# Human-aware Design Generation: Adding 3D Humans into Graphic Designs

Zijin Hou
ShanghaiTech University
Shanghai, China
houzj2025@shanghaitech.edu.cn

Ying Cao*
ShanghaiTech University
Shanghai, China
caoying59@gmail.com

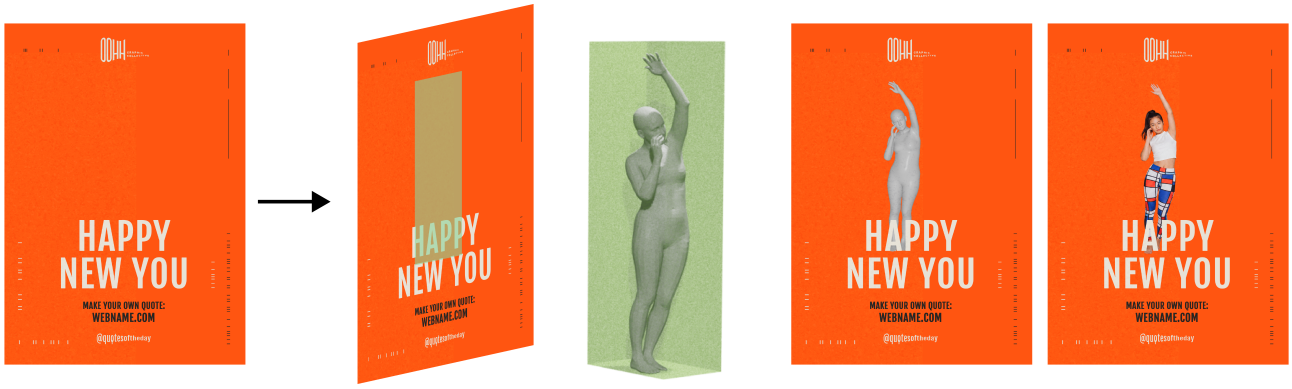



**Figure 1: Design-conditioned 3D human adding. Given an input design (1st column) with a set of elements, such as images, vector shapes and texts, our model adds a 3D human into the design (2nd column), by predicting its 3D pose and 2D framing, as well as the layout of the image containing the human in the input design (the 2D box on the design). The model results in a cohesive output design with the 3D human added (3th column). The 3D human representation can further be rendered into a realistic human image (4th column).**

## Abstract

Graphic designs nowadays predominantly contain human images, where people are posed and framed in a delicate way to guide viewers' visual exploration and evoke desired feelings. Despite recent progress in automatic graphic design generation, there has been no work on investigating the role of the human representations in design creation. In this paper, we seek to emphasize the importance of visual human representations in graphic design generation, by studying a novel task, dubbed as design-conditioned 3D human adding. To solve this task, we propose a model that adds 3D human representations into an existing partial design, by predicting their 3D poses, their 2D composition in the image containing them, as well as the spatial arrangement of the image in the input design, to generate a complete design. Our experiments show that our model outperforms baselines by predicting the 3D human poses, 2D framing and human image layout that better harmonize with the other existing design elements, and generating higher-quality holistic designs. Our work is the first to explore connections between 3D human generation and 2D graphic design generation, which would inspire future research in leveraging 3D information to build more powerful generative models for graphic designs in the 2D domain.

## CCS Concepts

• **Computing methodologies; Computer graphics**;

## Keywords

Graphic Design, Design Generation, Generative Model

**ACM Reference Format:**
Zijin Hou and Ying Cao. 2026. Human-aware Design Generation: Adding 3D Humans into Graphic Designs. In *Proceedings of Proceedings of the 34th ACM International Conference on Multimedia (MM '26).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3836019

## 1 Introduction

The use of human images is prevalent in graphic designs such as posters, magazines and advertisements. What poses people in an image take and how they are framed in the image play critical roles in affecting how viewers visually explore the entire design and what feelings are evoked. Consider the design in Figure 2 as an example. When looking at the design, one may first attend to the face of the woman, and then move the eyes to her hand near the face, Finally, the visual attention shifts to the left along the direction pointed by the finger, reaching the potted plant. This means that the pose of

*Corresponding author.

This work is licensed under a Creative Commons Attribution 4.0 International License.
*MM '26, Rio de Janeiro, Brazil.*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3836019




**Figure 2: An example graphic design with a human image (left) and the design with the 3D representation of the human (right).**

the woman can effectively direct one's visual flow throughout the design. Furthermore, the woman is framed in a medium full shot that starts from the top of the head and ends just below the waist, which conveys the feelings of "strong" and "confident". However, creating a human image with ideal posing and framing for a design is particularly challenging, requiring extensive design knowledge and experience, along with some creativity.

In recent years, there has been a growing interest in developing generative models for graphic design with some encouraging results. However, the role of visual human representations has not yet been considered explicitly, which hampers the generation of designs where human images are employed effectively.

In view of the issue above, in this paper, we take an initial step towards *human-aware design generation*, by investigating a new task, *design-conditioned 3D human adding*. As shown in Figure 1, given an input partial design without a human image, the task aims to add 3D human representations into the input design to form a cohesive complete design, where the added human representations and the other elements are harmonious with each other. The generated 3D representations can further be rendered into a realistic human image. The purpose of generating 3D representations explicitly, instead of only 2D images, is to offer great controllability, which allows users to easily change the poses and composition of people in the rendered image through manipulating the generated 3D representations directly. Such controllability is imperative for the practical deployment of design generative models in the design workflow, since designers wish to have a higher degree of control over the generation process when working with the models.

We propose *POFRAY*, **Po**se-**Fr**ame-L**ay**out, a model to solve this new task. POFRAY uses a partial layered design with a set of multimodal elements as a context, and add 3D humans into it by predicting what *poses* they should take in 3D space, how they are *framed* within the 2D image containing them, as well as how the image should be *laid out* in the design. Specifically, POFRAY employs a multimodal encoder to process the input elements to extract design context features. The features are then used to condition two separate generators: a pose and framing generator, a layout generator. The pose and framing generator employs a causal Transformer to autoregressively generate a sequence of tokens that control the 3D poses of the added humans and their 2D framing in the image. The layout generator uses a bidirectional Transformer to generate the layout of the added human image in the input design. To train and evaluate our model, we construct *3DHumanInDesign*, a benchmark of ~10k layered graphic designs. For each design, we remove the human image to form the input, and use the estimated human pose and framing parameters from the human image alongside the original layout of the the human image as the prediction targets.

Experimental results on our collected benchmark demonstrate that our model outperforms alternative baselines, generating higher-quality complete designs. The generated human poses, framing and human image layout of our model are better aligned with the input design context, compared to those of the baselines.

Our major contributions are threefold:

- We formulate a new problem in the graphic design generation field: design-conditioned 3D human adding. The problem can be viewed as a specific form of human-aware design generation, where visual human representations need to be modeled explicitly. To the best of our knowledge, this is the first time that the two fields, 3D human generation and 2D graphic design generation, are bridged. We hope that our work can inspire future work on leveraging various 3D information for graphic design synthesis, which was previously addressed with 2D-only methods.
- We propose a model, POFRAY, which adds 3D humans into an input design by simultaneously generating 3D human poses, the 2D framing of the humans in an image, and the layout of the human image in the design.
- We build a benchmark, 3DHumanInDesign, to facilitate future research on design-conditioned 3D human adding.

## 2 Relate Work

### 2.1 Graphic Design Generation

Automatic graphic design generation has emerged as a crucial research area aimed at facilitating the design process. CanvasVAE [33] introduces a variational autoencoder to generate vector graphic designs represented as sets of canvas and element attributes. GOL [34] learns a design element ordering strategy to optimize the performance of design generative models. In recent years, a large body of methods on text-to-design generation has emerged to convert textual design intentions into multi-layer graphic designs, e.g., by building a cascaded system comprising multiple specialized models for different sub-tasks (e.g., background generation, typography generation) [13, 14], formulating layered design generation as multi-layer transparent image generation [25], utilizing a multimodal large language model to generate graphic designs as text-image documents [26], and training a simple diffusion model on sequences of discrete element attribute tokens [31].There is also a line of work on automatic design composition, which aims to compose a set of input design elements into a complete design [4, 22, 30]. FlexDM [12] is a unified model that can solve multiple design tasks (e.g., image filling, layout generation, typography generation), by representing a graphic design as a set of elements, each comprising a set of fields, and formulating different design tasks as masked field prediction problems with different masking patterns.

Recent years have also seen a growing interest in developing models for generating layouts of graphic design elements [1, 16, 19, 21, 29, 36, 39]. The success of Transformers in natural language processing has inspired their application to layout generation. LayoutTransformer [10] formulates layout generation as a sequence modeling problem and employs the self-attention mechanism to capture element relationships. BLT [17] proposes a bidirectional Transformer that generates all layout tokens in parallel to enable flexible controllable generation. Our method leverages Transformers to predict 3D human pose, 2D framing, and human image layout.

To our best knowledge, we are the first to model 3D human representations in 2D graphic designs, which has not been yet explored in previou studies.

## 2.2 Human Pose Estimation and Generation

3D Human pose and shape estimation have been extensively studied in computer vision, with a range of applications in behavior analysis, animation and gaming. Recent advances in expressive human pose and shape estimation (EHPS) enable the reconstruction of 3D whole-body meshes of a person—including the body, hands and face—from monocular images or videos using end-to-end models [3, 5, 32] that regress the parameters of a 3D parametric human model (e.g., SMPL-X [23].

Conditional 3D human pose generation based on input conditions of various modalities has recently gained increasing attention. For example, ChatPose [7] generates 3D human poses from natural language descriptions, optionally with image inputs. GenZI [20] proposes a method to synthesize a posed 3D human that interacts with a given 3D scene according to an input text description. A recent study [15] focuses on dynamic human-scene interaction modeling, generating 3D human motions (sequences of 3D poses) that are plausible in a given 3D scene context. NIFTY [18] addresses the problem of object-conditioned human motion synthesis, which generates realistic 3D motions of a person interacting with a given 3D object.

In this work, we focus on generating 3D human poses in the context of graphic designs.

# 3 Method

## 3.1 Problem Formulation

Let $\mathcal{D} = \{e_1, e_2, \ldots, e_N\}$ denote an input graphic design consisting of $N$ design elements, where each element $e_i$ is described by a set of attributes: appearance attributes including font, color (font color or background color), opacity; content (image or text content); layout attributes $(c_i, l_i, t_i, w_i, h_i)$ specifying its type $c_i$, left position $l_i$, top position $t_i$, width $w_i$, and height $h_i$, respectively. $(l_i, t_i, w_i, h_i)$ are normalized bounding box parameters (with respect to the canvas) in the range $[0, 1]$. Our objective is to build a model that takes as input a design $\mathcal{D}$, without a human image, and produces the following outputs:

- **Pose**. A sequence of $M$ 3D human poses $P = \{\theta_1, \ldots, \theta_M\}$, where each pose is parameterized by the pose parameters $\theta_i \in \mathbb{R}^{52\times3}$ of the SMPL-H model [28], which specify 3D angular rotations for 52 joints (21 body joints, 30 hand joints, 1 root joint).
- **Framing**. Framing parameters $F = \{f_1, \ldots, f_M\}$ for $m$ 3D humans. The framing of human $j$ is parameterized by $f_j = (x_j, y_j, z_j)$, which is the global 3D position of the root joint (relative to the camera frame), which determines how the human is framed in the 2D image.
- **Layout.** A human image layout $L = (l, t, w, h, z)$ that specifies the left position $l$, top position $t$, width $w$, height $h$, and z-order $z$ of the image containing the 3D humans in the design.

The generated poses $P$ are rendered into an image using the predicted framing parameters $F$, and placed into the input design according to the predicted image layout $L$, to create a cohesive human-included graphic design.

## 3.2 Our Model

As illustrated in Figure 3, our model consists of three main components: 1) a design encoder that extracts design context representations from the input design, 2) a pose and framing generator that autoregressively generates human pose tokens and framing tokens , and 3) a layout generator that predicts the spatial arrangement of the human image. We now describe each component in detail.

*3.2.1 Design Encoder.* The design encoder follows the architecture of FlexDM [12], projecting the elements in the input design into a set of element-wise embeddings. In particular, for each element, the design encoder extracts three types of embeddings from its attributes:

- *Appearance embeddings.* We perform discretization on the font, color and opacity, and encode discrete font, color and opacity into separate embeddings through learned embedding layers. When an attribute is missing for an element (e.g., an image element does not have font color), an all-zeros embedding is used.
- *Content embeddings.* The visual content is processed through the OpenCLIP [11] image encoder to obtain an image embedding, while the text content is encoded using OpenCLIP text encoder to produce a text embedding. Then, the image or text embedding is projected into a content embedding through a learned linear project layer.
- *Layout embeddings.* Each of the four bounding box parameters is discretized. The element type and the discrete bounding box parameters are encoded into layout embeddings through learned embedding layers.

The embedding of the element is obtained by adding up its attribute embeddings. All the element embeddings are then processed through a stack of transformer encoder blocks with self-attention layers to capture inter-element relationships. The encoder finally outputs a sequence of contextualized design element embeddings, which we refer to as *a design context representation* $\mathbf{C}$ that will be used to condition subsequent generators on information in the input partial design.

*3.2.2 Pose and Framing Generator.* To generate human poses, we adopt a discrete pose representation inspired by TokenHMR [6]. We first train a Vector Quantized Variational Autoencoder (VQ-VAE) tokenizer specifically for the SMPL-H pose parameters. The VQ-VAE encoder maps continuous SMPL-H pose parameters $\theta \in \mathbb{R}^{52\times3}$

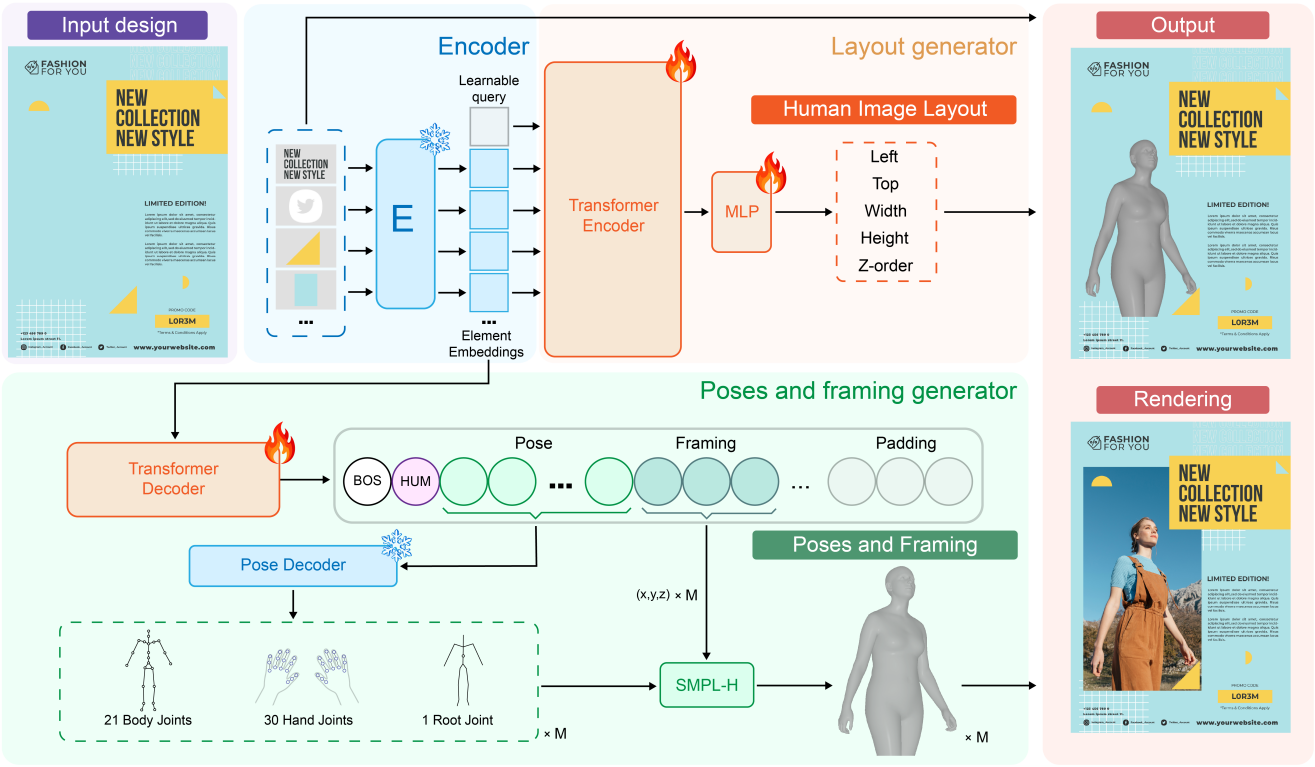



**Figure 3: Overview of our model. It takes as input a partial graphic design with a set of multimodal elements (such as images, texts, vector shapes), and generates a complete design with one or several 3D humans added. An encoder is used first used to extract element-wise design embeddings, which are then fed into: a pose and framing generator, a layout generator. The pose and framing generator autoregressively generates pose tokens, which are then decoded into 3D human poses using a pre-trained pose decoder. It also predicts framing tokens that control the composition of the 3D humans in an image. The layout generator takes as input a learnable query and all the element embeddings to predict the bounding box parameters and z-order of the human image in the design. With the predicted pose, framing and layout information, the 3D humans are composed into the design, and the 3D human representations can further be rendered as a realistic image.**

to a discrete token sequence $\mathbf{z}_p = (z_1, z_2, \ldots, z_{400})$, where each token $z_i$ corresponds to a code vector in a learned codebook of size 64. The VQ-VAE decoder can reconstruct the pose parameters from the tokens: $\hat{\theta} = \mathcal{D}_{\text{pose}}(\mathbf{z}_p)$. For framing parameters $f = (x, y, z)$, we discretize each coordinate into 256 bins by applying the K-means clustering on the training set, resulting in a sequence of framing tokens $\mathbf{z}_f = (z_x, z_y, z_z)$.

For $m$ humans, their poses and framing configurations can then be represented as a sequence of discrete tokens:

$$\mathbf{Z} = (\text{BOS}, \text{HUM}, \mathbf{z}_{p_1}, \mathbf{z}_{f_1}, \ldots, \mathbf{z}_{p_m}, \mathbf{z}_{f_m}, \text{EOS}) \tag{1}$$

where BOS and EOS are special tokens denoting the start and end of the sequence, and the HUM token is used to signals the start of a human.

Our pose decoder employs an autoregressive Transformer architecture trained to generate tokens in the sequence $\mathbf{Z}$ iteratively conditioned on the design context representation $\mathbf{C}$ from the design encoder. The design context representation is injected into the pose decoder through cross-attention layers. The generated pose token sequence $\hat{\mathbf{z}}_p$ is decoded back to the SMPL-H pose parameters using the pre-trained VQ-VAE decoder $\mathcal{D}_{\text{pose}}$, and the generated framing tokens are converted to continuous framing parameters by taking the bin center values of the tokens.

*3.2.3 Layout Generator.* The layout decoder predicts the spatial arrangement of the human image within the design. Our layout decoder utilizes a bidirectional Transformer architecture (with a stack of Transformer encoder blocks), whose inputs include a learnable query (a learnable embedding) and the design context representation $\mathbf{C}$ (a sequence of element embeddings). The query attends to all the element embeddings through the self-attention layers to extract an output query representation (the output embedding of the last Transformer block for the query) that contains necessary information for layout prediction. The output query representation is finally decoded into the layout parameters $(\hat{l}, \hat{t}, \hat{w}, \hat{h}, \hat{z})$ with a two-layer MLP.

*3.2.4 Human Image Generation and Composition.* Given the generated pose, framing and layout parameters, we can create a human

image and compose it into the input design. Specifically, we feed the pose and framing parameters into the SMPL-H model to generate posed 3D human meshes placed in the camera coordinate system, and project the meshes onto an image of size $(W \cdot \hat{w}, H \cdot \hat{h})$, where $(W, H)$ is the canvas size of the input design and $(\hat{w}, \hat{h})$ is the predicted human image size. This results in a transparent human mesh image (where the alpha values of pixels outside the human regions are set to 0), which can be composed into the input design according to the predicted position and z-order $(\hat{l}, \hat{t}, \hat{z})$. To generate a realistic human image, we create a 2D pose map using the 2D projections of the 3D joint positions of the human meshes, and employ LayerDiffuse [37] (with SDXL [24] to generate a transparent image conditioned on the pose map, the input design, and a text prompt. The text prompt is generated by Qwen3-VL [2] that is queried to describe the human appearance according to the input design. The pose map is incorporated into the SDXL through ControlNet [38]; the input design is sent into the SDXL through IP-Adapter [35].

### 3.3 Training

Our model is trained with two task-specific loss functions:

- *Pose and framing loss.* The pose and framing generator is trained with a cross-entropy loss for next-token prediction. For a sequence $\mathbf{Z}$ of length $L$. The loss is defined as:

$$\mathcal{L}_{\text{PF}} = -\sum_{t=2}^{L} \log p_\phi(\mathbf{Z}_t | \mathbf{Z}_{<t}, \mathbf{C}). \tag{2}$$

- *Layout loss.* The layout generator is supervised with a combination of a L1 loss and a Generalized IoU (GIoU) loss:

$$\mathcal{L}_{\text{L}} = \mathcal{L}_{L1} + \mathcal{L}_{\text{GIoU}}. \tag{3}$$

  $\mathcal{L}_{L_1} = \left|\hat{L} - L\right|$, where $\hat{L}$ and $L$ are the predicted and ground truth human image layout parameters, respectively. $\mathcal{L}_{\text{GIoU}} = 1 - \text{GIoU}(\hat{B}_h, B_h)$, where $\hat{B}$ and $B$ are the predicted and ground truth bounding boxes of the human image, respectively.

During training, we first fine-tune FlexDM on the training component of our dataset (Section 4.1) and use the obtained corresponding weights to initialize our design encoder. Then, we freeze the design encoder weights, and train our model via a three-stage training procedure:

- *Stage 1: Pose and framing generator pre-training.* We first train the pose and framing generator using the loss $\mathcal{L}_{\text{PF}}$. This stage allows the pose and framing generator to capture the dependency of human pose and framing on design context.
- *Stage 2: Layout generator pre-training.* We train the layout generator using the loss $\mathcal{L}_{\text{L}}$. This stage enables the layout generator to learn spatial placement of human images based on design context features.
- *Stage 3: Joint Fine-tuning.* Finally, we jointly fine-tune the two generators, with a total loss:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{PF}} + \mathcal{L}_{\text{L}}. \tag{4}$$

  This joint fine-tuning stage allows the two generators to produce the pose, framing and layout that are harmonious with each other.

## 4 Experiments

### 4.1 Dataset

Our *3DHumanInDesign* dataset consists of ∼10k *layered* graphic designs with human images, carefully curated from two sources to ensure diversity and quality. Specifically, we collect 4,479 samples from the *Crello* dataset [33], which covers different kinds of graphic designs such as web banners, social media posts and posters. Additionally, we gather 5,460 magazine covers and posters from *Freepik*[1]. For the human image on each design, we extract SMPL-H pose parameters and 3D root joint positions (framing parameters) using a previous human pose and shape estimation method [32]. We remove human images from the original designs to create input designs (without human images). We partition our dataset into training and test splits with a ratio of 8:2, while ensuring that each split has 40% and 60% of its samples from Crello and Freepik, respectively.

### 4.2 Implementation Details

For all the three training stages, we optimize our model using AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.1) with a base learning rate of $1 \times 10^{-4}$, using 1,000 warm-up iterations followed by a cosine decay schedule. We use a batch size of 16, a gradient-norm clip of 1.0, a label smoothing of 0.05, and maintain an exponential moving average of the weights with a decay of 0.995.At inference time, We employ top-$p$ sampling with temperature 0.8 and $p = 0.8$.

### 4.3 Baselines

Since there is no prior work on design-conditioned 3D human adding, we compare our method against two baselines:

- *FlexDM*. FlexDM [12] is a unified model for solving different design tasks. We fine-tune FlexDM on our training set, and adapt it to our problem by empolying a two-stage approach. In the first stage, we utilize FlexDM to solve an image filling task to produce an image embedding, by feeding all the elements in the input design, along with an added image element (with all the fields masked out except its category), into FlexDM. The output image embedding is used to retrieve an existing image from a database of realistic human images (taken from the designs in the test set) via cosine similarity between image embeddings. Then, the pose and framing parameters are estimated from the retrieve image using an off-the-shelf human pose and shape estimation method [32]. In the second stage, we use FlexDM to perform layout generation for the retrieved image. Since FlexDM can not predict the z-order of an element, for a fair comparison, we use the z-order generated by our model when composing the human image into the input design.
- *MLLM.* We design a baseline leveraging multimodal large language models (MLLM) for design-conditioned 3D human adding. Specifically, we provide the input design image along with its layout information (in textual format) to Gemini 3 Pro [8], and prompt it to predict: 1) a text description for a human image; 2) the bounding box parameters and z-order of the human image in the design. The image description is then fed into a text-to-image model, Nano Banana Pro [9],

[1] https://www.freepik.com

**Table 1: Quantitative evaluation results of different methods. The best results are in bold, while the second best results are underlined.**

| Method | HD-Align ↑ | D-FID ↓ | MPJPE ↓ | Align ↓ | Overlap ↓ |
|---|---|---|---|---|---|
| FlexDM | 4.65 | <u>3.90</u> | <u>2176.29</u> | <u>0.17</u> | 213.39 |
| MLLM | <u>5.26</u> | 7.95 | 2829.97 | 0.20 | **196.03** |
| Ours | **6.52** | **0.07** | **385.12** | **0.17** | <u>209.11</u> |
| Real Data | 7.06 | – | – | 0.16 | 210.48 |

to generate a realistic human image. Finally, the pose and framing parameters are extracted from the generated image using an off-the-shelf human pose and shape estimation method [32].

## 4.4 Evaluation Metrics

We quantitatively evaluate the performance of our method using four metrics that cover complementary aspects: human-design coherence, overall design quality, pose accuracy, and layout quality.

- *Human-Design Alignment (HD-Align).* Inspired by CLIP [27], we train a contrastive model to learn a joint embedding space of human representations and partial designs (with human images excluded), where matched human and design lie close and unmatched ones fall far apart. Our human representation combines the 3D pose, framing, and human image layout information. We compute cosine similarity between human and design embeddings as a HD-Align score, with higher HD-Align indicating better human-design coherence.
- *Design FID (D-FID).* We train a variational autoencoder that encodes both a human representation (combining 3D pose, framing, and human image layout information) and a partial design (without any human image) into a holistic design embedding. We use these embeddings to compute Fréchet Inception Distance (FID) between generated and real designs as a measure of overall design quality. Lower D-FID indicates better holistic design quality.
- *MPJPE.* We adopt Mean Per Joint Position Error (MPJPE), a common metric for 3D human pose estimation, to evaluate predicted 3D poses, which computes the mean Euclidean distance between predicted and ground-truth 3D joint positions.
- *Alignment, Overlap.* We evaluate the quality of predicted human image layouts (bounding boxes) with two widely used metrics in the layout generation literature, *Alignment* and *Overlap* [19]. Both metrics are computed on the layouts of complete designs with human images included.

## 4.5 Quantitative Results

Table 1 reports the quantitative results of different methods on the test set. The HD-Align score of our method is better than those of the baselines by significant margins, while being competitive with that of the read data (samples in the test set). This suggest great alignment of our generated human pose, framing and human image layout with the design context. Our method also shows a noticeable improvement in D-FID compared to the other methods, highlighting the superior quality of holistic designs (with added human images) generated by our method. Our method significantly outperforms the baselines in MPJPE, confirming that its generated poses are closer to the ground truth. In addition, our method achieves the best alignment score and the second best overlap score, which means that our method can result in spatially cohesive design layouts after adding human images into the partial designs. Although MLLM yields the lowest overlap score, the overlap score of our method (209.11) is quite close to that of the real data (210.48), which suggests that our resulting layouts encompass a similar amount of overlap between elements to real designs.

**Table 2: User study results. For each aspect, the percentage of the time that each method is chosen as the best is reported.**

| Method | Human-Design Align | Aesthetic Quality |
|---|---|---|
| FlexDM | 11.9% | 17.8% |
| MLLM | 27.0% | 22.5% |
| Ours | **61.1%** | **59.7%** |

## 4.6 Qualitative Results

Figure 4 shows visual comparison of results from different methods. As compared with the baselines, our method can generate 3D human poses and 2D framing that better harmonizes with the design context and more closely resemble what professional designers create. For example, in the first row, our method add multiple humans whose *body orientations* are similar to those of humans in the ground truth design, whereas FlexDM and MLLM generates a single human facing outwards. In the third row, our method frames the added human in the same way as the real design—a *medium full shot* is used, capturing the human from the top of the head to just below the waist, while FlexDM and MLLM employ a *medium shot* that starts just above the head and ends around the waist. In addition, our method is able to generate human image layouts that are spatially compatible with the remaining design elements. MLLM sometimes falls shot of predicting proper bounding box coordinates and z-order for the added human image, causing visual conflict between the human image and other design elements; for example, a large portion of the added human is occluded by the texts in the fourth row.

## 4.7 User Study

We conduct a user study comparing our method with FlexDM and MLLM for *human-design alignment* and *aesthetic quality*. We apply each method to 100 input designs from our test set to generate complete designs (with realistic human images added). Our study involves 23 professional graphic designers. The participants are shown three generated complete designs by different methods in randomized order, along with the input design and ground truth design for reference, and asked to evaluate the generated designs in terms of two aspects: 1) human-design alignment (whether the added human image fits well the remaining part of the design in terms of human pose, framing and image layout; 2) aesthetic quality (whether the design looks visually aesthetic). For each aspect, the participants are asked to choose the best generated design. The results are shown in Table 2. Our method is significantly preferred over the other methods on both aspects, suggesting that our method




**Figure 4: Qualitative comparison of different methods.**

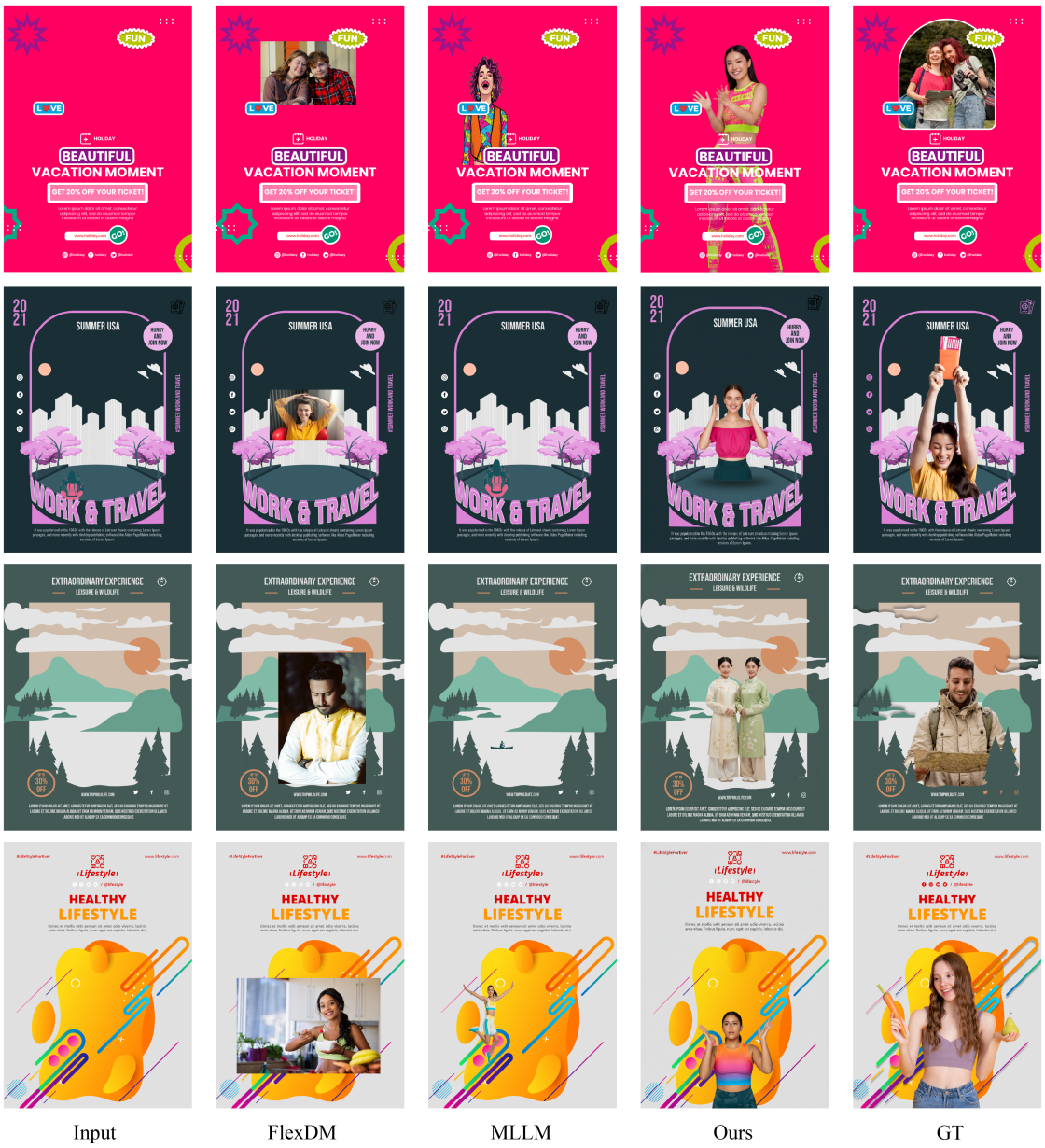



**Figure 5: Results of different methods where realistic human images are added.**

is able to achieve better human-design alignment and aesthetic quality. Figure 5 shows some results of different methods that are displayed in the user study.

## 5 Conclusion

In this work, we make the first step to bring 3D human modeling into the creation of 2D graphic designs, by formulating and investigating the problem of design-conditioned 3D human adding, where we aim to add 3D human representations into a partial input design. Our experimental results demonstrate the superiority of our proposed model in generating the 3D human poses, 2D framing and human image layout that better fit with the remaining elements in the input design, compared to several baselines. We believe that our effort in this work can inspire future work in building 3D-aware generative models of graphic design.

## References


[1] Diego Martín Arroyo, Janis Postels, and Federico Tombari. 2021. Variational Transformer Networks for Layout Generation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 13637–13647. https://api.semanticscholar.org/CorpusID:233033690

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025. Qwen3-VL Technical Report. arXiv:2511.21631 [cs.CV] https://arxiv.org/abs/2511.21631

[3] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. 2023. SMPLer-X: Scaling Up Expressive Human Pose and Shape Estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[4] Yutao Cheng, Zhao Zhang, Maoke Yang, Hui Nie, Chunyuan Li, Xinglong Wu, and Jie Shao. 2025. Graphic Design with Large Multimodal Model. In *AAAI Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:277752700

[5] Jing de Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Y. Li. 2023. One-Stage 3D Whole-Body Mesh Recovery with Component Aware Transformer. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 21159–21168. https://api.semanticscholar.org/CorpusID:257771455

[6] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J. Black. 2024. TokenHMR: Advancing Human Mesh Recovery with a Tokenized Pose Representation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), 1323–1333. https://api.semanticscholar.org/CorpusID:269362645

[7] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. 2024. ChatPose: Chatting about 3D Human Pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20178–20188.

[8] Gemini Team, Google. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805* (2023).

[9] Google DeepMind. 2025. Nano Banana Pro. Accessed: 2026-03-31. https://blog.google/innovation-and-ai/products/nano-banana-pro/

[10] Akash Gupta, Aditya Behl, Abhinav Dhall, Angela Yao, and Niloy J Mitra. 2021. LayoutTransformer: Scene Layout Generation with Conceptual and Spatial Diversity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 3732–3741.

[11] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. *OpenCLIP*. If you use this software, please cite it as below.. doi:10.5281/zenodo.5143773

[12] Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. 2023. Towards Flexible Multi-modal Document Models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 14287–14296. https://api.semanticscholar.org/CorpusID:257900939

[13] Naoto Inoue, Kento Masui, Wataru Shimoda, and Kota Yamaguchi. 2024. OpenCOLE: Towards Reproducible Automatic Graphic Design Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

[14] Peidong Jia, Chenxuan Li, Yuhui Yuan, Zeyu Liu, Yichao Shen, Bohan Chen, Xingru Chen, Yinglin Zheng, Dong Chen, Ji Li, Xiaodong Xie, Shanghang Zhang, and Baining Guo. 2024. COLE: A Hierarchical Generation Framework for Multi-Layered and Editable Graphic Design. arXiv:2311.16974 [cs.CV] https://arxiv.org/abs/2311.16974

[15] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. 2024. Scaling Up Dynamic Human-Scene Interaction Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1737–1747.

[16] Akash Abdu Jyothi, Thibaut Durand, Jiawei He, Leonid Sigal, and Greg Mori. 2019. LayoutVAE: Stochastic Scene Layout Generation From a Label Set. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 9895–9904.

[17] Xiang Kong, Lu Jiang, Huiwen Chang, Han Zhang, Yuan Hao, Haifeng Gong, and Irfan Essa. 2022. BLT: Bidirectional Layout Transformer for Controllable Layout Generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 474–490.

[18] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. 2024. NIFTY: Neural Object Interaction Fields for Guided Human Motion Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18062–18072.

[19] Jianan Li, Jimei Yang, Aaron Hertzmann, Jianming Zhang, and Tingfa Xu. 2020. LayoutGAN: Synthesizing Graphic Layouts With Vector-Wireframe Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43 (2020), 2388–2399. https://api.semanticscholar.org/CorpusID:209894294

[20] Lei Li and Angela Dai. 2024. GenZI: Zero-Shot 3D Human-Scene Interaction Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20465–20474.

[21] Jiawei Lin, Jiaqi Guo, Shizhao Sun, Zijiang James Yang, Jian-Guang Lou, and Dongmei Zhang. 2023. LayoutPrompter: awaken the design ability of large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 1902, 28 pages.

[22] Jiawei Lin, Shizhao Sun, Danqing Huang, Ting Liu, Ji Li, and Jiang Bian. 2024. From Elements to Design: A Layered Approach for Automatic Graphic Design Composition. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), 8128–8137. https://api.semanticscholar.org/CorpusID:280016385

[23] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body From a Single Image. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 10967–10977. https://api.semanticscholar.org/CorpusID:109932872

[24] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[25] Yifan Pu, Yiming Zhao, Zhicong Tang, Ruihong Yin, Haoxing Ye, Yuhui Yuan, Dong Chen, Jianmin Bao, Sirui Zhang, Yanbin Wang, Lin Liang, Lijuan Wang, Ji Li, Xiu Li, Zhouhui Lian, Gao Huang, and Baining Guo. 2025. ART: Anonymous Region Transformer for Variable Multi-Layer Transparent Image Generation. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), 7952–7962. https://api.semanticscholar.org/CorpusID:276580289

[26] Yadong Qu, Shancheng Fang, Yuxin Wang, Xiaorui Wang, Zhineng Chen, Hongtao Xie, and Yongdong Zhang. 2025. Igd: Instructional Graphic Design With Multimodal Layer Generatio. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)* (2025), 18218–18228. https://api.semanticscholar.org/CorpusID:280232207

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

[28] Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (Nov. 2017).

[29] Jaejung Seol, Seojun Kim, and Jaejun Yoo. 2024. PosterLlama: Bridging Design Ability ofnbsp;Language Model tonbsp;Content-Aware Layout Generation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXII* (Milan, Italy). Springer-Verlag, Berlin, Heidelberg, 451–468. doi:10.1007/978-3-031-73007-8_26

[30] Mohammad Amin Shabani, Zhaowen Wang, Difan Liu, Nanxuan Zhao, Jimei Yang, and Yasutaka Furukawa. 2024. Visual Layout Composer: Image-Vector Dual Diffusion Model for Design Layout Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9222–9231.

[31] Xinya Song, Bo Yang, and Ying Cao. 2026. Harnessing Layered Graphic Designs with Real Intentions for Text-to-Design Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4729–4738.

[32] Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, C.K.S. Leung, Ziwei Liu, Lei Yang, and Zhongang Cai. 2024. AiOS: All-in-One-Stage Expressive Human Pose and Shape Estimation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), 1834–1843. https://api.semanticscholar.org/CorpusID:268691528

[33] Kota Yamaguchi. 2021. CanvasVAE: Learning to Generate Vector Graphic Documents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5471–5480.

[34] Bo Yang and Ying Cao. 2025. Order Matters: Learning Element Ordering for Graphic Design Generation. *ACM Trans. Graph.* 44, 4, Article 34 (2025), 16 pages. doi:10.1145/3730858

[35] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. arXiv:2308.06721 [cs.CV] https://arxiv.org/abs/2308.06721

[36] Junyi Zhang, Jiaqi Guo, Shizhao Sun, Jian-Guang Lou, and D. Zhang. 2023. LayoutDiffusion: Improving Graphic Layout Generation by Discrete Diffusion Probabilistic Models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), 7192–7202. https://api.semanticscholar.org/CorpusID:257636725

[37] Lvmin Zhang and Maneesh Agrawala. 2024. Transparent Image Layer Diffusion using Latent Transparency. *ACM Transactions on Graphics (TOG)* 43 (2024), 1 – 15. https://api.semanticscholar.org/CorpusID:268032388

[38] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), 3813–3824. https://api.semanticscholar.org/CorpusID:256827727

[39] Xinru Zheng, Xiaotian Qiao, Ying Cao, and Rynson W. H. Lau. 2019. Content-aware generative modeling of graphic design layouts. *ACM Trans. Graph.* 38, 4, Article 133 (July 2019), 15 pages. doi:10.1145/3306346.3322971