We thank reviewers for their insightful comments and are happy that they find the problem significant (**R1**) and challenging (**R2**), the proposed decomposition novel (**R2**) and the results compelling (**R1**). In the following, we address concerns and provide new experimental evidence. Fig. 1-8 and Tab. 1-2 are in the main paper and Fig. 9-15 can be found in the appendix. We kindly ask the reviewers and the AC zoom into the figures.

R1: Suggested papers and Video. Thank you for pointers to additional papers; we will include a discussion. Note that a video *is* available in the *supplementary materials*.

R2: Stochasticity of task. Our task shares similarities to NLP problems such as text auto-completion in gmail. In text prediction, localization hints are provided by positional encoding, and the "starting position" is the last token; the attention model in transformers allows the model to determine the relevant *local* context to predict the next token. In drawings, on the other hand, the starting position is *not* fixed and an important degree of freedom. Hence the attention model in CoSE-$\mathcal{R}_\theta$ allows the prediction to focus on a *local* context by conditioning on the starting position. This allows our model to perform effectively. To show the importance of the initial stroke positions, we trained a model without conditioning on them and see the CD nearly double from 0.0442 (Tab. 2) to 0.0790 (new). Fig. 16 also shows that conditioning on the start position helps to attend to the nearby strokes, which is increasingly important as the number of strokes gets larger.

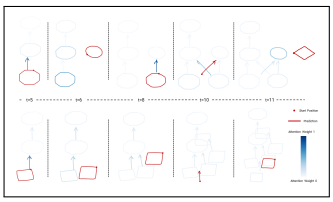

Figure 16: Average attention visualization over time with (top) and without (bottom) conditioning on the start position. Please enlarge Fig.

R2: Relational model ablation. Note that predicting starting positions alone is not enough. A crucial component in capturing pairwise dependencies is the proposed relational model CoSE-$\mathcal{R}_\theta$. Performance degrades substantially if we replace CoSE-$\mathcal{R}_\theta$ with an LSTM, receiving stroke embeddings in drawing order (Tab. 4). Sketch-RNN models the data as a sequence of points in contrast to our compositional approach.

Table 4: Ablation on $\mathcal{R}_\theta$

$\mathcal{E}_\theta/\mathcal{D}_\theta$	$\mathcal{R}_\theta$	CD↓
CoSE-$\mathcal{E}_\theta/\mathcal{D}_\theta$	CoSE-$\mathcal{R}_\theta$	**0.0442**
CoSE-$\mathcal{E}_\theta/\mathcal{D}_\theta$	RNN	0.0713
Sketch-RNN		0.0679

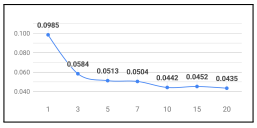

Figure 17: Pred. CD vs. # GMM.

R2: Diversity of the predictions. Given an initial position, the GMM contains a diverse set of predictions (Fig. 4). In Fig. 17, we ablate wrt the number of components as requested. The ability of our model to generate similar diversity to the test set is also visible in Sec. 9: mode collapse would incur a visible difference in the distribution of ground-truth (blue) vs. predicted (yellow) embeddings (cf, Fig.11-left). We quantify this effect by calculating the Earth-Mover distance (EMD) between the two embedding distributions. Fig.11, left-to-right: EMDs of 1797, 251 and 155 (ours). The EMD decreases as the GT and predicted distributions become more similar.

R2: Stroke discontinuity. Note that this is emergent behavior from the dataset which contains many such examples.

R2: Experimental design. The results summarized in Fig. 16 & Tab. 4 show that modeling of pairwise dependencies and predicting the next embedding are crucial. Our experiments assess different models under that assumption and we focus on the task of predicting the next stroke giving a partial drawing. To control high variability in the predictions across different generative models, we feed ground-truth starting positions in our quantitative analysis (note that the qualitative results rely only on the *predicted* starting positions). We furthermore use a stochastic metric (Eq. 5) to ensure fairness. Moreover, our final metric, the chamfer distance (CD) of the strokes, allows us to compare models trained with different objectives (e.g., next point prediction as in SketchRNN) and different representations (e.g., velocity).

R3: Gradients. We aim to decouple the local stroke from the global drawing structure. We train via the *reconstruction* loss only, and do not back-propagate the relational model's gradients. Doing so would force the encoder to use some capacity to capture global semantics. Training our best model with all gradients flowing to the encoder, the error (Recon. CD) increases from 0.0136 to 0.0162 and the prediction error (Pred. CD) from 0.0442 to 0.0470.

R3: Embedding size. We compare CoSE-$\mathcal{E}_\theta/\mathcal{D}_\theta$ and the baseline seq2seq with varying embedding size; see Tab. 5. We use CoSE-$\mathcal{R}_\theta$ to evaluate the predictive power of the corresponding embeddings. For both models, the reconstruction performance improves with increasing embedding size. However, it also results in a less compact representation space, making the prediction task *more challenging*.

Table 5: Ablation on embedding size D

$\mathcal{E}_\theta/\mathcal{D}_\theta$	D	Recon. CD ↓	Pred. CD↓	SC ↑
CoSE-$\mathcal{E}_\theta/\mathcal{D}_\theta$	8	0.0136	**0.0442**	**0.361**
CoSE-$\mathcal{E}_\theta/\mathcal{D}_\theta$	16	0.0091	0.0481	0.335
CoSE-$\mathcal{E}_\theta/\mathcal{D}_\theta$	32	0.0081	0.0511	0.314
seq2seq	8	0.0138	0.0540	0.276
seq2seq	16	0.0076	0.0783	0.253
seq2seq	32	**0.0047**	0.0848	0.261

R4: Novelty. We respectfully disagree with **R4** on the limited novelty. We don't simply replace RNNs with transformers but propose a novel task decomposition that we show to be important and propose a novel architecture to capture stroke dependencies in an unordered fashion. Further, we quote from the *official* reviewing guidelines that "`excuse authors for not knowing all non-refereed work (e.g, ArXiv)`". Both references were recently published (2/3 months) on ArXiv at submission time (see below for differences).

R4: Baselines. Sketchformer learns sketch representations for image retrieval (SBIR) using full supervision whereas our task is fully unsupervised. The suggested $mAP\%$ metric requires *labels* for evaluation. We emphasize that our goal is to learn the compositions of strokes into drawings, rather than the entire sketch, to allow for scalability wrt to sketch complexity.

Our approach can generalize to different domains, we provide qualitative results on QuickDraw sketch (Fig. 5) and IamOnDB handwriting datasets (Fig. 18)

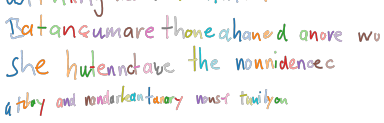

Figure 18: Given the **first** and **second** strokes, unconditional handwriting samples generated by our model.