A. Method implementation details

MaskSketch is implemented in Jax [1] / Flax [21] similarly to the official implementation of MaskGIT. We will release the implementation of MaskSketch upon acceptance. We used an ImageNet-pretrained 256×256 VQGAN encoder-decoder and a 24-layer BERT transformer in all experiments.[2] In all experiments, we used the following parameters:

- layers $1, 3, 16, 18, 20, 21, 22$ for the structure distance objective in Eq. (1)
- Gumbel temperature 0 for ImageNet-Sketch and 0.001 for Pseudosketches experiments.
- 4 sampling trials for ImageNet-Sketch and 3 sampling trials for the Pseudosketches.
- 1000 iterations for ImageNet-Sketch and 500 iterations for Pseudosketches.
- Classifier-free guidance scales of $(0., 0.1, 0.25, 0.5)$ for ImageNet-Sketch and $(0., 0.05, 0.1)$ for Pseudosketches, varied for each iteration trial accordingly.
- λ_s is set to 0.9 for ImageNet-Sketch and to 0.95 for Pseudosketches.
- Starting mask rate is set to 0.95 for both datasets, and the end mask rate is 0.25 for ImageNet-Sketch and 0.33 for Pseudosketches.
- Token-Critic parameters: We used the Token-Critic refinement ratio $r_{tc} = 0.5$ and $r_{tc} = 0.6$ for ImageNet-Sketch and Pseudosketches experiments, respectively, and set the number of refinement steps to $N_{tc} = 32$ (explained in Appendix D).

B. Structure-guided sampling

Please see Fig. 10 for more examples of the structure-guided sampling across the first and last layers of MaskGIT (extending Fig. 4). Fig. 8 shows results on the examples from other modalities, such as cartoons and 3D models.

C. Results on ImageNet-Sketch and Pseudosketches

Unfortunately, we cannot include the illustration on ImageNet-Sketch and Pseudosketches in the main manuscript due to copyright concerns.

[2] The VQGAN and transformer model checkpoints used in our experiments are found in https://github.com/google-research/maskgit.

	3/3	2/3	Overall
ImageNet-Sketch 10-class			
MUNIT	10.70%	13.44%	13.80%
CUT	19.78%	24.90%	21.42%
VQ-I2I	0%	0.%	0.33%
CoGS	8.55%	16.60%	15.24%
MaskSketch (ours)	59.35%	40.71%	44.22%
Pesudosketches 10-class			
MUNIT	23.08%	22.88%	22.23%
CUT	25.00%	24.94%	23.57%
VQ-I2I	0.64%	2.75%	4.75%
CoGS	14.10%	16.02%	16.76%
MaskSketch (ours)	35.25%	31.35%	27.96%

Table 3. User preference study: ratios of unanimous votes (3/3), exactly two out of three votes (2/3) as well as the overall preference on the 10-class subsets of ImageNet-Sketch and Pseudosketches datasets.

	3/3	2/3	Overall
ImageNet-Sketch 10-class			
MUNIT	20	34	210
CUT	37	63	326
VQ-I2I	0	0	5
CoGS	16	42	232
MaskSketch (ours)	111	103	673
No selection	3	11	76
Pesudosketches 10-class			
MUNIT	36	100	528
CUT	39	109	560
VQ-I2I	1	12	113
CoGS	22	70	398
MaskSketch (ours)	55	137	664
No selection	3	9	112

Table 4. User preference study: number of the unanimous votes (3/3), exactly two out of three votes (2/3) as well as the overall number of votes on the 10-class subsets of ImageNet-Sketch and Pseudosketches datasets. The participants were asked to select the "No selection" option on the examples on which all methods performed comparatively poorly or the sketch content was unclear. We excluded the "No selection" examples from the statistics in Tab. 3 and Tab. 1.

C.1. Ablation of CLIP-based rejection

Tab. 5 shows that multi-trial CLIP-based rejection sampling achieves a significantly better trade-off between structure fidelity and realism than single-trial sampling.

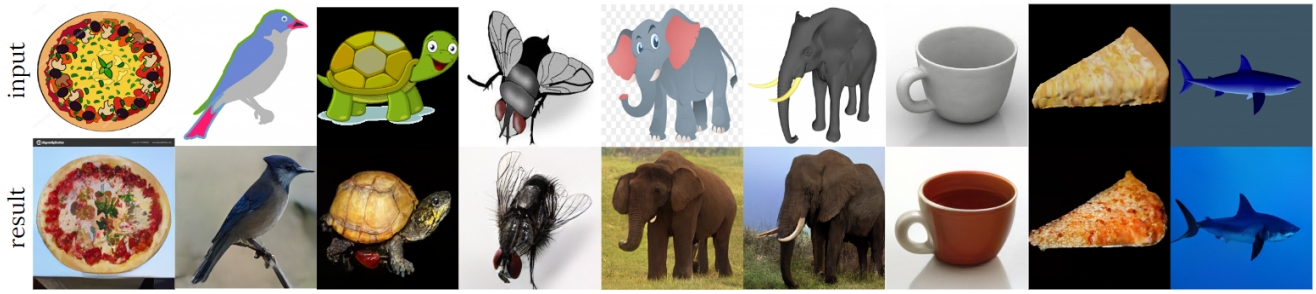

Figure 8. Results on cartoons (*left*) and 3D models (*right*).

	FID ↓	LPIPS ↑	CLIP pt. ↑	CLIP ft. ↓
ImageNet-Sketch 10-class				
No sel.	34.23	0.77	**71.87**	27.17
4-trial sel.	**33.24**	**0.78**	67.10	**26.63**
Pseudosketches 10-class				
No sel.	60.44	0.78	56.31	26.85
3-trial sel.	**56.55**	**0.78**	**59.48**	**25.60**

Table 5. CLIP-based rejection sampling ablation study. *No-sel.* indicates no rejection sampling was used. *4-trial sel. and 3-trial sel.* indicates selecting one sample out of 4 and 3 trials, respectively. *CLIP pt.* is the CLIP prompt similarity between the translation result and the prompt *"Photo of a* **c***"*, where **c** is the input class name *CLIP ft.* is the CLIP feature distance between the input sketch and the corresponding translation.

C.2. Ablation study on the parameter λ_s

Fig. 9 illustrates the effect of traversing the parameter λ_s on realism (CLIP prompt similarity) and structure fidelity (CLIP feature distance).

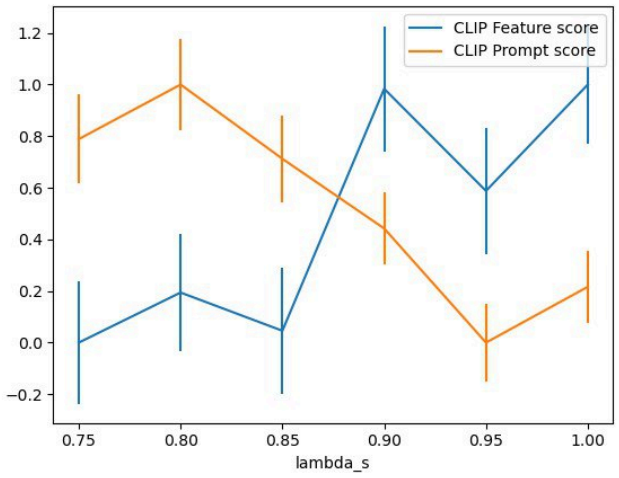

Figure 9. Ablation study result: CLIP feature score (structure fidelity) and CLIP prompt similarity score (realism) w.r.t. λ_s.

D. Token-Critic refinement

In our experiments, we used the ImageNet-trained Token-Critic [29] refinement to further improve realism of the translation results. In Token-Critic refinement, the tokens of a sampled image are passed to a critic transformer model that outputs a conditional likelihood score for each token. The score is high for tokens that are likely under the data distribution and low otherwise. We refine a sampled image by using the Token-Critic scores as the confidence scores in Algorithm 1, and setting $\lambda_s = 0$ (no structure guidance). The refinement process uses a mask rate of r_{tc}. We used $r_{tc} = 0.5$ and $r_{tc} = 0.6$ for ImageNet-Sketch and Pseudosketches experiments, respectively, and set the number of refinement steps to $N_{tc} = 32$, and in both experiments, the mask ratio varies across iterations according to the cosine schedule.

E. User preference study

For all the validation images in the ImageNet 10-classes and Pseudosketches 10-classes datasets, we asked the participants to pick one option that best answers the question: *"Given the task of converting the sketch shown on the left into a realistic photo, which result do you prefer?"*. For each example, we got the answers from three participants, and we report the unanimous voting results (3/3) in Tab. 1. We report the ratios of choices of the user preference study in Tab. 3: statistics for the unanimous votes (3/3), exactly two out of three votes (2/3) as well as the overall preference. We also report the total number of choices in Tab. 4.

F. CLIP-based metrics

Structure distance To estimate structure similarity between the input sketch $\boldsymbol{x}$ and the translation result $\boldsymbol{y}$, we compute L_1-distance between the ResNet101-based CLIP image encoder intermediate layer features: $\text{CLIP}^s(\boldsymbol{x}, \boldsymbol{y}) = ||\,\text{CLIP}_l(\boldsymbol{x}) - \text{CLIP}_l(\boldsymbol{y})||_1$, where l is the ResNet-101 layer block index. In our experiments, we use the last layer block ($l = 4$).

Figure 10. Structure-guided sampling examples using layers $\mathcal{L} = \{1, 2, 3\}$ (top of each row) and layers $\mathcal{L} = \{16, 18, 20\}$ (bottom of each row).

Prompt similarity To asses realism and semantic accuracy of the translation result, we use CLIP zero-shot classification to estimate the relative similarity between the translated image and the prompt "*Photo of a* $\boldsymbol{c}$", where c is the ground truth class label index corresponding to the input sketch. Therefore, given an input sketch $\boldsymbol{x}$ of class c, the prompt similarity is computed as:

$$\text{CLIP}^r(c, \boldsymbol{y}) = \text{softmax}\{CLIP(\boldsymbol{y})^T CLIP(\boldsymbol{p})\}[c]$$

where $\boldsymbol{p} =$ ["Photo of a **m**" $\forall \boldsymbol{m} \in \Omega$], Ω is the set of class labels in the dataset.

G. Comparison with PITI

In this section, we provide the quantitative and qualitative comparison with the concurrent *supervised* image-to-image translation method PITI [51]. For a fair comparison, we compared the generation results on the four classes from the intersection of classes of the MS COCO [30] dataset that was used to train PITI and ImageNet-Sketch 10 classes we used to compare with the other baseline methods. Since PITI is sensitive to the modality of the input (e.g., it produces subpar results on inverted sketches), we used PITI 's edge extraction pipeline on the input sketches before translating with PITI. The CLIP-based evaluation results on Tab. 6 show that PITI results are slightly better in terms of structure fidelity, however they are generally less realistic than MaskSketch translation results. An important disadvantage of PITI is its sensitivity to the domain shift: the edge extraction method HED [53] that was used to train PITI removes some edges in the given sketch, which results in errors in structure and even misclassification of the input sketch (e.g. PITI typically confuses the round pizza shape with other round objects, such as watch or bowl).

	CLIP ft.	CLIP pt.
PITI	**25.0**	59.1
MaskSketch (ours)	27.3	**68.2**

Table 6. CLIP-based evaluation (Sec. 3.4) on 4 classes from the intersection of classes in MS-COCO [30] and ImageNet-Sketch 10-class datasets: zebra, pizza, songbird, door.