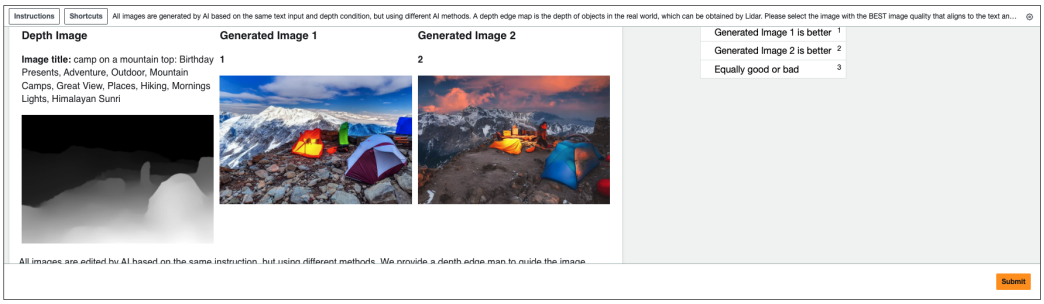


Figure 8: Mturk interface to select the better generated image.

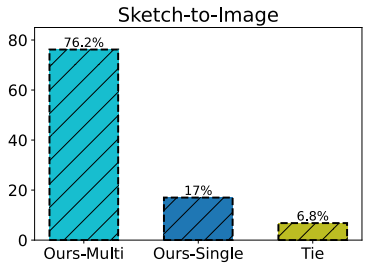


Figure 9: User study results of User Sketch to image generation.

F Additional Results

We illustrate more visualized results in this section on tasks Canny (Fig. 14), HED (Fig. 15), Depth (Fig. 16), Surface Normal (Fig. 17), Human Skeleton (Fig. 18), Bounding Box (Fig. 19), Segmentation (Fig. 20) and Outpainting (Fig. 21). These results further demonstrate the effectiveness of our proposed method. Moreover, due to the space limitation in the main paper, we report results of the last task, User Sketch. Given a sketched image, UniControl is able to achieve promising realistic images. The visualized results are in Fig. 22. The user study result can be found in Fig. 9, where it is observed that UniControl obtains significantly more votes than the single task model.

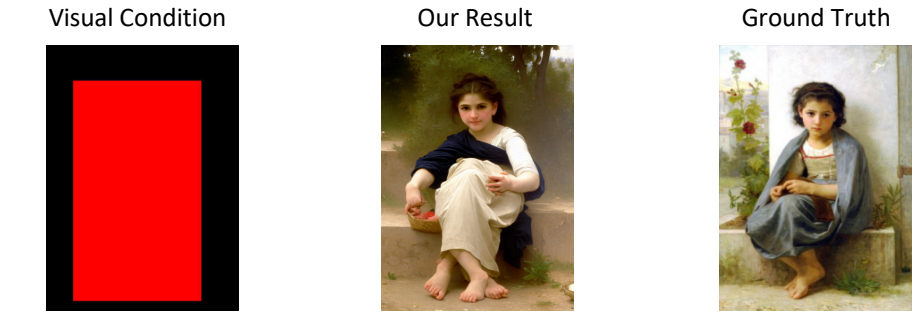
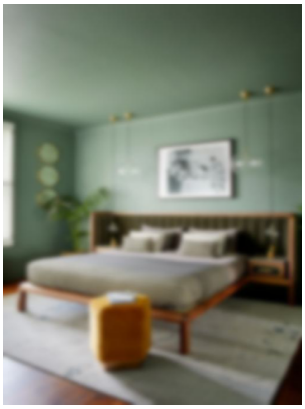
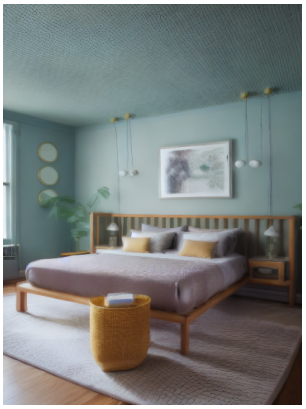


Figure 10: Failure Cases: distorted body (row one); blurred faces (row two); incorrect creation (row three).

Blurred Image

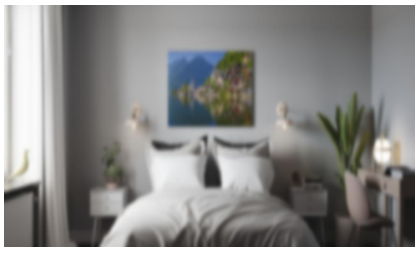
Our Result





“Bedroom Colour Ideas 25 Paint Colours With Impact Living”

“Christa McAuliffe (right, sat with her backup crew member Barbara Morgan) was a social studies teacher who had won NASA's Teacher in Space contest and earned herself a spot on the mission”



“Mountain Village in the Alps - Canvas print – Bedroom”

Figure 11: More zero-shot-task deblurring results.

Gray Image

Our Result

“Long White Casual Wedding Dress”

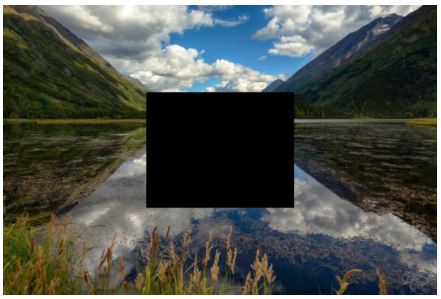
“Pixie Cropped Short Layered Synthetic Wig for Women-KAMI WIGS”

“Early morning view over the town of Tinerhir, south of the Todra Gorge, Morocco, North Africa, Africa”

Figure 12: More zero-shot-task gray-to-RGB colorization results.

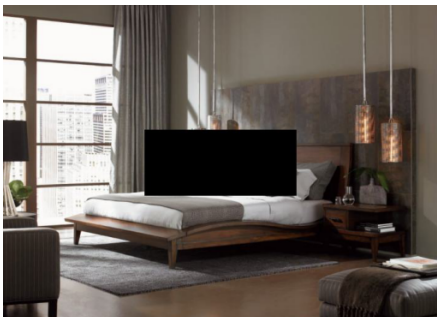
Cropped Image

Our Result



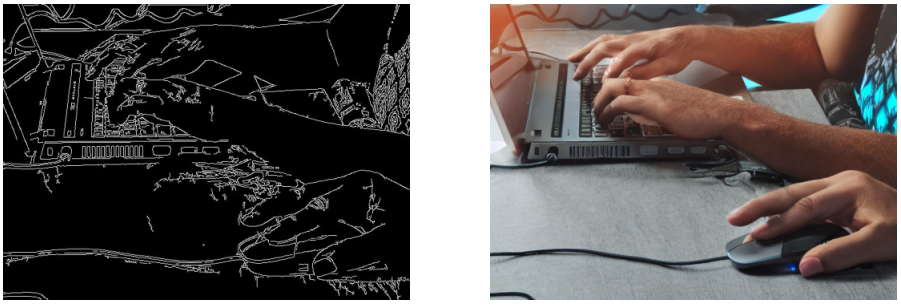
“A lone duck basks in the calm lake's mirror reflection of the Chugach mountain valley”

“Chancellor of the Exchequer Rishi Sunak was the most high-profile, and unexpected, appointment of the day”



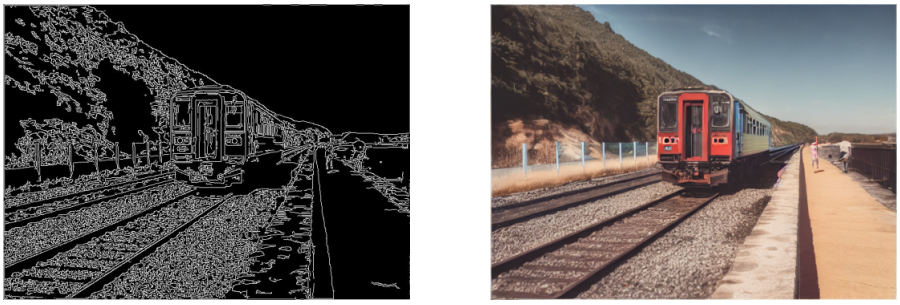
“Contemporary Bedroom Designs 2015 modern bedroom designs intended design “

Figure 13: More zero-shot-task image in-painting results. The in-painting MOE adapter weights are directly inherited from outpainting.



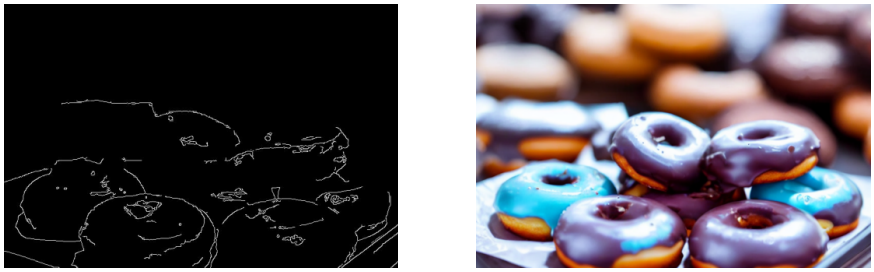
Input Image Our Method Output

(a) “Two arms typing on a laptop and one hand on a mouse”



Input Image Our Method Output

(b) “Two people walking along a side walk next to a train on the tracks.”



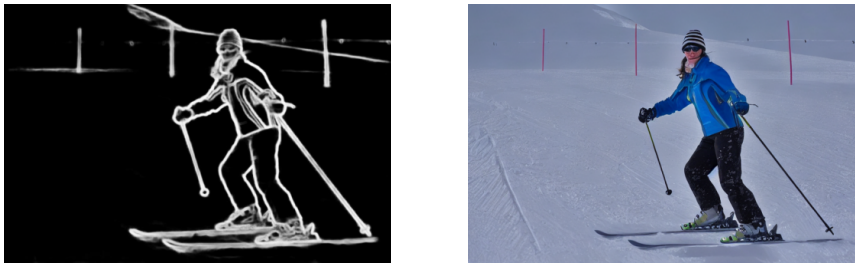
Input Image Our Method Output

(c) “A close up of glazed donuts that are plain or with chocolate.”

Input Image Our Method Output

(d) “A group of elephants with water in front and trees behind.”

Figure 14: Canny to Image Generation

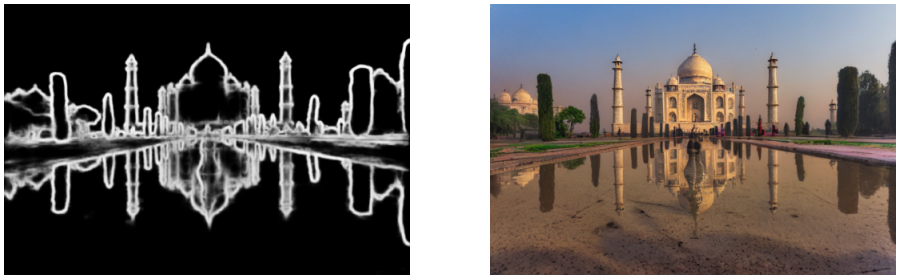


Input Image Our Method Output

(a) “A person on skis makes her way through the snow”

Input Image Our Method Output

(b) “Three zebras grazing in a grassy area near shrubs”



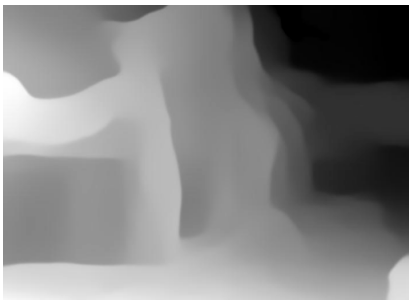
Input Image Our Method Output

(c) “The Taj Mahal mirrored by a water fountain’s reflection. - Agra, Uttar Pradesh, India - Daily Travel Photos”

Input Image Our Method Output

(d) “A young girl who is brushing her teeth with a toothbrush.”

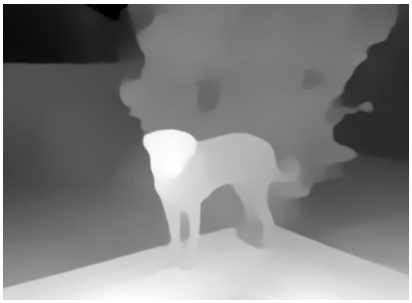
Figure 15: HED to Image Generation



Input Image

Our Method Output

(a) “The Ta Prohm Temple Located at Angkor in Cambodia by Kyle Hammons”



Input Image

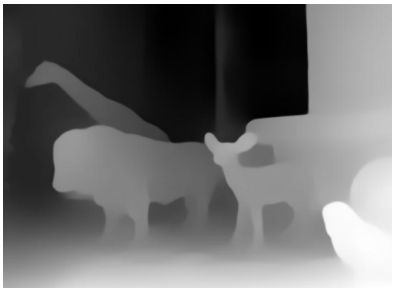
Our Method Output

(b) “A brown dog standing on a wooden bench near a lemon tree.”

Input Image

Our Method Output

(c) “A Polar Bear walks toward water, while a large bird lands on the opposite bank.”



Input Image

Our Method Output

(d) “A display of vintage animal toys on the floor.”

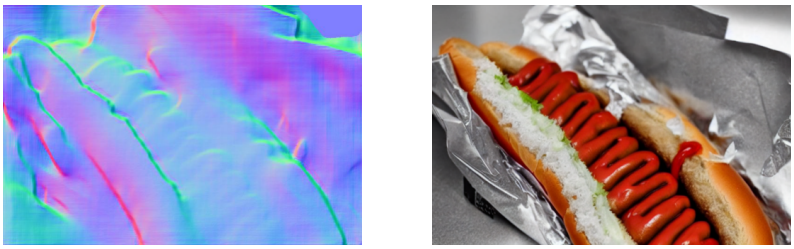
Figure 16: Depth to Image Generation

Input Image Our Method Output

(a) “A line of small teddy bears are in front of several DVD cases.”

Input Image Our Method Output

(b) “A man in the kitchen standing with his dog”



Input Image Our Method Output

(c) “A hot dog sitting on top of a bun in a wrapper”

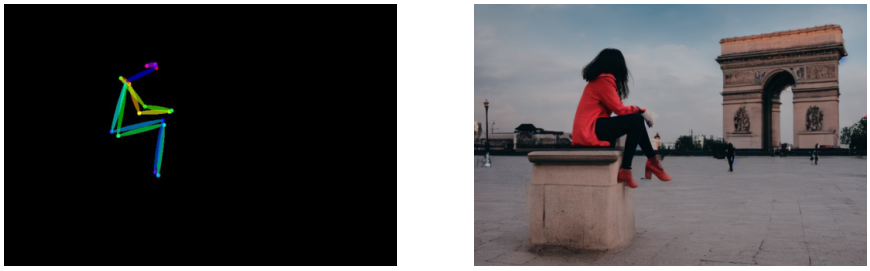
Input Image Our Method Output

(d) “Several people waiting on the side of train tracks as a train with it’s lights on comes down the track”

Figure 17: Surface Normal to Image Generation

Input Image Our Method Output

(a) “Photo of handsome man in black leather jacket”

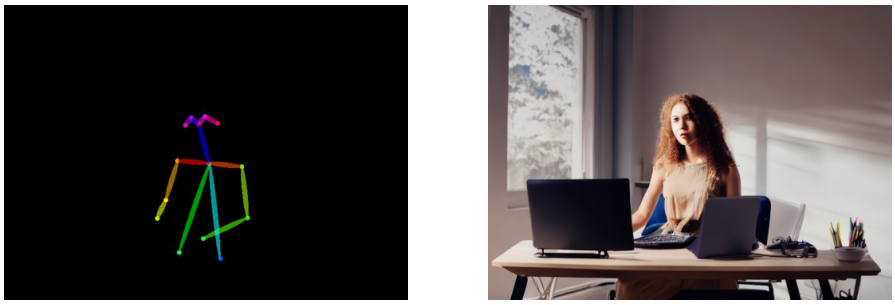


Input Image Our Method Output

(b) “A woman is sitting near a prominent landmark”

Input Image Our Method Output

(c) “A man that has ski’s and is standing in the snow.”



Input Image Our Method Output

(d) “A woman is sitting in front of a desk”

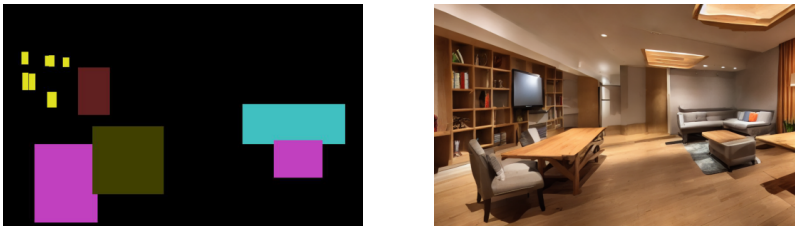
Figure 18: Human Pose Skeleton to Image Generation

Input Image Our Method Output

(a) “A woman is walking two dogs in the snow”

Input Image Our Method Output

(b) “Simone Righi frasi glasses linen suit menswear streetstyle icon fashion florence.”



Input Image Our Method Output

(c) “The large room has a wooden table with chairs and a couch.”

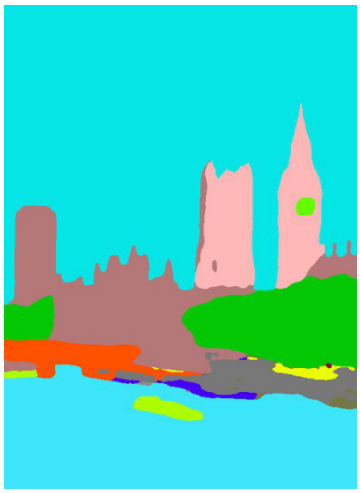
Input Image Our Method Output

(d) “Two black bags placed standing on the ground”

Figure 19: Bounding Box (by YOLO-V4-MSCOCO) to Image Generation

Input Image Our Method Output

(a) “A bench at the beach next to the sea”

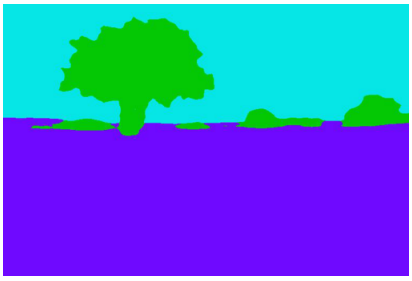


Input Image Our Method Output

(b) “Water traffic along the Thames by Big Ben”

Input Image Our Method Output

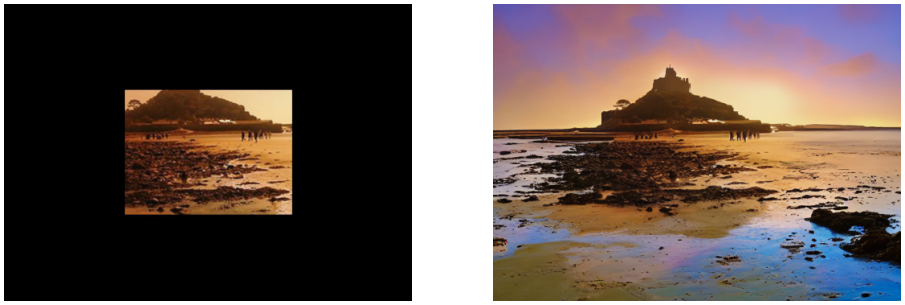
(c) “A well-lit and well-decorated living room shows a glimpse of a glass front door through the corridor.”



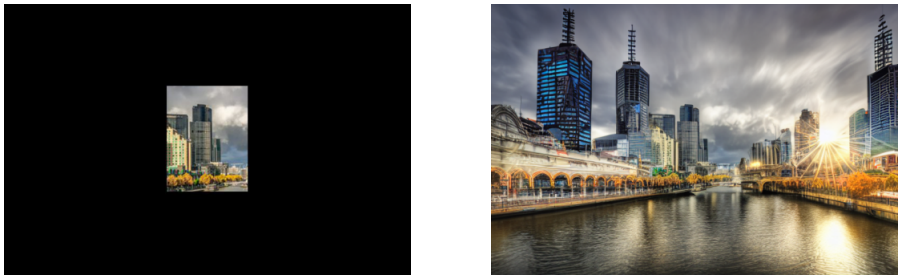
Input Image Our Method Output

(d) “Blue Hour Barley”

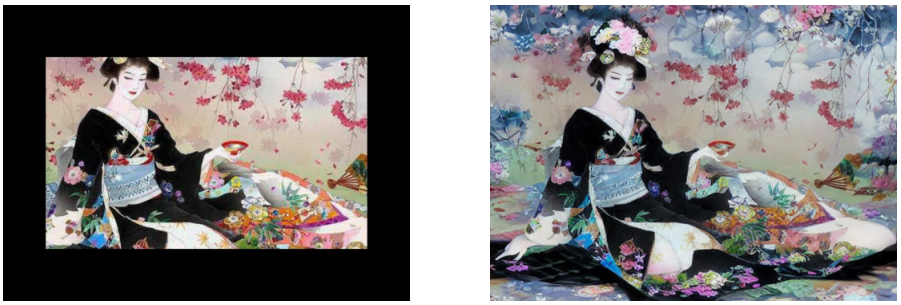
Figure 20: Segmentation Map (by Uniformer-ADE20K) to Image Generation



(a) “St, Michaels Mount, Cornwall”



(b) “Melbourne by teekay 72”



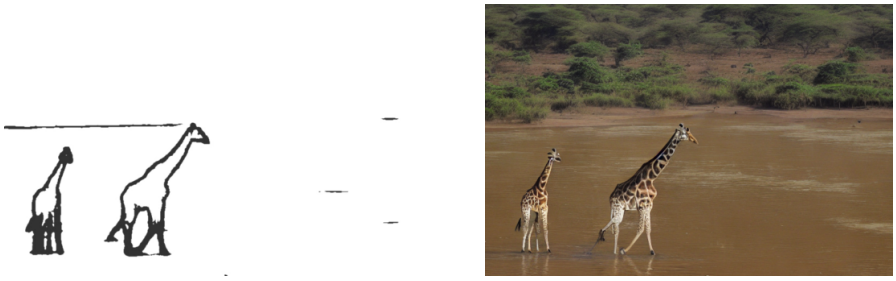
(c) “Lady in Black Kimono Paint by Diamonds”

(d) “Beautiful kitchen grand scale living pinterest for Kitchen cabinets lowes with old world metal wall art”

Figure 21: Image Outpainting

Input Image Our Method Output

(a) “A Limited Edition, Fine Art photograph of a beautiful sunrise at Lake Jackson in Sebring, Florida. Available as a Fine Art print”

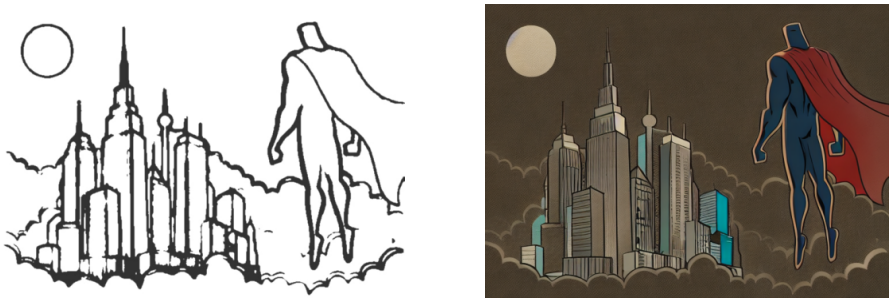


Input Image Our Method Output

(b) “Giraffes in Lake Manyara national park”

Input Image Our Method Output

(c) “Wild Ones 1991 Limited Edition Print - Frank McCarthy”



Input Image Our Method Output

(d) “Superhero watching over city. No transparency used. Basic (linear) gradients. A4 proportions.”

Figure 22: User Sketch to Image Generation