CollageNoter: Real-Time and Adaptive Collage Layout Design for Screenshot-Based E-Note-Taking

Qiuyun Zhang [1], Bin Guo [1*], Lina Yao [2], Xiaotian Qiao [3, 4], Ying Zhang [1], Zhiwen Yu [5,1]

[1] School of Computer Science, Northwestern Polytechnical University
[2] CSIRO's Data61
[3] School of Computer Science and Technology, Xidian University
[4] Guangzhou Institute of Technology, Xidian University
[5] College of Computer Science and Technology, Harbin Engineering University
qiuyunzhang@mail.nwpu.edu.cn, guob@nwpu.edu.cn, lina.yao@data61.csiro.au, qiaoxt1992@gmail.com, izhangying@nwpu.edu.cn, zhiwenyu@nwpu.edu.cn

Abstract

To enhance the processing of complex multi-modal documents (*e.g.* e-books, long web pages, etc.), it is an efficient way for users to take digital screenshots of key parts and reorganize them into a new collage E-Note. Existing methods for assisting collage layout design primarily employ a semantic relevance-first strategy, with arranging related contents together. Though capable, it can not ensure the visual readability of screenshots and may conflict with human natural reading patterns. In this paper, we introduce CollageNoter for real-time collage layout design that adapts to various devices (*e.g.* laptop, tablet, phone, etc.), offering users visually and cognitively well-organized screenshot-based E-Notes. Specifically, we construct a novel two-stage pipeline for collage design, including 1) readability-first layout generation and 2) cognitive-driven layout adjustment. In addition, to achieve real-time response and adaptive model training, we propose a cascade transformer-based layout generator named CollageFormer and a size-aware collage layout builder for automatic dataset construction. Extensive experimental results have confirmed the effectiveness of our CollageNoter.

Introduction

Note-taking is vital for comprehending complex information by aggregating multiple contents into an integrated view. When users are dealing with multi-modal digital documents (*e.g.* long web pages, e-books, etc.), note-taking benefits them from quickly processing information to creating comprehensive study summarization (Qiao, Cao, and Lau 2022). With the increasing of various E-Note tools such as Notability [1], Microsoft onenote [2], etc., it is convenient for users to capture screenshots of crucial contents and then arrange them upon a new digital canvas for note-taking. Given the collage layout is suitable for organizing rich information (Dayama et al. 2020), existing tools provide basic collage layout assistance like automatic grid line alignment, size adjustment, etc. However, it is still time-consuming for users to make a well-organized layout for notes, as cognitive coherence and readability concerns exert different influences on the layout design. The former focuses on organizing images in a cognitive-aware manner that is easy for readers to understand, which is related to the logical relations among screenshots and humans' reading habits on the collage layout. The latter seeks to avoid visual distortion by optimizing the use of empty spaces, which may sometimes conflict with the cognitive-aware manner. Therefore, it is crucial to achieve a trade-off on these factors for high-quality collage layout generation of screenshot-based E-Note-taking.

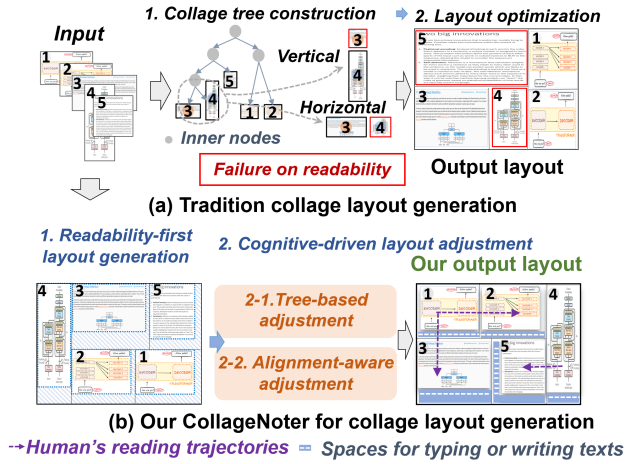

Figure 1: Comparison of tradition collage layout generation method(a) with our novel CollageNoter(b). While tradition methods tend to cause visual distortion due to their relevance-first layout optimization strategy, CollageNoter adopts a novel two-stage procedure including 1) data-driven readability-first layout generation and 2) cognitive-driven layout adjustment, which can produce visually appealing and cognitively intuitive layouts in real-time.

Studies in the field of natural image collage layout generation(Fan 2012; Liang et al. 2017) and graphic layout generation (Li et al. 2018; Jyothi et al. 2019) are highly relevant to ours. As for the former, tree-based methods achieve

*Corresponding author

[1]https://notability.com
[2]https://www.onenote.com

Dataset	Scale	Canvas	Empty↓	Over↓	Align↓
Magazine	4K+	(225, 300)	0.01-0.89 (M=0.249, SD=0.03)	0.00-0.20 (M=0.028, SD=0.01)	0.00-17.51(M=0.582, SD=2.63)
Rico	30K+	(440, 2392)	0.02-0.94 (M=0.419, SD=0.10)	0.00-0.97 (M=0.539, SD=0.22)	0.00-74.25(M=0.869, SD=18.4)
PublayNet	360K+	(791,613)	0.15-0.99 (M=0.458, SD=0.13)	0.00-0.17 (M=0.001, SD=1.38)	0.00-42.01(M=0.201, SD=1.46)
PKU	9K+	(750, 513)	0.03-0.96 (M=0.835, SD=0.03)	0.00-0.87 (M=0.080, SD=0.01)	0.00-49.01(M=1.394, SD=5.04)
CollageLayout	30K+	HD, UHD, UR	0	0	0

Table 1: Comparison of datasets for graphic layout generation, where lower Empty, Over, and Align indicate better canvas utilization, lower overlap, and improved alignment, respectively. Existing datasets are designed for a single fix-sized canvas and not suitable for training collage layout generator, while our CollageLayout is alignment, filling the canvas, and none-overlap, which contains layouts of high-definition (HD), Ultra high-definition (UHD) and ultra-wide resolutions (UR).

the best performances (Gan et al. 2020; Pan et al. 2019; Yu et al. 2022), which obtain the layout based on a two-step procedure as shown in Fig. 1(a). Initially, a binary tree is constructed for input images in a bottom-up manner by iteratively merging the two most semantically related nodes to a new one, where the semantic feature for the new node is the fusion of its children. Then, by predicting the placement (*e.g.* vertical or horizontal) operation of its children for each inner node, the collage layout can be determined, which is optimized under the constraints of minimizing the deformation of images and blank spaces as much as possible. In the latter field, advancements in deep learning have led to the successful application of generative neural networks, particularly transformers (Gupta et al. 2021; Jiang et al. 2023), in graphic layout generation, which determine the size and position for each input design element upon a fixed-size canvas conditioned on their category labels, spatial relation constraints, etc. Therefore, deep neural networks have already been applied to numerous graphic layout design scenarios.

Although existing works in these two fields have provided valuable insights, each has its own limitations. ***First***, the widely adopted semantic relevance-first tree for collage layout generation may cause unavoidable visual distortion and cognitive gaps. As shown in Fig. 1(a), when two elements with significantly different aspect ratios have to be placed together (element-3 and element-4) due to their highest semantic relevance, both vertical and horizontal arrangements may cause deformation. In addition, as existing layout optimization ignores human reading habits like the preferred reading trajectories illustrated with purple dotted arrows in Fig. 1, viewers have to spend extra effort to discern the correct logical order of screenshots when they read the collage design, thereby increasing the cognitive burden. ***Second***, as for deep-learning-based methods, existing open-source layout datasets summarized in Table. 1 are not suitable for note-taking and tail to handle diverse device sizes. What's more, since there exists a trade-off between readability and cognitive coherence, it is challenging to solely depend on data-driven methods to address E-note design.

In this paper, we propose the CollageNoter with the following improvements to solve the aforementioned issues. ***First***, we introduce a novel two-stage pipeline as shown in Fig. 1(b), including 1) readability-first layout generation and 2) cognitive-driven layout adjustment. The first stage places images into a collage layout while preserving their original aspect ratio and size [3]. The second stage refines the layout based on two human-centric strategies including the tree-based order adjustment strategy and the alignment-aware grouping adjust strategy, which aim to align the logical order of screenshots with human natural reading order of the layout. ***Second***, as optimization-based layout searching is time-consuming due to the vast solution space for possible layouts, we introduce transformer neural networks for readability-first layout generation. Specifically, as shown in Fig. 1(b) on the left, when the input screenshots can not fully cover the whole canvas, the layout generator should be able to arrange the remaining black regions for each element (illustrated with the blue dotted line filled areas) to form the overall alignment collage layout. Therefore, we propose the CollageFormer with a generative adversarial network, which includes a cascade generator decoupled into position and size prediction, and a discriminator to make generated layouts more closely aligned with the real-world designs.

Our contributions are summarized as follows:

- We propose a novel two-stage procedure for screenshot collage layout design, considering both readability and cognitive coherence to ensure that generated layouts are both visually appealing and cognitively accessible.
- To generate high-quality results and achieve real-time response, we introduce a novel data-driven cascade transformer-based generator with an additional discriminator into collage layout generation.
- To achieve adaptive E-note generation, we build a collage layout dataset builder to construct training dataset for different device sizes, and a collage layout dataset including layout instances of commonly used device sizes.
- Both qualitative and quantitative experiments validate the effectiveness of our approach compared to several strong baselines. Moreover, we conduct user studies to demonstrate the advantages our methods provide to users.

Related Work

Optimization Based Collage Layout Generation Existing image collage generation studies focus on the photo album arrangement (Liang et al. 2017) to plan a group of photos

[3]Preserving aspect ratio ensures visual consistency. Maintaining size prevents disrupting the original hierarchy of text in screenshots due to scaling, which is typically indicated by font size

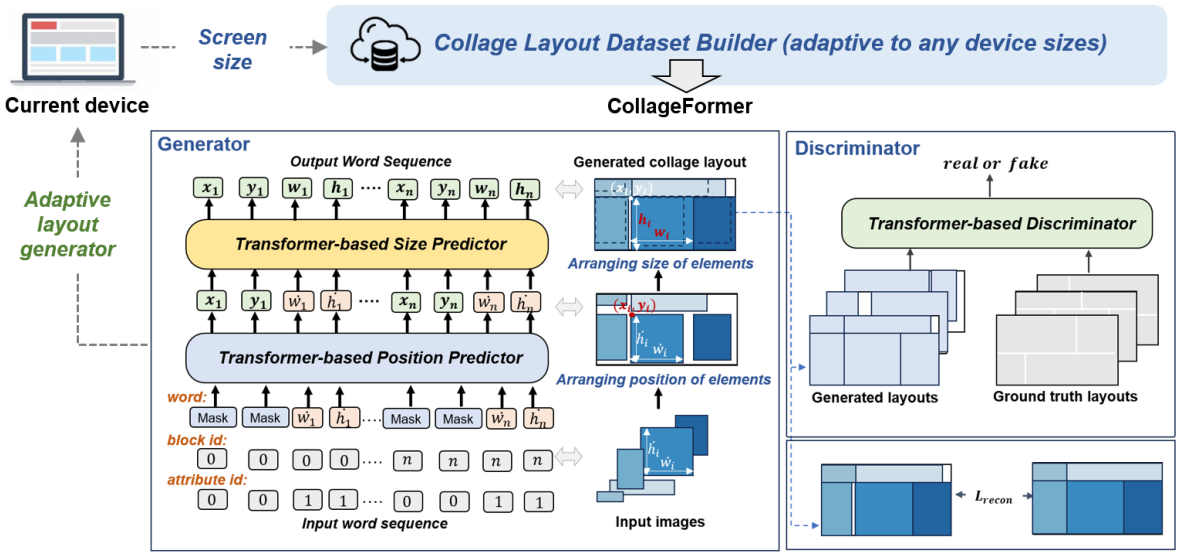

Figure 2: Illustration of the CollageFormer for readability-first layout generation, which includes a cascade generator decoupled into position prediction and size prediction stages, a discriminator taking generated and real-world collage layouts as input, and a supervised reconstruction loss measuring the accuracy of generated layouts compared to the corresponding ground truths. Specifically, we construct a device-adaptive layout dataset builder to automatically provide training data for any devices.

on a fixed-size canvas, which can be mainly divided into the following two categories. First, parametric-based methods (Wei, Matsushita, and Yang 2009) model the collage generation task with manually defined objective functions on a set of geometric variables of input images, which include position, size, orientation, etc. They depend on the probabilistic graphical framework (Rother et al. 2006; Liu et al. 2009) or the heuristic manner (Gan et al. 2020) to solve the multi-objective problem and struggle to deal with multiple input images. Second, partition-based methods (Fan 2012; Geigel and Loui 2000; Yu et al. 2022) generate the collage layout in a top-down manner, which first iteratively divide the canvas into two parts with a vertical or horizontal cut line until enough regions are obtained, and then put images into corresponding regions. However, though such top-down manner can eliminate the overlap of images, regions not generated according to the aspect ratio of input images tend to arouse image distortions. Therefore, genetic algorithm (Fan 2012), greedy strategy (Wu and Aizawa 2016), back propagation (Pan et al. 2019; Yu et al. 2022), etc. are introduced to enhance the layout generation, which refine the partition process for better ratio preserving.

Deep-learning Based Layout Generation The successful applications of deep generative models have greatly benefited the graphic layout generation. As generative adversarial networks (GANs) and variational auto-encoders (VAEs) can learn effective features for generating layouts similar to real-world data, early works (Li et al. 2018; Zhou et al. 2022) depend on them to build layout generators. Recently, Gupta et al. (Gupta et al. 2021) utilize a classic Transformer network to auto-regressively generate layouts, whose experimental results confirm that self-attention is able to extract features representing layout design context. Yang et al. (Yang et al. 2021) build the LayoutTransformer to handle scene graphs. Very recent methods focus on further improving transformer-based networks. Kong et al. (Kong et al. 2022) introduce bi-directional transformers with a hierarchical sampling policy. Tang et al. (Tang et al. 2023) further improve the layout generation quality with a graph Transformer generative adversarial network. Jiang et al. (Jiang et al. 2023) propose a unified representation for diverse layout generation tasks with constrained decoding.

Method

As shown in Fig. 1(b), our CollageNoter contains two modules. First, the readability-first layout generation is based on our novel CollageFormer, which takes a sequence of screenshots as input and predicts a layout with only considering the aspect ratio and size of each screenshot. Second, the cognitive-driven layout adjustment refines the generated layout to enhance cognitive coherence with the tree-based order adjustment strategy and alignment-aware grouping adjustment strategy. Details are as follows.

Readability-first Layout Generation

Overview of CollageFormer. As shown in Fig. 2, in light of the great improvements made by transformer-based layout generation methods, our CollageFormer contains three modules including a cascade transformer-based generator decoupled into position prediction and size prediction stages, a discriminator taking generated and real-world collage layouts into consideration, and a supervised reconstruction loss measuring the accuracy of generated layouts compared to the corresponding ground truth.

Problem Formulation. The input screenshots are denoted as $I = \{I_1, I_2, ..., I_n\}$. For each screenshot I_i, corre-

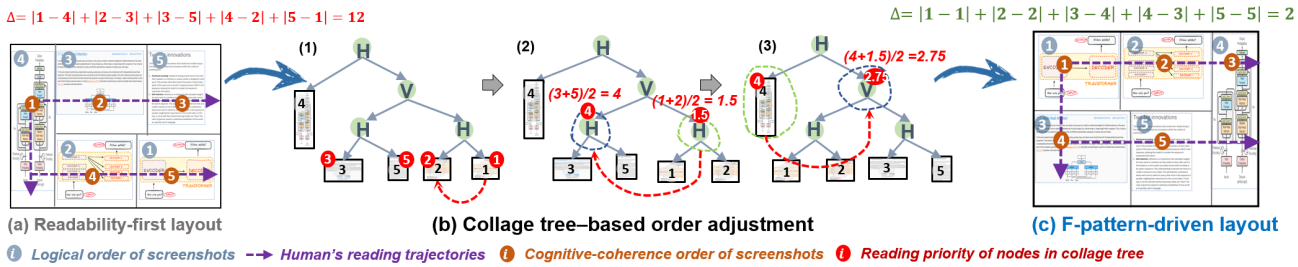

Figure 3: Illustration of Tree-based order adjustment strategy, which includes refactoring collage tree from the generated layout and bottom-up-based tree adjustment. Purple dashed arrows indicate human-preferred reading trajectories following the 'F-pattern' for informative collage designs, blue and orange circles denote the logical and cognitive-coherence order of screenshots, respectively. Obviously, our Tree-based order adjustment strategy significantly reduces the cognitive distance Δ.

sponding original $(\dot{w}_i, \dot{h}_i)$ for its width and height are automatically detected. The output layout is formulated as $\{(x_i, y_i, w_i, h_i)\}_{i=1}^{n}$, where $(x_i, y_i), w_i, h_i$ are the top-left coordinates, width and height of each element. The deviation from $(\dot{w}_i, \dot{h}_i)$ to (w_i, h_i) accounts for additional black spaces required in some regions to ensure overall alignment.

Model Details. The input word sequence is formulated as $X = \{unk, unk, \dot{w}_1, \dot{h}_1, ...unk, unk, \dot{w}_n, \dot{h}_n\}$. As shown in Fig. 2, we decompose the embedding of each word X_i into three components covering intra-instance and inter-instance information. First, the ***word embedding***, noted as E_w, captures semantic features of each word in the vector space. Second, we define each $[unk, unk, w_i, h_i]$ as a block corresponding to the same instance I_i. The ***block-id embedding***, noted as E_b, identifies different blocks. Third, the ***attribute-id embedding***, noted as E_a, differentiates words related to size or position attributes. The final input embedding is formulated by Eq. 1, where $\bigoplus$ is the concatenation operation. The CollageFormer first generates the position of each instance through a transformer-based position predictor as Eq. 2, where f_{e1} is a standard transformer encoder.

$$E_{i=1\sim 4n} = E_w \bigoplus E_b \bigoplus E_a \tag{1}$$

$$L_{position} = \{(x_i, y_i, \dot{w}_i, \dot{h}_i)\}_{i=1}^{n} = f_{e1}(E_{i=1\sim 4n}) \tag{2}$$

Based on the predicted $L_{position}$, the $\dot{E}_{i=1\sim 4n}$ is updated and the final size of each instance is generated according to Eq. 3 by the transformer-based size predictor, where f_{e2} is another standard transformer encoder.

$$L_{final} = \{(x_i, y_i, w_i, h_i)\}_{i=1}^{n} = f_{e2}(\dot{E}_{i=1\sim 4n}) \tag{3}$$

As for loss functions, the widely used reconstruction loss (Jiang et al. 2023; Kong et al. 2022) for measuring the deviation between a generated layout and the corresponding ground truth can not effectively address overlapping among design elements. Given the combination of the reconstruction loss and adversarial objective benefits the model training (Isola et al. 2017), we introduce an additional discriminator to our CollageFormer, which further constrains the generator to create fake layouts that are sufficiently similar to the ground high-quality layouts and deceives the discriminator into recognizing them as true. The transformer-based discriminator takes generated collage layout and real samples as input, with the same embedding function as the generator. The final loss is calculated by Eq. 4, where L_{recon} is the reconstruction loss, L_{adv} is the adversarial loss.

$$L = L_{recon} + \lambda L_{adv} \tag{4}$$

Cognitive-driven Layout Adjustment

Though there exists extensive research on collage layout generation, the human-centric concern has not been taken into consideration. Based on studies in the field of cognitive science (Mikolov et al. 2013; Perls, Hefferline, and Goodman 1951) explaining humans reading habits on collage informative designs, we first introduce two concepts for cognitive-coherence collage layout generation.

- Logic order of screenshots for users to consistently understand them, which is formulated as the chronological order of screenshots in which they were captured by users in this paper [4].
- Human-preferred reading order of regions within a collage layout, which can be inferred by the structure of the layout and human reading habits when they process informative collage designs.

The cognitive-coherence layout should align the logical order of screenshots with the human reading order, which can help users understand the information continuously rather than in a disjointed manner. In other words, the screenshots should be arranged in logical order along the collage regions in human-preferred reading order.

Given the readability-first layout only considers the aspect ratio and size of each screenshot, our CollageNoter introduces the following layout adjustment strategies to rearrange the layout to a both visually and cognitively appealing result (details are added in the supplementary).

Tree-based Order Adjustment Strategy. As illustrated with purple dotted arrows in Fig.3, humans tend to read complex information according to the F-pattern reading trajec-

[4]Other measurements could be easily integrated into our layout adjuster

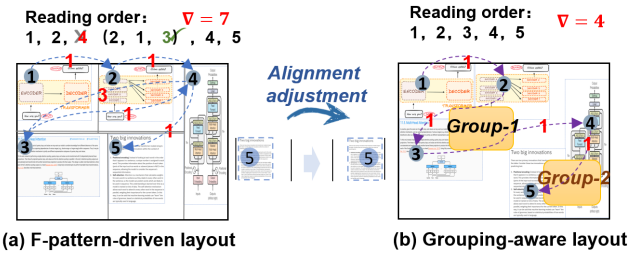

Figure 4: Illustration of alignment-aware grouping adjustment strategy, where refining the direction of the alignment helps to distinguish different semantic groups. Aligning image 5 to the right enhances the distinction between Group 1 (images 1,2, and 3) and Group 4 (images 4 and 5).

tories (Moen and Fee 2000; Mikolov et al. 2013), processing the collage regions row by row from top-left to right-down. Therefore, a cognitively coherent layout should bring the logical order (illustrated with blue circles) of screenshots close to the F-pattern reading order (illustrated with orange circles). To measure the differences between the two orders, as shown in Fig.3, we introduce a cognitive reading distance, noted as Δ, which sums the absolute differences between the numbers of the reading order and the logical order for each element. As shown in Fig.3(a), the cognitive distance for the readability-first layout is 12, requiring significant effort from users to comprehend the overall design.

However, directly mapping screenshots to their corresponding reading regions will cause visual distortion. Therefore, a tree-based order adjustment strategy is proposed for the collage layout to maintain the readability of each screenshot while narrowing the gap between reading order and logic order. As illustrated in Fig. 3(b), a binary collage tree is first refactored from the generated layout, where inner nodes represent vertical or horizontal placements and leaf nodes indicate corresponding screenshots. Then, we assign the reading priority order to each node (illustrated with red circles), with smaller values indicating higher priority. For each leaf node, the reading priority order equals the logic number of the corresponding screenshot. For each inner node, the reading priority order is the average of its children's. Then, the readability-first collage tree is refined layer by layer in a bottom-up manner, iteratively ensuring that higher-priority nodes become the left children, which arranges these nodes further to the left and above in the collage layout, making them more accessible to readers following the reading order. Finally, as shown in Fig. 3(c), the F-pattern-driven layout reduces the cognitive reading distance from 12 to 2.

Alignment-aware Grouping Adjustment Strategy. To further reduce users' effort in understanding screenshots, we propose the alignment-aware grouping adjust strategy, which is inspired by the gestalt cognitive principle (Perls, Hefferline, and Goodman 1951; Mann 2020) that humans tend to perceive adjacent elements as a whole. To further measure the reading efforts that viewers need to afford, we introduce the actuarial reading distance noted as ∇, which indicates the total distances users need to go through to process all screenshots in a logically correct manner. As shown in Fig. 4(a), all screenshots are displayed with left alignment, where the F-patter reading order is aroused and users read 1-2-4(wrong)-2-1-3(correct)-4-5 to cover all contents. By changing the alignment direction of screenshots, the visual grouping will be aroused where humans tend to first notice the adjacent elements, and thus F-patter reading trajectories are changed. As shown in Fig. 4(b), changing screenshot 5 to be right-aligned could guide users to distinguish two groups easily, which reduces the ∇ from 7 to 4.

Benchmark Construction

Many public datasets related to different graphic layout design tasks have been proposed including Magazine (Zheng et al. 2019) for magazine, Rico (Deka et al. 2017) for user interface, PubleyNet (Zhong, Tang, and Yepes 2019) for document, and PKU (Zhou et al. 2022) for advertisements, etc. However, existing datasets are not suitable for the collage layout design. As illustrated in Table 1, it is evident that magazines and advertisements suffer from excessive overlapping. Though the document layout is well-aligned, it is constructed for only one type of canvas size.

To achieve adaptive model training and inference, we propose a new collage layout dataset builder to automatically construct datasets for any device. In addition, we collect a dataset named CollageLayout, which includes layouts for high-definition (HD, 1920x1080), Ultra high-definition (UHD, 3840x2160) and ultra-wide resolutions (UR, 2560x1080) screen sizes, which are commonly used in daily life. Each collage layout is accompanied by the geometric parameters for design elements, noted as their bounding boxes $bb = \{[x_i, y_i, w_i, h_i]\}$, standing for top-left coordinates, width and height. In addition, as shown in Table. 1, collage layouts in our dataset strictly adhere to the requirements of no overlapping, proper alignment, and full canvas utilization (detailed in supplementary).

Experiment

Evaluation Metrics

First, we employ the metrics including ***Miou***, ***Over*** and ***Align*** commonly used in the literature (Kong et al. 2022; Jiang et al. 2023). Miou measures how well the generated layout matches the corresponding ground truth. Over calculates the total overlapping area between any pair of bounding boxes within a layout. Align evaluates the extent to which elements in graphic design are aligned by their center or edge. Second, since collage design focuses on making full use of the canvas, we introduce the occupancy ratio to gauge the proportion of canvas occupied by all elements by Eq. 5, where $\phi_{polygon}$ calculates the total area covered by the bounding boxes, noted as $\{(x_i, y_i, w_i, h_i)\}_{i=1}^{n}$, counting overlapping areas only once. C_w and C_h note the width and height of the canvas. The ***Empty*** metric is $1 - Occupy$.

$$Occupy = \frac{\phi_{polygon}(\sum_i^n bb_i)}{C_w \times C_h} \tag{5}$$

Method	HD				UHD				UR			
	Miou↑	Over↓	Align↓	Empty↓	Miou↑	Over↓	Align↓	Empty↓	Miou↑	Over↓	Align↓	Empty↓
Trans (Gupta et al. 2021)	0.633	0.015	0.041	0.064	0.541	0.043	0.051	0.073	0.603	0.025	0.059	0.066
BLT (Kong et al. 2022)	0.831	0.011	0.027	0.033	0.686	0.037	0.045	0.049	0.812	0.019	0.039	0.033
Former++ (Jiang et al. 2023)	0.797	0.015	0.033	0.049	0.645	0.039	0.049	0.064	0.751	0.026	0.037	0.042
Ours	**0.873**	**0.007**	**0.015**	**0.026**	**0.723**	**0.031**	**0.036**	**0.041**	**0.849**	**0.012**	**0.023**	**0.029**

Table 2: Quantitative evaluation of different layout generation methods on original inputs.

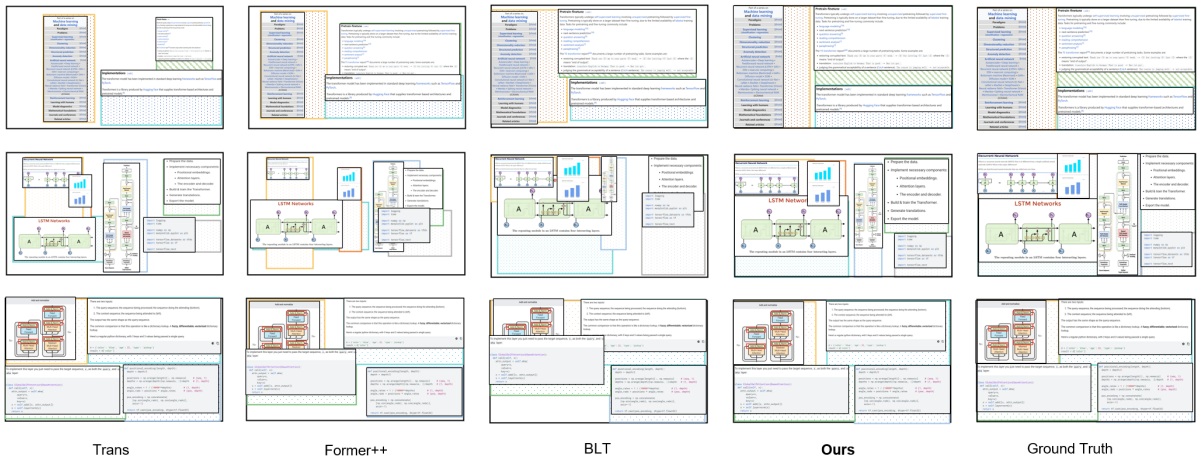

Figure 5: Qualitative comparison on layouts generated by data-driven baselines and our method, where the original input images are illustrated on the ground truth layouts. It is obvious that layouts generated by ours are most closely to ground truth ones, with the best alignment and most minimal overlapping compared to the baselines.

Method	Canvas	Number of input images			
		3	5	7	9
Softcollage	(128, 128)	61	123	301	488
Semi	(4,4)	8	28	35	60
Ours	(1920, 1080)	$< \mathbf{2}$			

Table 3: Comparison of running time (in seconds) on different methods.

Implementation

Each transformer block has 4 layers with a hidden size of 512, 4-head attention, and feed-forward dim of 1028. The CollageFormer is trained for 100 epochs. The Adam optimizers are used, and the initial learning rate is 10^{-3} for both generator and discriminator. All experiments are carried out with Pytorch framework and NVIDIA 3080 Ti GPUs.

Comparison with Data-driven Methods

We compare our CollageFormer with the following methods. Trans (Gupta et al. 2021) depends on a transformer-based encoder-decoder framework to predict layouts autoregressively. BLT (Kong et al. 2022) (BLT) is the first attempt to depend on only a transformer encoder to predict a layout for all elements in parallel. Former++ (Jiang et al. 2023) is the state-of-the-art method introducing special tokens into transformers for layout generation.

Table. 2 illustrates the metrics calculated on generated layouts. It is obvious that our model outperforms the rest methods overall metrics, where higher Miou and lower Over, Align, and Empty indicate better performances. The great improvements validate the effectiveness of our cascade predicting framework and the additional discriminator. To compare the layout generation quality of ours and baseline methods, Fig. 5 displays generated layouts randomly sampled from the model outputs, where the corresponding input screenshots are illustrated in the ground truth layouts. Obviously, our layouts are the most similar to ground truth ones, which are more alignment and tidy. Notably, our CollageFormer can handle a larger number of input screenshots with diverse aspect ratios and sizes.

Comparison with Optimization-based Methods

We compare our CollageNoter with the Softcollage (Yu et al. 2022) which depends on the gradient-based optimization to search for a reasonable layout, and Semi-Automatic Layout Adaptation(Semi) (Zeng et al. 2023) which depends on the simulated annealing algorithm to search for solutions.

First, as shown in Table. 3, our data-driven method significantly outperforms these optimization-based methods in terms of running time. The inference time for our neural network model is notably shorter, as it only involves the inference phase, whereas the optimization-based methods require extensive time to search for a suitable layout solution.

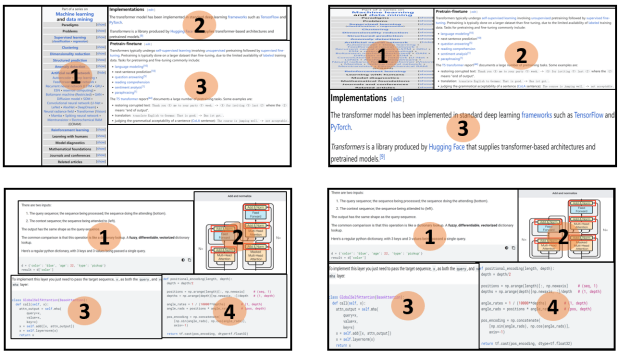

Figure 6: Qualitative comparison on layouts generated from our method and optimization-based method.

Method	HD			
	Miou↑	Over↓	Align↓	Empty↓
Ours-cascade	0.847	0.009	0.019	0.027
Ours-dis	0.839	0.010	0.023	0.031
Ours(full)	0.873	0.007	0.015	0.026
Method	**UHD**			
	Miou↑	Over↓	Align↓	Empty↓
Ours-cascade	0.706	0.033	0.041	0.042
Ours-dis	0.689	0.035	0.049	0.046
Ours(full)	0.723	0.031	0.036	0.041

Table 4: Ablation study on CollageFormer.

Second, Fig. 6 illustrates the generated E-Notes using both the relevance-first strategy and our method. It is evident that the relevance-first approach leads to more visual distortions due to its fixed relevance-driven constraints, while our method effectively preserves the original forms of the images, maintaining visual consistency.

Ablation Study

As our method includes a cascade generator and an additional discriminator, we study the following ablations. Ours-cascade indicates only one transformer-based layout generator with the additional discriminator. Ours-dis represents the CollageFormer with only the reconstruction loss. As shown in Table. 4, comparing Ours-cascade and Ours-dis, the latter performs better, highlighting the importance of the cascade framework in enhancing layout generation quality. In addition, both variations generate layouts better than baselines, demonstrating the effectiveness of these components.

User Study

To assess the real-world usability of our CollageNoter compared to baseline methods, we conduct two user studies with inviting 12 participants aged from 20 to 45 (50% female) to join our tests. ***First***, we collect 6 groups of screenshots and display the E-Notes generated by different methods to participants, respectively. Each participant is required to select the most visually appealing one, which requires the least efforts to read and understand. As shown in Fig. 7(a), 83.3% participants preferred the layouts generated by our method. ***Second***, we collect 4 groups of input screenshots and display the generated results with an additional question for each collage design, which is related to the last screenshot in the logical order simulating the most challenging information searching scenario. We record the time each participant takes to answer the questions based on corresponding generated E-Notes. As shown in Fig. 7(b), with more input screenshots, the difficulty of finding the answer is also increasing, where E-Notes designed by our CollageNoter obviously help users find the answer more quickly.

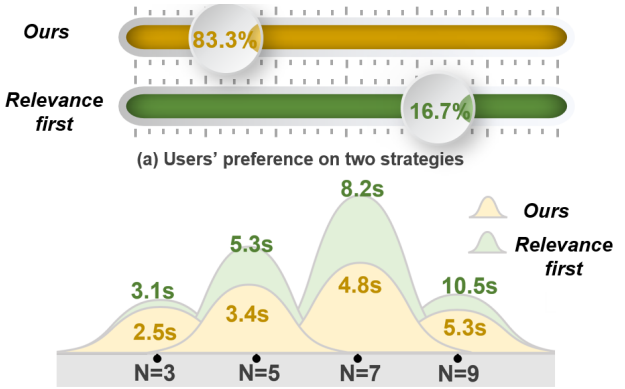

Figure 7: Illustration of user study.

Conclusion and Discussion

In this work, we focus on screenshot layout design for E-Note-taking and propose a novel readability-then-cognitive coherence strategy to ensure that the generated notes are both visually and cognitively easy to understand. Considering that optimization-based methods are time-consuming, we propose a novel data-driven generator named CollageFormer responding to different inputs in real time, with a novel cascade framework and an additional discriminator. Specifically, we develop a dataset builder to provide training data for any devices, enabling adaptive model training. Given users may struggle with extensive information, a potential improvement for CollageNoter would be to partition the numerous input screenshots into multiple sets, for creating an E-note document with several pages.

Acknowledgments

This work was partially supported by the National Science Fund for Distinguished Young Scholars (No.62025205), National Natural Science Foundation of China (No.62432007, No.62272390, No.62302356), and Guangdong Basic and Applied Basic Research Foundation (No.2022A1515110740).

References

Dayama, N. R.; Todi, K.; Saarelainen, T.; and Oulasvirta, A. 2020. Grids: Interactive layout design with integer programming. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.

Deka, B.; Huang, Z.; Franzen, C.; Hibschman, J.; Afergan, D.; Li, Y.; Nichols, J.; and Kumar, R. 2017. Rico: A mobile app dataset for building data-driven design applications. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, 845–854.

Fan, J. 2012. Photo layout with a fast evaluation method and genetic algorithm. In *2012 IEEE International Conference on Multimedia and Expo Workshops*, 308–313. IEEE.

Gan, Y.; Zhang, Y.; Sun, Z.; and Zhang, H. 2020. Qualitative photo collage by quartet analysis and active learning. *Computers & Graphics*, 88: 35–44.

Geigel, J.; and Loui, A. C. 2000. Automatic page layout using genetic algorithms for electronic albuming. In *Internet Imaging II*, volume 4311, 79–90. SPIE.

Gupta, K.; Lazarow, J.; Achille, A.; Davis, L. S.; Mahadevan, V.; and Shrivastava, A. 2021. Layouttransformer: Layout generation and completion with self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1004–1014.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Jiang, Z.; Guo, J.; Sun, S.; Deng, H.; Wu, Z.; Mijovic, V.; Yang, Z. J.; Lou, J.-G.; and Zhang, D. 2023. Layoutformer++: Conditional graphic layout generation via constraint serialization and decoding space restriction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18403–18412.

Jyothi, A. A.; Durand, T.; He, J.; Sigal, L.; and Mori, G. 2019. Layoutvae: Stochastic scene layout generation from a label set. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9895–9904.

Kong, X.; Jiang, L.; Chang, H.; Zhang, H.; Hao, Y.; Gong, H.; and Essa, I. 2022. BLT: bidirectional layout transformer for controllable layout generation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, 474–490. Springer.

Li, J.; Yang, J.; Hertzmann, A.; Zhang, J.; and Xu, T. 2018. LayoutGAN: Generating Graphic Layouts with Wireframe Discriminators. In *International Conference on Learning Representations*.

Liang, Y.; Wang, X.; Zhang, S.-H.; Hu, S.-M.; and Liu, S. 2017. PhotoRecomposer: Interactive photo recomposition by cropping. *IEEE transactions on visualization and computer graphics*, 24(10): 2728–2742.

Liu, T.; Wang, J.; Sun, J.; Zheng, N.; Tang, X.; and Shum, H.-Y. 2009. Picture collage. *IEEE Transactions on Multimedia*, 11(7): 1225–1239.

Mann, D. 2020. *Gestalt therapy: 100 key points and techniques*. Routledge.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Moen, D. R.; and Fee, F. 2000. Newspaper layout & design: A team approach. *Newspaper Research Journal*, 21(4): 113.

Pan, X.; Tang, F.; Dong, W.; Ma, C.; Meng, Y.; Huang, F.; Lee, T.-Y.; and Xu, C. 2019. Content-based visual summarization for image collections. *IEEE transactions on visualization and computer graphics*, 27(4): 2298–2312.

Perls, F.; Hefferline, G.; and Goodman, P. 1951. Gestalt therapy. *New York*, 64(7): 19–313.

Qiao, X.; Cao, Y.; and Lau, R. W. 2022. Design Order Guided Visual Note Layout Optimization. *IEEE Transactions on Visualization and Computer Graphics*, 29(9): 3922–3936.

Rother, C.; Bordeaux, L.; Hamadi, Y.; and Blake, A. 2006. Autocollage. *ACM transactions on graphics (TOG)*, 25(3): 847–852.

Tang, H.; Zhang, Z.; Shi, H.; Li, B.; Shao, L.; Sebe, N.; Timofte, R.; and Van Gool, L. 2023. Graph Transformer GANs for Graph-Constrained House Generation. In *CVPR*.

Wei, Y.; Matsushita, Y.; and Yang, Y. 2009. Efficient optimization of photo collage. *Microsoft Research, Redmond, WA, USA, MSRTR-2009-59*.

Wu, Z.; and Aizawa, K. 2016. Very fast generation of content-preserved photo collage under canvas size constraint. *Multimedia Tools and Applications*, 75: 1813–1841.

Yang, C.-F.; Fan, W.-C.; Yang, F.-E.; and Wang, Y.-C. F. 2021. LayoutTransformer: Scene Layout Generation With Conceptual and Spatial Diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3732–3741.

Yu, J.; Chen, L.; Zhang, M.; and Li, M. 2022. SoftCollage: A Differentiable Probabilistic Tree Generator for Image Collage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3729–3738.

Zeng, W.; Chen, X.; Hou, Y.; Shao, L.; Chu, Z.; and Chang, R. 2023. Semi-automatic layout adaptation for responsive multiple-view visualization design. *IEEE Transactions on Visualization and Computer Graphics*.

Zheng, X.; Qiao, X.; Cao, Y.; and Lau, R. W. 2019. Content-aware generative modeling of graphic design layouts. *ACM Transactions on Graphics (TOG)*, 38(4): 1–15.

Zhong, X.; Tang, J.; and Yepes, A. J. 2019. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1015–1022. IEEE.

Zhou, M.; Xu, C.; Ma, Y.; Ge, T.; Jiang, Y.; and Xu, W. 2022. Composition-aware Graphic Layout GAN for Visual-textual Presentation Designs. *arXiv preprint arXiv:2205.00303*.