# Disentanglement and Compositionality of Letter Identity and Letter Position in Variational Auto-Encoder Vision Models

**Bruno Bianchi[1,@], Aakash Agrawal[2], Stanislas Dehaene[2,3,4], Emmanuel Chemla[4,5], Yair Lakretz[4,5,@]**

**1.** CONICET-UBA, FCEN, ICC-Departamento de Computación. Buenos Aires, Argentina;
**2.** Cognitive Neuroimaging Unit, CEA, INSERM, Université Paris-Saclay, NeuroSpin, Gif/Yvette, France;
**3.** Collège de France, Paris, France; **4.** PSL University, Paris, France;
**5.** LSCP, Ecole Normale Supérieure, CNRS, Paris, France,
bbianchi@dc.uba.ar; yair.lakretz@gmail.com

**Abstract**

Human readers can accurately count how many letters are in a word (e.g., 7 in "buffalo"), remove a letter from a given position (e.g., "bufflo") or add a new one. The human brain of readers must have therefore learned to disentangle information related to the position of a letter and its identity. Such disentanglement is necessary for the compositional, unbounded ability of humans to create and parse new strings, with any combination of letters appearing in any positions. Do modern deep neural models also possess this crucial compositional ability? Here, we tested whether neural models that achieve state-of-the-art on disentanglement of features in visual input can also disentangle letter position and letter identity when trained on images of written words. Specifically, we trained beta variational autoencoder ($\beta$-VAE) to reconstruct images of letter strings and evaluated their disentanglement performance using CompOrth - a new benchmark that we created for studying compositional learning and zero-shot generalization in visual models for orthography. The benchmark suggests a set of tests, of increasing complexity, to evaluate the degree of disentanglement between orthographic features of written words in deep neural models. Using CompOrth, we conducted a set of experiments to analyze the generalization ability of these models, in particular, to unseen word length and to unseen combinations of letter identities and letter positions. We found that while models effectively disentangle surface features, such as horizontal and vertical 'retinal' locations of words within an image, they dramatically fail to disentangle letter position and letter identity and lack any notion of word length. Together, this study demonstrates the shortcomings of state-of-the-art $\beta$-VAE models compared to humans and proposes a new challenge and a corresponding benchmark to evaluate neural models.

**Keywords:** Reading, Deep Learning, Compositionality, Auto-Encoders.

## 1. Introduction

Reading is an invention of human modern culture. Unlike other domains of visual processing such as faces, reading skills are not innate and require extensive practice (Dehaene et al., 2015). Reading acquisition therefore must rely on the development of a new neural mechanism in the human brain. Given that letters are the building blocks of words, to process new words, the neural mechanism needs to identify single letters in the input, their positions in a word, and compose this information to process entire words. While brain imaging has localized where in the brain visual processing of words occurs (Cohen et al., 2002), what is the precise neural mechanism that enables us to recognize words is largely unknown.

Recent advances in deep neural models have drastically improved the accuracy on Optical Character Recognition (OCR) tasks (Li et al., 2023). Deep neural models can now achieve similar-to-human performance on a variety of tasks, including vision and language. Although neural models are often considered black boxes, full access to their neural computations during processing is possible. Analyzing the properties of these networks provides new opportunities to study neural mechanisms underlying orthographic processing, and in particular, into how letter-identity and letter-position information are extracted from raw images and then composed together to encode whole words.

To study this question in neural models, we developed *CompOrth* - a battery of tests, which evaluate compositionality in models and their generalization performance. CompOrth provides several tests, which can be used to both evaluate the 'behavioral' performance of the models, as well as study *neural* mechanisms in the model. The tests are designed in a way that directly probes the question of whether a neural model extracts and disentangles letter-identity and letter-position information from raw images, and whether it can compose them together to encode entire words.

The main hypothesis of this work is that vision models can achieve such *functional* disentanglement of letter identity and position, required to succeed on CompOrth, by *neurally* disentangling this information, representing it in separate units of the model. This is since neural disentanglement could facilitate composition of letter identity and letter po-



*Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026*, pages 293–308
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

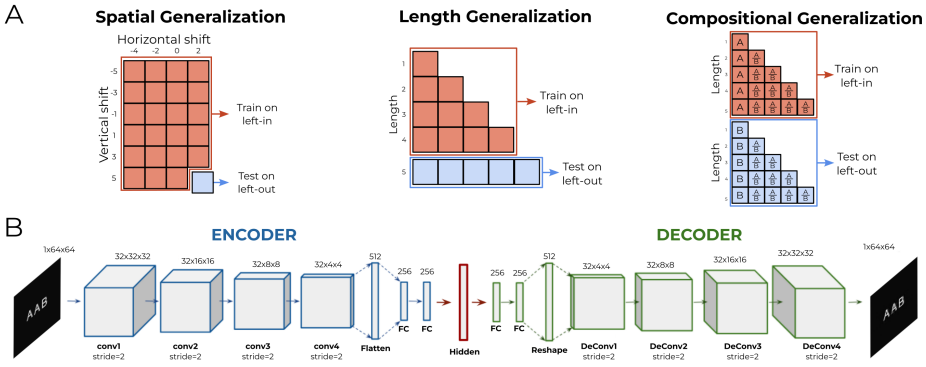


Figure 1: **(A)** The CompOrth Benchmark: schemes of the three types of generalization tests. Each test consists of several splits into training and test sets, ensuring that in each split, the test set contains images generated from different combinations of factors than those in the training set. **(B)** Model architecture: an illustration of the architecture of an auto-encoder for processing images of written words. The size of the hidden layer (bottleneck) was varied as part of the experimentation.

sition in downstream computations, thus improving generalizations to unseen combinations of letters and positions. Beta Variational Auto-Encoders ($\beta$-VAEs) are a leading model for neural disentanglement (Higgins et al., 2017), and, interestingly, they have been shown to also align with neural activity recorded from the primate cortex. Specifically, Higgins et al. (2021) showed that disentangled representations in neural models enhance prediction of Macaque neural activity, aligning artificial and biological processing. This suggested that the face region in the Macaque's brain disentangles such factors of variations. Here, we thus study $\beta$-VAEs using CompOrth, testing whether strong neural disentanglement in the models could lead to improved performance on CompOrth, given its requirement for both disentanglement and compositional abilities as of humans.

Our results show that $\beta$-VAEs learn to disentangle surface properties of written words, such as 'retinal' horizontal and vertical position. However, $\beta$-VAEs dramatically fail to disentangle identity and positional information of letters, and to combine them together. We show that such disentanglement of letter identity and position is lacking both at the behavioral and neural levels of the model. Furthermore, we show that $\beta$-VAEs lack a robust notion of word length, failing to generalize to unseen word lengths in CompOrth. Finally, we provide arguments for why other types of neural models might suffer from the same limitation, and therefore suggest CompOrth as a challenging benchmark for future models.

## 2. Related Literature

Literature on cognitive sciences contains competing theories about how words are neurally encoded in the human brain. Based on experiments in humans, early studies theorized the presence of letter combination detectors (Dehaene et al., 2005), i.e., letters combine to form bigrams, which are further combined to form larger n-grams and finally words. Other theories suggested neural encoding based on open bigrams - neurons that are tuned for higher-order combinations of letters, which are not necessarily adjacent (Grainger and Whitney, 2004). Compositional theories suggest a neural encoding scheme, whereby letter identities and letter positions are directly combined to form words (Agrawal et al., 2020), in line with early suggestions in the literature (Davis and Bowers, 2004). More recently, the lexical categorization model (LCM) was suggested, positing that linguistic processing is optimized by enabling swift retrieval of meaning for known words while effectively discarding meaningless letter combinations (Gagl et al., 2022). Predictions from such theories about the computational operations underlying orthographic processing were often contrasted using neural data recorded from brain data (e.g., Taylor et al., 2013).

Modern neural models now provide new opportunities to study how words can be neurally encoded, possibly informing the above debate. One of the computational tasks in the field, most closely related to reading, is OCR. Various neural models were suggested for OCR, including Recurrent Neural Networks (Breuel et al., 2013), Convolutional Neural Networks (CNNs) (Zhang et al., 2017), Transformers (Li et al., 2023; Azadbakht et al., 2022), and hybrid architectures (Naseer and Zafar, 2019; Jain et al., 2017). Some studies have also tried to identify neural mechanisms in models trained on OCR, or similar tasks. For example, Hannagan et al. (2021) studied neural activations of units of CNNs that were trained to recognize

images of objects and words. They found that a compositional encoding scheme emerged also in the models. During training, the models developed special units that encode either letter identity or ordinal letter position in a word (but not n-grams). This evidence from neural models thus provides support in favor of a compositional code based on single letters and their positions (Agrawal et al., 2020).

OCR focuses on developing models capable of extracting text from images, a task complicated by various sources of noise. This noise can stem from inherent variations in handwriting styles and scripts, as explored in the survey by Baldominos et al. (2019), which examines handwritten character recognition using the MNIST and EMNIST datasets (Cohen et al., 2017). Another significant source of noise is the degradation of physical documents due to age or environmental factors, as discussed by Fontanella et al. (2020), who investigate pattern recognition and AI techniques for cultural heritage preservation, including dealing with deteriorated materials. These challenges have driven the development of numerous benchmarks and approaches in scene text recognition, aiming to address specific issues. Shi et al. (2016) pioneered end-to-end trainable neural networks for image-based sequence recognition, a foundational work in the field. More recently, Lyu et al. (2022) introduced MaskOCR, leveraging masked encoder-decoder pretraining for improved text recognition. Du et al. (2022) proposed SVTR, a scene text recognition model based on a single visual model, simplifying the architecture. Wang et al. (2021) presented a novel scene text recognizer with a visual language modeling network, bridging the gap between visual and linguistic understanding.

The complexities of scene text recognition extend beyond just noise. Researchers have tackled the challenge of capturing global semantic context, as demonstrated by Yu et al. (2020) with their semantic reasoning networks, Wan et al. (2020) with TextScanner's ordered character reading, Cui et al. (2021) with their representation-enhanced encoder-decoder framework, and Bhunia et al. (2021b) who explore joint visual semantic reasoning with a multi-stage decoder. Handling text in diverse orientations is another key area, addressed by Zhang et al. (2020) with AutoSTR's efficient backbone search and Yan et al. (2021) who utilize primitive representation learning. The interplay between visual processing and language models is crucial, as investigated by Fang et al. (2021) who propose a human-like bidirectional and iterative language model, and Bautista and Atienza (2022) who explore scene text recognition with permuted autoregressive sequence models. Furthermore, researchers have focused on specific challenges like degraded image quality, as in Mou et al. (2020) PlugNet which uses a pluggable super-resolution unit, the misrecognition of contextless text images studied by Yue et al. (2020) with their RobustScanner that enhances positional clues, and unseen character sequences addressed by Bhunia et al. (2021a) through iterative text recognition by distilling from errors. Finally, the issue of limited labeled data has been tackled by Baek et al. (2021) who explore scene text recognition with fewer labels by leveraging real datasets.

However, in contrast to these other challenges in OCR, the primary objective of CompOrth is to identify models that achieve *compositionality*. That is, models that can recognize words with new combinations of letter identities and letter positions, unseen during training, similarly to how humans process new words (Barton et al., 2014). To achieve compositionality, a model would need to learn to identify single letters and then efficiently compose them with their positions, recursively, when encoding entire words. To focus on compositionality, CompOrth thus simplifies much of the problem by eliminating several sources of noise present in OCR corpora, as described below.

## 3. General Setup

### 3.1. The CompOrth Benchmark

#### 3.1.1. Stimuli:

The stimuli for CompOrth were designed to probe the encoding of single letters and their positions, minimizing other information. Strings ('words') in each test contain only two uppercase letters (e.g., 'A' and 'B') in Arial font, size 12. We generated all 62 possible words of 1 to 5 letters (e.g., 'A', 'B', 'AA', ..., 'BBBBB'). Images were created by varying word location ('retinal' location) and character spacing, resulting in 11,904 images. All featured white letters on a black background (Figure A.1).

Retinal location was adjusted by shifting the string vertically and horizontally; zero displacement centers the string, while a displacement of 1 moves it one pixel up and right. Strings were shifted up to 4 pixels from the center in all directions (Figure A.1). Spacing variation decoupled a letter's retinal (absolute) location from its (relative) word position by adjusting inter-character spacing from -2 to +1 pixels, avoiding overlap with 0 as the default.

#### 3.1.2. Three Generalization Tests:

CompOrth contains three generalization tests: (1) *Spatial Generalization* (Figure 1A-Left), across 'retinal' positions; (2) *Length Generalization* (middle), to unseen word lengths; and (3) *Compositional Generalization* (right), to unseen letter identity and position combinations. The latter evaluates if mod-

els develop abstract notions of position and identity. Each test used multiple training-test splits where the test set included factor combinations excluded from training. Specifically, one level of a generative factor (e.g., a specific retinal position) was reserved for testing (Figure 1A-Left, blue square).

**Spatial Generalization:** For each possible combination of x-y-shift, all words with that combination are used as the left-out while the rest of the words are used as the left-in set (Figure 1A, left).

**Length Generalization:** For each length, all words with that length are used as the left-out set, while the rest of the words are used as the left-in set (Figure 1A, middle).

**Compositional Generalization:** For each relative position of each letter (e.g., 'A' in $2^{nd}$ position), all words with that combination (e.g., "AA", "BA", "AAA", "BAB", etc.) are used as the left-out set, while the rest of the words are used as the left-in set (Figure 1A, right).

### 3.2. Model Architecture

We benchmarked CompOrth using Variational Auto-Encoders (VAEs; Kingma and Welling, 2013), including $\beta$-VAEs (Higgins et al., 2017). Following Higgins et al. (2017), models used 4 convolutional and 2 fully connected layers for the encoder, with a mirrored decoder architecture (Figure 1B). Auto-encoders can be tested on unseen words, whereas standard feed-forward classifiers have finite output units (Hannagan et al., 2021), making evaluation on unseen words difficult without retraining. Moreover, $\beta$-VAEs can be optimized for neural disentanglement, encouraging individual latent units to encode generative factors like letter identity and position. For training, we used a batch size of 64 samples. Training and Evaluation were conducted with Nvidia Quadro RTX 8000 48 GB GPUs. The whole experiment, including grid search, took about 72 hours. For the grid search, we optimized for the following hyper-parameters: Beta ($2^i, i \in \{0, ..., 7\}$), size of the latent layer ($2^i, i \in \{3, ..., 7\}$) and initial learning rate ($10^i, i \in \{-4, ..., -2\}$). The optimal learning rate across all combinations in the grid search was consistently $10^{-4}$. We set the maximal number of epochs to 1000, which we verified to be large enough to reach full convergence in all cases. We also verified that the hyper-parameters that led to the best performance lie within the range of explored values, and not at its edges (Figure A.1).

### 3.3. Model Evaluation

**Reconstruction Loss:** For model selection (Figure 3), we used the standard reconstruction loss for visual AEs, which is calculated based on the pixel-by-pixel difference between the original and reconstructed images Suganuma et al. (2018).

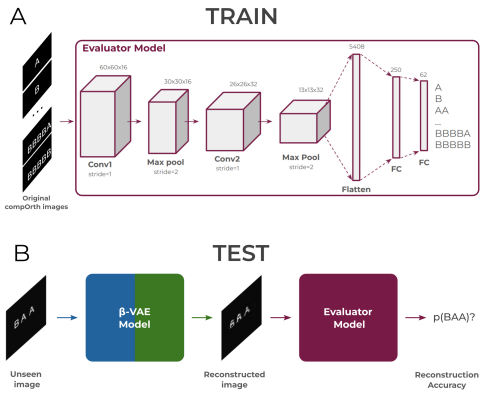


Figure 2: **Evaluator Model: (A)** Pixel-by-pixel loss poorly quantifies reconstruction quality; blurry images can yield low loss despite unrecognizable letters. To ensure a reliable measure of word recognizability, we trained an 'Evaluator Model'—a CNN-based classifier—on all CompOrth words. Its output probability serves as a proxy for the recognizability of $\beta$-VAE reconstructions. **(B)** Reconstruction Accuracy: This metric is defined as the classification accuracy of the Evaluator model on all reconstructed images.

**Reconstruction Accuracy:** However, to test compositionality, relying solely on reconstruction loss might not be sufficient. This is since a large number of pixels in the images could simply be black, leading to a low reconstruction loss even when the reconstructed word is poorly recognizable. Furthermore, blurry or unidentifiable letters in reconstructed images can reduce reconstruction loss without genuinely improving word recognizability. (see examples in Figure 3B). To address this, we defined a new evaluation metric, Reconstruction Accuracy, which quantifies word *recognizability*.

To compute Reconstruction Accuracy, we trained an *Evaluator*—a feed-forward CNN classifier—on all CompOrth words (Figure 2A). It was trained on original images, achieving perfect accuracy via 5-fold cross-validation. The Evaluator comprises two convolutional and two linear layers, with a softmax output across 62 words. This methodology is akin to a GAN 'discriminator' (Goodfellow et al., 2014), but used only for testing.

Reconstruction Accuracy was then defined as the probability assigned by the Evaluator to the target image (Figure 2B). For example, given an image from the test set, such as "ABABA", the image is presented first to the beta-VAE model. The output from the beta-VAE, the reconstructed image, is then given as input to the Evaluator. The Reconstruction Accuracy for this image is therefore the probability

that the Evaluator assigns, at its output layer, to the word "ABABA". In sum, the Evaluator produces a value in the range [0, 1] for how well the word image was reconstructed by the beta-VAE model. How recognizable it is.

**Metrics for Disentanglement (MIG and MIR):** To quantify neural disentanglement, Chen et al. (2018) used the Mutual Information Gap (MIG), whereas Whittington et al. (2022) suggested the Mutual Information Ratio (MIR). MIG scores high if each factor is encoded in a single neuron. In contrast, MIR scores high when each neuron responds to a single factor. MIR thus allows high scores when multiple neurons respond to the same unique factor. Since compositionality can be achieved even if several neurons extract the same information, we used MIR to test if neural disentanglement facilitates behavioral disentanglement of letter-identity and position in CompOrth.

# 4. Results

## 4.1. Model selection

We first optimized for the hyperparameters of the $\beta$-VAE models, using nested cross-validation. Figure 3A illustrates model selection, by showing both the reconstruction loss and the MIR for all models in the grid-search. Since both good reconstruction and disentanglement are desired properties of a model, the optimal models lie on the Pareto front (purple circles) of the problem (Goodarzi et al., 2014). No other models outperform them in both criteria simultaneously. In what follows, we therefore report results based on average performance across all optimal models on the Pareto front. We later analyze particular cases from these models.

To illustrate the reconstruction ability of the models, Figure 3B&C shows examples from two models - one from the Pareto front, having good reconstruction performance (with $\beta = 4$, *latent-layer size* $= 32$, *learning rate* $= 10^{-3}$) and the other with a low one ($\beta = 4$, *latent-layer size* $= 128$, *learning rate* $= 10^{-5}$). In both panels, the top row shows examples from the original images, and the bottom row shows their corresponding reconstructions. In the case of low reconstruction performance, the reconstructed images are blurry, with only 'retinal' location preserved.

## 4.2. Behavioral Evaluation using CompOrth

**Spatial Generalization – $\beta$-VAEs can generalize to unseen 'retinal' locations:** Figure 4A-Left shows the mean generalization performance to unseen retinal positions, for all models from the Pareto front (mean performance in black). On average, the models show good ability to reconstruct words in positions where they were not seen during training. Except for three models on the Pareto front ($layer - size(ls) = 16$, orange and green lines, and $beta = 64$, brown line) all other individual models achieve mean accuracy above $90\%$ for all spatial generalizations, also for vertical generalization (not shown) (Figure A.2). Figure 4A-Right further shows reconstruction examples: each of the plots shows 6 random samples (top) and their reconstruction (bottom) by a given model. Overall, the reconstruction is, qualitatively, similar to the original image, even for models with relatively low performance.

**Length Generalization – $\beta$-VAEs fail to generalize to longer word lengths:** Figure 4B-Left shows the reconstruction accuracy for all left-out word lengths, as measured with the Evaluator Model. Overall, reconstruction accuracy for short left-out word length is high, in particular for the four models with also better performance on spatial generalization. However, a qualitative inspection of random samples from these models (4B-Right, for examples) show that letter parts are nonetheless present in the reconstruction of images with a single letter, remnants from the longer words in the training data. In general, for longer word lengths, generalization performance decreases. In particular, generalization performance for words with five letters is lowest. A qualitative inspection shows that, indeed, the reconstruction of words with five letters contains in many cases only four letters (red rectangle).

**Compositional Generalization: $\beta$-VAEs fail to generalize to unseen compositions of letter identity and letter position:** Finally, the compositional test assessed the ability of the models to develop an abstract understanding of letter position. Overall, results show that the compositional test was most challenging for the models compared to the other two tests (Figure 4C-Left), with one model's performance approaching chance level. Figure 4C-Right illustrates the type of errors the models make (e.g., in red rectangles). For example, when presented with strings where 'B' was never present in the $5^{th}$ position in the training data, the models 'hallucinate' an 'A' in this position. Similar errors occur when 'B' or 'A' were omitted from other positions during training.

## 4.3. Neural Evaluation using Perturbation Experiments

We next studied to what extent beta-VAEs develop neural disentanglement of letter position and letter identity information. Figure 5A illustrates what such neural disentanglement could look like – it shows

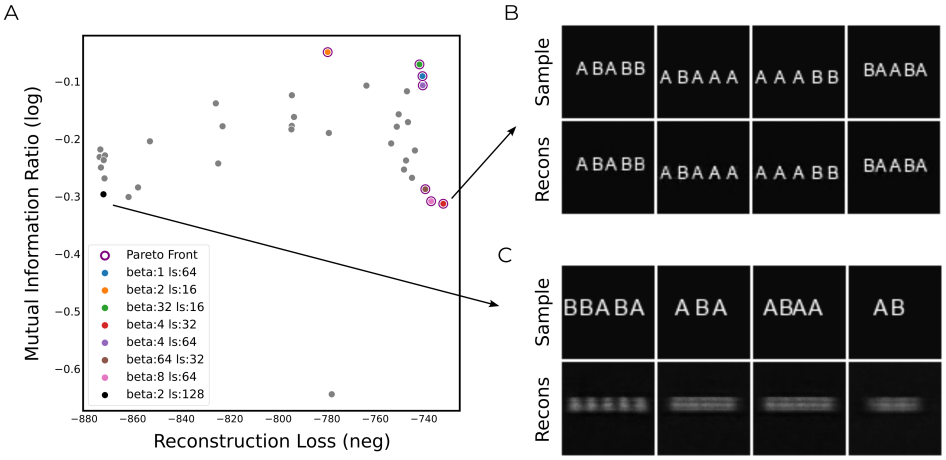


Figure 3: **Model Selection: (A)** We trained a large set of $\beta - VAEs$, with varying values of hyper-parameters, such as the beta value of the model or the size of their latent layer. An optimal $\beta - VAE$ would be one that achieves both good reconstruction *and* strong disentanglement. Here, we thus show the reconstruction loss against Mutual Information Ratio for all models. Each dot represents a model. Orange dots represents models from which samples were taken as examples in the right panels. **(B)** Reconstruction examples from the model with best reconstruction loss and MIG ($\beta$=4, Latent Size=32, Learning Rate=0.0001). **(C)** Reconstruction examples from a model with relatively poor reconstruction loss ($\beta$=2, Latent Size=128, Learning Rate=0.0001). Models marked with purple circles represent the Pareto Front.

a possible encoding scheme, where different units of the latent layer encode for different positions in the word, and different levels of activity encode for different level identities. Therefore letter identity is independently encoded of letter position.

To test whether such neural disentanglement emerges in the model, even partly, we conducted a perturbation experiment. That is, given an input image from the training set, we computed the neural activations at the latent layer of the model, and then separately for each unit, we systematically perturbed its activity to different levels. After each perturbation, we then reconstructed the image. The difference between the input image and the reconstructed image (after perturbation) is revealing about the information that the perturbed unit encodes. For example, if a model developed neural disentanglement of letter position and letter identity (Figure 5A), then perturbing one of the latent units can cause a replacement of one letter with another one, in only the perturbed position.

Figures 5B-D show examples from the perturbation experiments, from the model with the best reconstruction loss and strong performance on CompOrth ($\beta = 4$, *latent-size*$= 32$; red lines in Figure 4). Each panel corresponds to a latent unit; rows of each panel correspond to different samples (input images) and columns to different levels of perturbation. Unit 22 (Panel B) encodes vertical translation, partly entangled with letter identity. Unit 3 (Panel C) modulates word length, partly entangled as well with letter identity. Unit 23 (Panel D) mainly encodes letter identity, without spatial effects. Perturbation effects for all 32 units are shown in Figures A.3 & A.4.

However, analyzing all 32 latent units in the model, none fully disentangled letter identity and position (Figure 5A). No model on the Pareto-front achieved strong neural disentanglement. This is, in fact, consistent with their poor performance on CompOrth's Compositional-Generalization test – Limited neural disentanglement is consistent with poor compositionality and thus low performance on CompOrth.

### 4.4. The Relationship between Neural Disentanglement and Compositionality

While we haven't discovered strong neural disentanglement of identity and position in the previous section, a weak neural disentanglement might have nonetheless emerged in some of the models, which is hard to detect with mere perturbation experiments. Such weak disentanglement would possibly lead to a small, yet significant, improvement in performance on the Compositional-Generalization test in CompOrth.

We therefore next tested the hypothesis that neural disentanglement facilitates the separation of letter-positions and letter-identity information, and therefore, in turn, their composition. This predicts that models that achieve high neural disentangle-

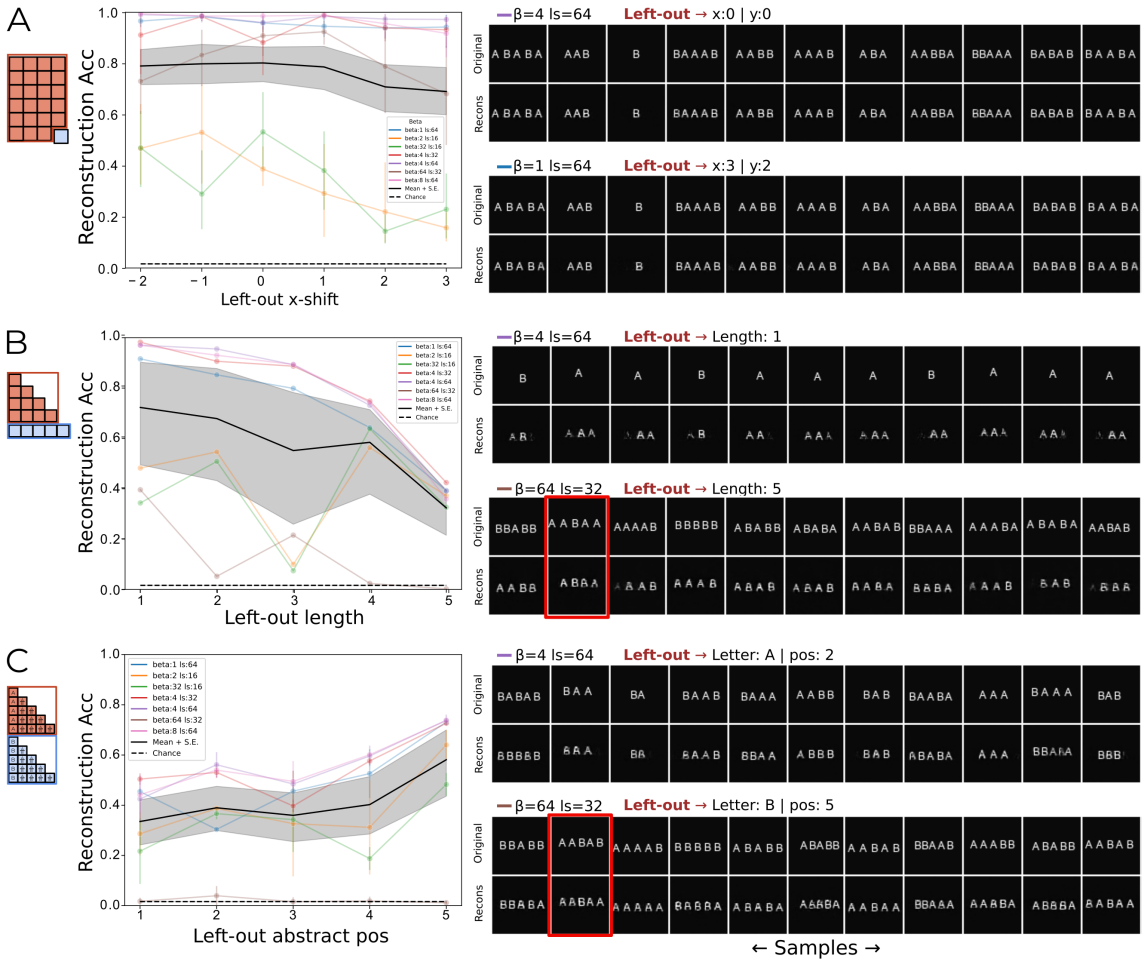


Figure 4: **Results on CompOrth tests.** Average Reconstruction Accuracy on the **(A)** Retinal-Position Test (for vertical axis see Figure A.2), **(B)** Word-Length Test, and **(C)** the Abstract-Position Test. Each of these tests evaluates the optimal models (Figure 3) on a left-out portion of the dataset, to assess its generalization performance, and in particular, its compositionality ability, as tested in panel C. For each model, dots represents the average across all the left-out splits and error bar represents the Standard Error. Black line shows the average across all the selected models with the Standard Error in gray shadow. Dashed lines mark the chance level for the classifier ($chance = 1/62$). On the right of each panel, several examples are shown for how the model reconstruct test images. Note the red marking on the images, which highlight the type of errors the model makes. Information about the generalization test is given at the top of each panel. For example, `Left-out → Letter: A | pos: 2` means that strings with the letter 'A' at the second position of the word were not seen during training.

ment, as measured by MIR, will achieve better performance on compositional generalization, as measured by the Compositional-Generalization test in CompOrth. To test this, we computed the correlation between the MIR and reconstruction accuracy on CompOrth for all models on the Pareto front. We found a weak correlation (Pearson coefficient $\rho = 0.13$). (Figure A.5). However, it was not statistically significant ($p - value = 0.26$).

## 5. Summary and Discussion

We introduced CompOrth, a novel benchmark for evaluating orthographic processing in visual models. The primary goal of CompOrth is to assess compositionality — the ability of a model to generalize to new combinations of letter identities and positions beyond the training set. This task is considered trivial for humans, so passing the CompOrth test is essential for a model to be regarded as achieving human-like performance.

We tested numerous VAEs, including $\beta$-VAEs, which encourage neural disentanglement. All models generalized well to unseen spatial locations

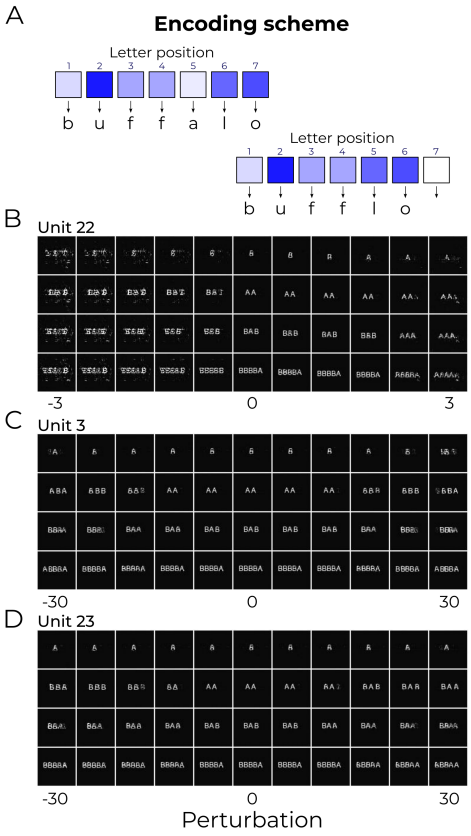


Figure 5: **Neural Perturbation Analyses (A)** Examples of a hypothetical encoding scheme: single neurons for positions, where the degree of activation indicates which letter is present in that position. **(B-D)** Perturbation results for example units from a model with strong performance on CompOrth ($\beta = 4$, *latent-size*$= 32$). Each row represents different samples (word images), columns represents different levels of perturbation.

and disentangled spatial factors in their latent layers. These results align with CNNs' inherent translation invariance—driven by weight sharing and pooling—allowing them to process patterns regardless of position (but see, Azulay and Weiss, 2019; Khemakhem et al., 2020). The match between the model's inductive bias and the dataset's spatial manipulations likely supported this disentanglement (Locatello et al., 2019).

However, all models failed CompOrth's compositional tests, failing to generalize to new letter-position combinations and longer word lengths. This shows models did not capture the underlying generative process—sampling combinations of letters in varying positions. Instead, results suggest the models rely on *memorizing* training data. This is observed when the model 'hallucinates' or adds letters, failing to reconstruct the input image.

The inductive biases of autoencoders—specifically the architecture's bottleneck and $\beta$-VAE regularization—could encourage models to learn underlying patterns through abstract representations rather than memorization. However, an overly constrained bottleneck can reduce reconstruction accuracy (Figure 3). Disentanglement (MIR) showed no correlation with generalization, consistent with prior work (Montero et al., 2021; Schott et al., 2021). The explored VAEs' inductive biases might not sufficiently align with the data's generative process, hindering the emergence of disentanglement (Locatello et al., 2019). While more data could theoretically compensate for weak priors (Welling, 2019; Goyal and Bengio, 2022), this is often infeasible, necessitating architectures with better-adapted inductive biases.

Future directions for orthographic processing include incorporating dual-route processing into computational models. Cognitive science distinguishes between habitual (System 1) and controlled (System 2) processing—fast and unconscious versus slow and attention-demanding. In orthography, behavioral and neuroimaging data suggest two distinct routes: lexical and sublexical, supported by memory-based and rule-based processing (e.g., Marshall and Newcombe, 1973; Coltheart et al., 2001; Jobard et al., 2003; Fiebach et al., 2002). Integrating these cognitive insights into inductive biases may improve out-of-distribution generalization (Goyal and Bengio, 2022). We suggest integrating dual-route architectures where one route relies on implicit symbol-based rules. Other future avenues include disentanglement via distributed equivariant operators (Bouchacourt et al., 2021), alternative notions of disentanglement (O'Neill et al., 2024), and vision transformers with contextual positional encoding (CoPE; Golovneva et al., 2024) for ordinal position.

We further hypothesized that neural disentanglement, which tends to emerge in $\beta$-VAEs with high values of $\beta$, would facilitate the separation of letter-positions and letter-identity information, and in turn, their composition. To test this, we conducted perturbation experiments with $\beta$-VAEs, to see whether some of the units disentangle identity and position information. Exploring all units in the model, we have not identified any such units, which is consistent with the failure of the models on CompOrth. However, a weak neural disentanglement of identity and position may have emerged in some of the models, unobserved by our perturbation experiments. We therefore tested whether there exists a correlation between MIR (a metric that quantifies neural disentanglement based on mutual information) and CompOrth performance across all our VAE models. We found no significant correlation between MIR and reconstruction accuracy on CompOrth.

The observed failure of the models in this study

is one more example of the shortcoming of artificial neural networks to dynamically and flexibly bind information, which might be distributively encoded in the network (Greff et al., 2020), even when neural disentanglement is explicitly optimized, as in $\beta$-VAEs. This binding problem affects the capacity of the models to achieve compositional ability by manipulating symbols (letters) and combining them in various, unbounded ways, as humans (Fodor and Pylyshyn, 1988; Lake and Baroni, 2018; Baroni, 2020). Similar limitations were also observed, for similar reasons, for shapes (Montero et al., 2021; Schott et al., 2021) and in language models (Delétang et al., 2022).

However, unlike existing benchmarks for testing compositionality in language models (e.g., Lake and Baroni, 2018), for orthographic processing and OCR, there are no existing targeted benchmarks. CompOrth therefore aims to fill this gap by providing means to evaluate compositionality in vision models.

## Ethical statement

This paper presents work whose goal is to bridge closer the fields of Machine Learning and Psycholinguistics; being theoretical in nature, we believe that no societal risks need to be specifically highlighted.

## Limitations

This study investigates the ability of a neural architecture to disentangle relevant information in the input for compositional generalization. One possible limitation is that the models explored here were trained on the CompOrtho dataset only. However, we note that while the models were trained from scratch on CompOrth, the challenges posed by the benchmark are also applicable to pre-trained models. These models can be refined and evaluated using the same approach (Figure 1A) to assess their compositional generalization capabilities.

## Acknowledgments

B.B. received a travel grant from the University of Buenos Aires (UBAint Docentes 2023) that made the completion of this study possible. B.B. was also supported by funding from CONICET.

## 6. Bibliographical References

Aakash Agrawal, KVS Hari, and SP Arun. 2020. A compositional neural code in high-level visual cortex can explain jumbled word reading. *Elife*, 9:e54846.

Alireza Azadbakht, Saeed Reza Kheradpisheh, and Hadi Farahani. 2022. Multipath vit ocr: A lightweight visual transformer-based license plate optical character recognition. In *2022 12th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 092–095. IEEE.

Aharon Azulay and Yair Weiss. 2019. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20(184):1–25.

Jeonghun Baek, Yusuke Matsui, and Kiyoharu Aizawa. 2021. What if we only use real datasets for scene text recognition? toward scene text recognition with fewer labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3113–3122.

Alejandro Baldominos, Yago Saez, and Pedro Isasi. 2019. A survey of handwritten character recognition with mnist and emnist. *Applied Sciences*, 9(15):3169.

Marco Baroni. 2020. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375(1791):20190307.

Jason JS Barton, Hashim M Hanif, Laura Eklinder Björnström, and Charlotte Hills. 2014. The word-length effect in reading: A review. *Cognitive neuropsychology*, 31(5-6):378–412.

Darwin Bautista and Rowel Atienza. 2022. Scene text recognition with permuted autoregressive sequence models. In *European conference on computer vision*, pages 178–196. Springer.

Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, and Yi-Zhe Song. 2021a. Towards the unseen: Iterative text recognition by distilling from errors. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14950–14959.

Ayan Kumar Bhunia, Aneeshan Sain, Amandeep Kumar, Shuvozit Ghose, Pinaki Nath Chowdhury, and Yi-Zhe Song. 2021b. Joint visual semantic reasoning: Multi-stage decoder for text recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14940–14949.

Diane Bouchacourt, Mark Ibrahim, and Stéphane Deny. 2021. Addressing the topological defects of disentanglement via distributed operators. *arXiv preprint arXiv:2102.05623*.

Thomas M Breuel, Adnan Ul-Hasan, Mayce Ali Al-Azawi, and Faisal Shafait. 2013. High-performance ocr for printed english and fraktur

using lstm networks. In *2013 12th international conference on document analysis and recognition*, pages 683–687. IEEE.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. 2018. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. 2017. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages 2921–2926. IEEE.

Laurent Cohen, Stéphane Lehéricy, Florence Chochon, Cathy Lemer, Sophie Rivaud, and Stanislas Dehaene. 2002. Language-specific tuning of visual cortex? functional properties of the visual word form area. *Brain*, 125(5):1054–1069.

Max Coltheart, Kathleen Rastle, Conrad Perry, Robyn Langdon, and Johannes Ziegler. 2001. Drc: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, 108(1):204.

Mengmeng Cui, Wei Wang, Jinjin Zhang, and Liang Wang. 2021. Representation and correlation enhanced encoder-decoder framework for scene text recognition. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part IV 16*, pages 156–170. Springer.

Colin J Davis and Jeffrey S Bowers. 2004. What do letter migration errors reveal about letter position coding in visual word recognition? *Journal of Experimental Psychology: Human Perception and Performance*, 30(5):923.

Stanislas Dehaene, Laurent Cohen, José Morais, and Régine Kolinsky. 2015. Illiterate to literate: behavioural and cerebral changes induced by reading acquisition. *Nature Reviews Neuroscience*, 16(4):234–244.

Stanislas Dehaene, Laurent Cohen, Mariano Sigman, and Fabien Vinckier. 2005. The neural code for written words: a proposal. *Trends in cognitive sciences*, 9(7):335–341.

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, et al. 2022. Neural networks and the chomsky hierarchy. *arXiv preprint arXiv:2207.02098*.

Yongkun Du, Zhineng Chen, Caiyan Jia, Xiaoting Yin, Tianlun Zheng, Chenxia Li, Yuning Du, and Yu-Gang Jiang. 2022. Svtr: Scene text recognition with a single visual model. *arXiv preprint arXiv:2205.00159*.

Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. 2021. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7098–7107.

Christian J Fiebach, Angela D Friederici, Karsten Müller, and D Yves Von Cramon. 2002. fmri evidence for dual routes to the mental lexicon in visual word recognition. *Journal of cognitive neuroscience*, 14(1):11–23.

Jerry A Fodor and Zenon W Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Francesco Fontanella, Francesco Colace, Mario Molinara, A Scotto Di Freca, and Filippo Stanco. 2020. Pattern recognition and artificial intelligence techniques for cultural heritage.

Benjamin Gagl, Fabio Richlan, Philipp Ludersdorfer, Jona Sassenhagen, Susanne Eisenhauer, Klara Gregorova, and Christian J Fiebach. 2022. The lexical categorization model: A computational model of left ventral occipito-temporal cortex activation in visual word recognition. *PLOS Computational Biology*, 18(6):e1009995.

Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. 2024. Contextual position encoding: Learning to count what's important. *arXiv preprint arXiv:2405.18719*.

Ehsan Goodarzi, Mina Ziaei, and Edward Zia Hosseinipour. 2014. *Introduction to optimization analysis in hydrosystem engineering*, volume 25. Springer.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Anirudh Goyal and Yoshua Bengio. 2022. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266):20210068.

Jonathan Grainger and Carol Whitney. 2004. Does the huamn mnid raed wrods as a wlohe? *Trends in cognitive sciences*, 8(2):58–59.

Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. 2020. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*.

T Hannagan, A Agrawal, L Cohen, and S Dehaene. 2021. Emergence of a compositional neural code for written words: Recycling of a convolutional neural network for reading. *Proceedings of the National Academy of Sciences*, 118(46):e2104779118.

Irina Higgins, Le Chang, Victoria Langston, Demis Hassabis, Christopher Summerfield, Doris Tsao, and Matthew Botvinick. 2021. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature communications*, 12(1):6456.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*.

Mohit Jain, Minesh Mathew, and CV Jawahar. 2017. Unconstrained ocr for urdu using deep cnn-rnn hybrid networks. In *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 747–752. IEEE.

Gaël Jobard, Fabrice Crivello, and Nathalie Tzourio-Mazoyer. 2003. Evaluation of the dual route theory of reading: a metanalysis of 35 neuroimaging studies. *Neuroimage*, 20(2):693–712.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. 2020. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pages 2207–2217. PMLR.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13094–13102.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR.

Pengyuan Lyu, Chengquan Zhang, Shanshan Liu, Meina Qiao, Yangliu Xu, Liang Wu, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. 2022. Maskocr: Text recognition with masked encoder-decoder pretraining. *arXiv preprint arXiv:2206.00311*.

John C Marshall and Freda Newcombe. 1973. Patterns of paralexia: A psycholinguistic approach. *Journal of psycholinguistic research*, 2:175–199.

Milton Llera Montero, Casimir JH Ludwig, Rui Ponte Costa, Gaurav Malhotra, and Jeffrey Bowers. 2021. The role of disentanglement in generalisation. In *International Conference on Learning Representations*.

Yongqiang Mou, Lei Tan, Hui Yang, Jingying Chen, Leyuan Liu, Rui Yan, and Yaohong Huang. 2020. Plugnet: Degradation aware scene text recognition supervised by a pluggable super-resolution unit. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 158–174. Springer.

Asma Naseer and Kashif Zafar. 2019. Meta features-based scale invariant ocr decision making using lstm-rnn. *Computational and Mathematical Organization Theory*, 25:165–183.

Charles O'Neill, Christine Ye, Kartheik Iyer, and John F Wu. 2024. Disentangling dense embeddings with sparse autoencoders. *arXiv preprint arXiv:2408.00657*.

Lukas Schott, Julius Von Kügelgen, Frederik Träuble, Peter Gehler, Chris Russell, Matthias Bethge, Bernhard Schölkopf, Francesco Locatello, and Wieland Brendel. 2021. Visual representation learning does not generalize strongly within the same domain. *arXiv preprint arXiv:2107.08221*.

Baoguang Shi, Xiang Bai, and Cong Yao. 2016. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304.

Masanori Suganuma, Mete Ozay, and Takayuki Okatani. 2018. Exploiting the potential of standard convolutional autoencoders for image

restoration by evolutionary search. In *International Conference on Machine Learning*, pages 4771–4780. PMLR.

JSH Taylor, Kathleen Rastle, and Matthew H Davis. 2013. Can cognitive models explain brain activation during word and pseudoword reading? a meta-analysis of 36 neuroimaging studies. *Psychological bulletin*, 139(4):766.

Zhaoyi Wan, Minghang He, Haoran Chen, Xiang Bai, and Cong Yao. 2020. Textscanner: Reading characters in order for robust scene text recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12120–12127.

Yuxin Wang, Hongtao Xie, Shancheng Fang, Jing Wang, Shenggao Zhu, and Yongdong Zhang. 2021. From two to one: A new scene text recognizer with visual language modeling network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14194–14203.

Max Welling. 2019. Do we still need models or just more data and compute. *University of Amsterdam*, 7.

James CR Whittington, Will Dorrell, Surya Ganguli, and Timothy Behrens. 2022. Disentanglement with biological constraints: A theory of functional cell types. In *The Eleventh International Conference on Learning Representations*.

Ruijie Yan, Liangrui Peng, Shanyu Xiao, and Gang Yao. 2021. Primitive representation learning for scene text recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 284–293.

Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. 2020. Towards accurate scene text recognition with semantic reasoning networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12113–12122.

Xiaoyu Yue, Zhanghui Kuang, Chenhao Lin, Hongbin Sun, and Wayne Zhang. 2020. Robustscanner: Dynamically enhancing positional clues for robust text recognition. In *European Conference on Computer Vision*, pages 135–151. Springer.

Haochen Zhang, Dong Liu, and Zhiwei Xiong. 2017. Cnn-based text image super-resolution tailored for ocr. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE.

Hui Zhang, Quanming Yao, Mingkun Yang, Yongchao Xu, and Xiang Bai. 2020. Autostr: efficient backbone search for scene text recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 751–767. Springer.

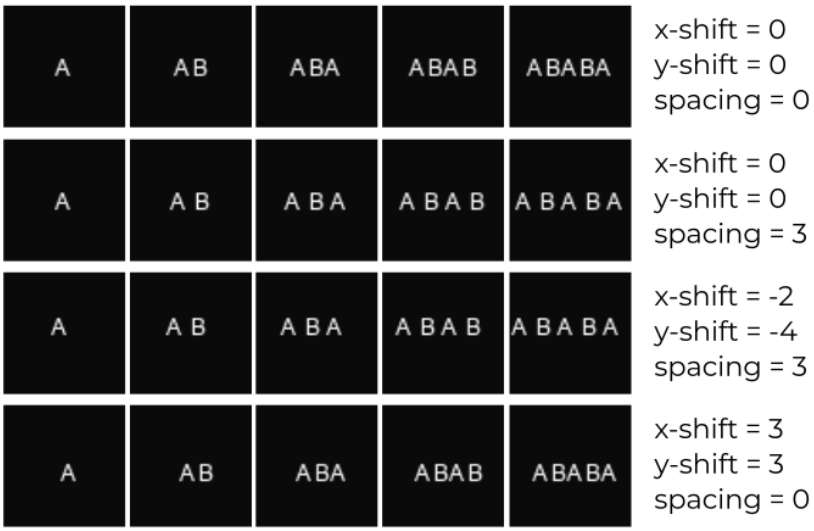


Figure A.1: **Examples from the generated dataset and model training.** All the images are comprised by a string of 1 to 5 letters, using only the uppercase characters A and B. To generate variations of this strings the spacing and the x and y position were modified.

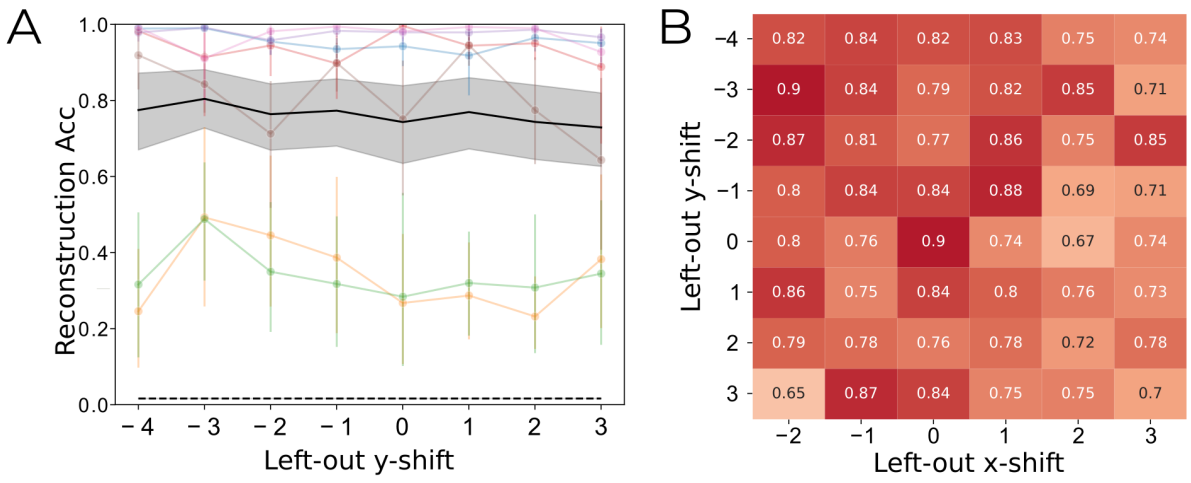


Figure A.2: **Reconstruction accuracy for Spatial Generalization test. (A)** averaged by model, across y-shift; and **(B)** average for each combination of x- and y-shift, across models.

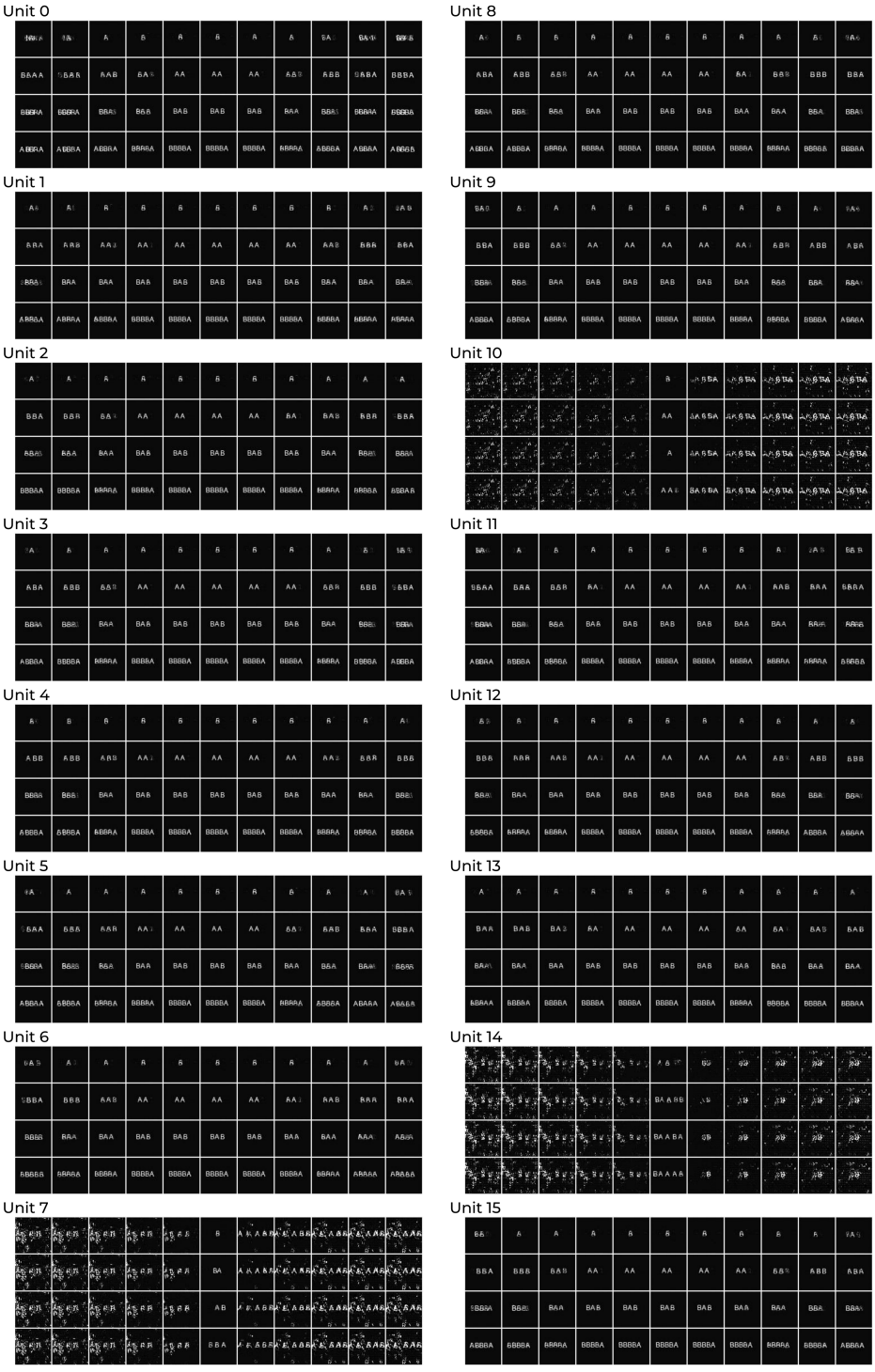


Figure A.3: **Neural Perturbation Analyses.** Perturbation results for example units from a model with strong performance on CompOrth ($\beta = 4$, $latent - size = 32$). First 16 neurons of the model.

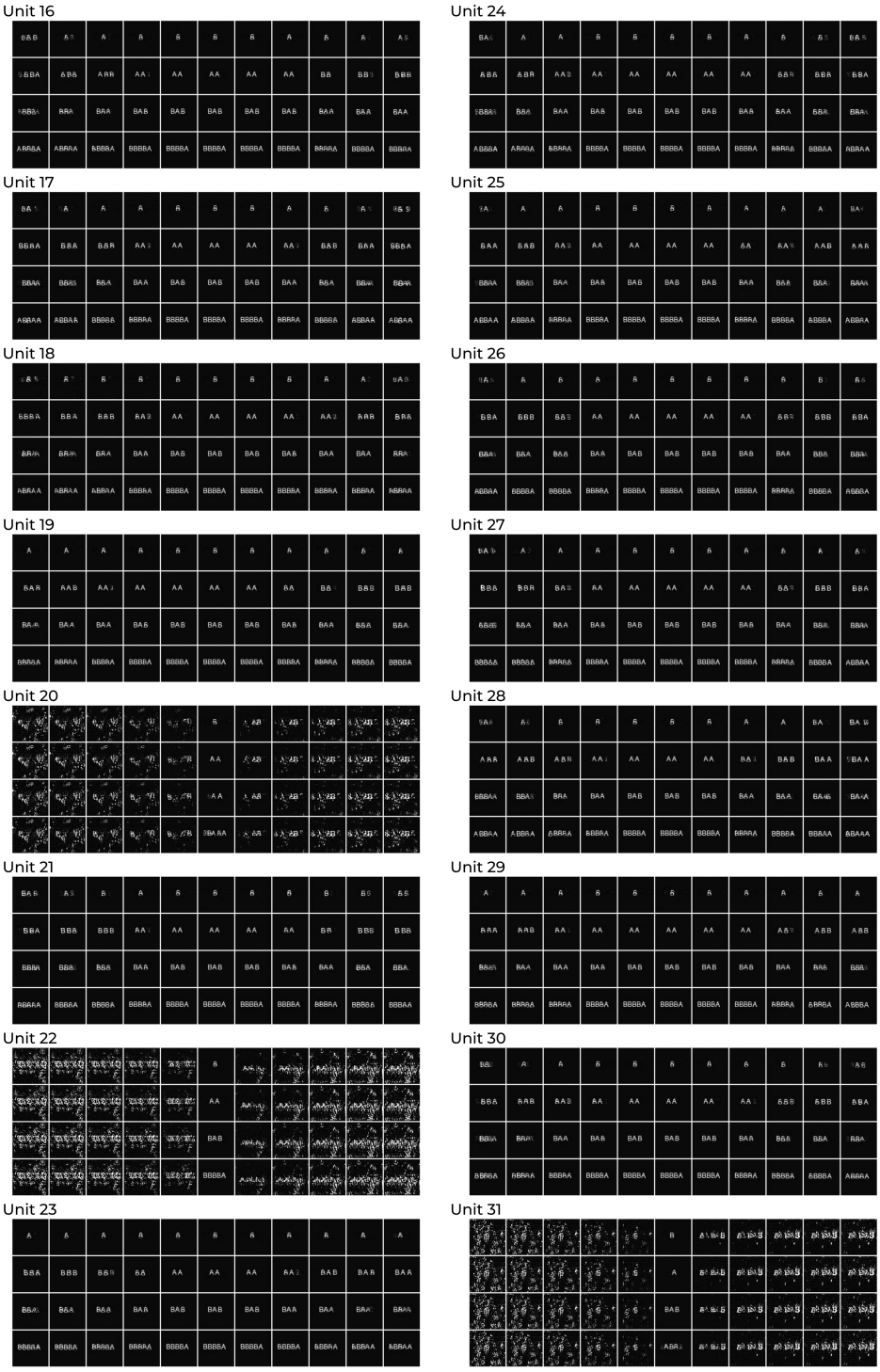


Figure A.4: **Neural Perturbation Analyses.** Perturbation results for example units from a model with strong performance on CompOrth ($\beta = 4$, $latent - size = 32$). Last 16 neurons of the model.

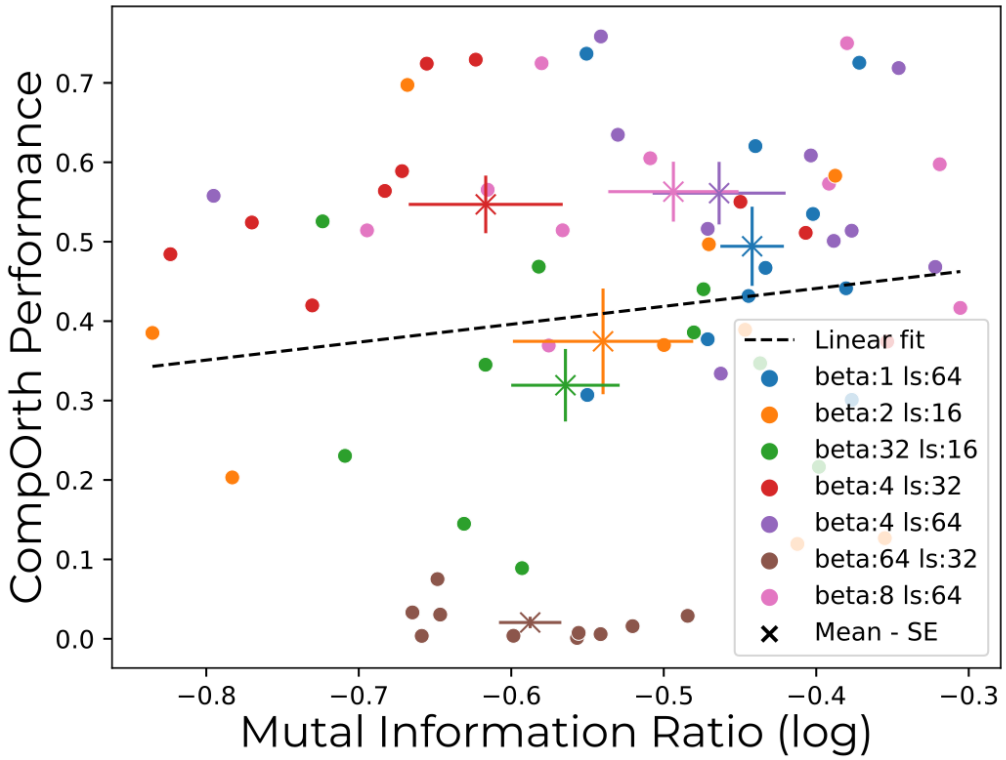


Figure A.5: **Mutual Information Ratio (MIR) vs Reconstruction Accuracy for Compositional Generalization analyses.** We computed MIR using the 5-letter word identity encoding (Figure 5A) across all Pareto Front models. Each model's results per test split are shown as colored circles, with crosses indicating their mean performance across splits. Error bars reflect SEM, and the dashed line shows the linear trend. Pearson correlation resulted in $\rho = 0.12$ and $p - val = 0.30$.