WAFFLE: Multimodal Floorplan Understanding in the Wild

Keren Ganon*[1] Morris Alper*[1] Rachel Mikulinsky[1] Hadar Averbuch-Elor[1,2]
[1]Tel Aviv University [2]Cornell University

https://tau-vailab.github.io/WAFFLE

Abstract

Buildings are a central feature of human culture and require significant work to design, build, and maintain. As such, the fundamental element defining their structure – the floorplan *– has increasingly become an object of computational analysis. Existing works on automatic floorplan understanding are extremely limited in scope, often focusing on a single semantic category and region (e.g. apartments from a single country). This contrasts with the wide variety of shapes and sizes of real-world buildings which reflect their diverse purposes. In this work, we introduce* WAFFLE*, a novel multimodal floorplan understanding dataset of nearly 20K floorplan images and metadata curated from Internet data spanning diverse building types, locations, and data formats. By using a large language model and multimodal foundation models, we curate and extract semantic information from these images and their accompanying noisy metadata. We show that* WAFFLE *serves as a challenging benchmark for prior computational methods, while enabling progress on new floorplan understanding tasks. We will publicly release* WAFFLE *along with our code and trained models, providing the research community with a new foundation for learning the semantics of buildings.*

1. Introduction

"Life is chaotic, dangerous, and surprising. Buildings should reflect that."
—Frank Gehry

Buildings come in all shapes and sizes, from the tiny cottages dotting the English countryside to the imposing galleries of the temple of Angkor Wat. The diverse architectural designs of buildings have been influenced by their purposes, geographical locations, and changing trends throughout history. Recent years have seen a growing interest in the development of computational tools for architecture, which promise to aid experts engaged in the design and maintenance of buildings. Of particular interest is the automatic analysis of floorplans, the most fundamental element defining the structure of buildings which communicate rich schematic and layout information.

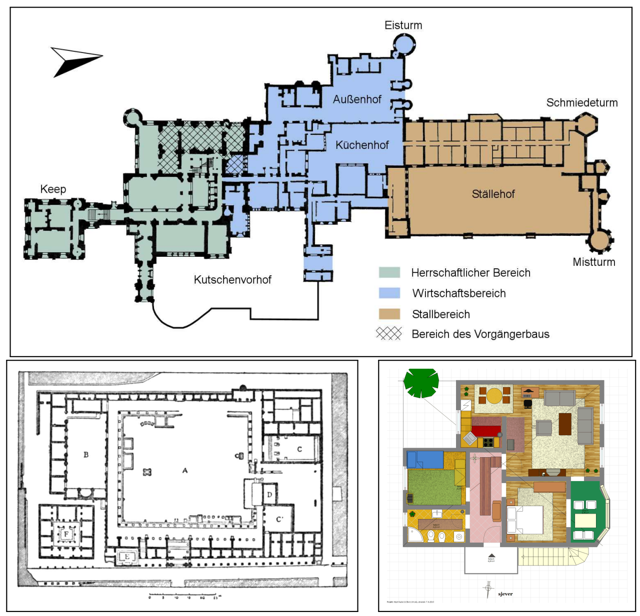


Figure 1. What can we understand from looking at these images? For instance, do we have a sense of what type of buildings these floorplans depict? Floorplans provide multimodal cues over the semantics and structure of buildings; however, they are often opaque for non-professionals, particularly for images lacking textual descriptions (such as the bottom images). We propose *WAFFLE*, a new multimodal dataset depicting floorplan images associated with rich textual descriptions. Our dataset allows for understanding *in-the-wild* floorplan imagery illustrating a wide array of building types. For example, a vision-and-language model finetuned on our data can correctly predict the building types for the examples depicted above (answers are provided below*).

*Denotes equal contribution

*From left to right: castle, temple, residential building. These samples (depicting the *Penrhyn Castle* in Wales, the *Forum at Timgad* in Algeria and a house floorplan in Bosnia and Herzegovina, respectively) were taken from the *WAFFLE* test set.

Prior works have tapped into the vast visual knowledge encoded by floorplans for various applications, such as 3D reconstruction [22] and floorplan-guided building navigation [24, 36]. However, prior data-driven techniques operating on floorplans mostly focus on extremely limited semantic domains (e.g. apartments) and geographical locations (often a single country), failing to cover the diversity needed for automatic understanding of floorplans in an unconstrained setting.

In this work, we introduce *WAFFLE* (***W**ikipedi**A**-**F**ueled **FL**oorplan **E**nsemble*), a multimodal floorplan understanding dataset comprised of diverse imagery spanning a variety of building types, geographical regions, historical eras, and data formats (as illustrated in Figure 1), along with comprehensive textual data. *WAFFLE* is derived from freely-available Internet images and metadata from the Wikimedia Commons platform. To turn noisy Internet data into this curated dataset with rich semantic annotations, we leverage state-of-the-art foundation models, using large language models (LLMs) and vision-language models (VLMs) to perform curation tasks with little or no supervision. This includes a decomposition of floorplans into visual elements; and structuring textual metadata, code and OCR detections with LLMs. By combining these powerful tools, we build a new dataset for floorplan understanding with rich and diverse semantics.

In addition to serving as a challenging benchmark for prior work, we show the utility of this data for various building understanding tasks that were not feasible with previous datasets. By using high-level and localized semantic labels along with floorplan images in *WAFFLE*, we learn to predict building semantics and use them to generate floorplan images with the correct building type, along with optional conditioning on structural configurations. Grounded labels within images also provide supervision to segment areas corresponding to domain-specific architectural terms. As shown by these applications, *WAFFLE* opens the door for semantic understanding and generation of buildings in a diverse, real-world setting.

2. Related Works

Floorplans in Computer Vision. Floorplans are a fundamental element of architectural design; as such, automatic understanding and generation of floorplans has drawn significant interest from the research community.

Several works aim to reconstruct floorplans, either from 3D scans [18, 44], RGB panoramas [2, 30, 43], room layouts [13] or combined modalities, such as sparse views and room-connectivity graphs [11]. Prior works also investigate the problem of alignment between floorplans and 3D point clouds depicting scenes [17]. Martin *et al*. [22] leverage floorplans of large-scale scenes to produce a unified reconstruction from disconnected 3D point clouds. Floorplans have also been utilized for navigation tasks. Several works predict position over a given floorplan, for a single image [36] or video sequences [6] depicting regions of the environment. Narasimhan *et al*. [24] train an agent to navigate in new environments by predicting corresponding labeled floorplans.

	Rent3D++	CubiCasa5K	RPLAN	SD	R2V	FloorPlanCAD	***WAFFLE***
Building types	— Residential buildings —					$\sim 4^{\diamond}$	$>1K$
Countries	UK	Fin.	Asia	Switz.	Japan	?[†]	>100
#Categories*	$\sim$15	$\sim$80	$\sim$13	91	22	35	∞[‡]
Real images?	✓	×	×	✓	✓	✓	✓

*Number of unique annotation values for labeled grounded regions or objects.

[†] Unspecified data source

[‡] Free text, on a subset of images

$^{\diamond}$Contains floorplans of 100 buildings spanning residential buildings, schools, hospitals, and shopping malls

Table 1. A comparison between *WAFFLE* and other floorplan datasets. SD above stands for Swiss Dwellings. We can see that, in contrast to our proposed *WAFFLE* dataset, most existing datasets focus on a single building type in a specific area in the world, and consider a small, closed list of annotation values.

Some works specifically target recognition of semantic elements over both rasterized [9, 45] and vectorized [42] floorplan representations, as well as applying this to perform raster-to-vector conversion [16, 19, 21]. In our work, we are interested in understanding Internet imagery of diverse data types such as raster graphics and photographs or scans of real floorplans. In contrast to prior work that mostly focuses on a fixed set of semantic elements in residential apartments, such as walls, bathrooms, closets, and so on, we are interested in acquiring higher-level reasoning over a wide array of building types.

The problem of synthesizing novel floorplans, and other types of 2D layouts such as documents [27, 47], has also received considerable interest (see the recent survey by Weber *et al*. [38] for a comprehensive review). Earlier works generate floorplans from high-level constraints, such as room adjacencies [15, 23]. Later works are able to generate novel floorplans in more challenging settings, *e.g*. only given their boundaries [14, 40]. In our work, we show that SOTA text-to-image generation tools can be fine-tuned for generating floorplans of diverse building types, not only residential buildings, as explored by prior methods.

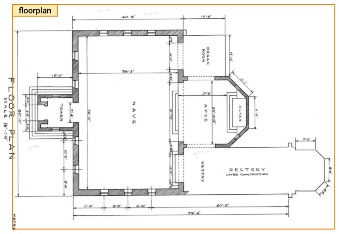
Floorplan Datasets. Prior datasets containing floorplan data are limited in structural and semantic diversity, typically being limited to residential building types such as apartments from specific geographic locations, often mined from real estate listings. For example, Rent3D++ [35] contains floorplans of 215 apartments located in London, and CubiCasa5K [16] contains floorplans of 5K Finnish apartments. The RPLAN [40] dataset contains 80K floorplans of apartments in Asia, further limited by various size and structural requirements (e.g., having only 3–9 rooms with specific proportions relative to the living room). The Swiss Dwellings [31] includes floorplan data for 42K Swiss apartments, and the Modified Swiss Dwellings [34] dataset provides a filtered subset of this data with additional access graph information for floorplan auto-completion learning. The R2V [19] dataset introduces 815 Japanese residential building floorplans.

Additionally, *WAFFLE* differs substantially from prior works with regards to the sourcing and curation of data. Datasets of real floorplans, such as those previously-mentioned, are constructed with tedious manual annotation. For example, specialists spent over 1K hours in the construction of FloorPlanCAD [10] to provide annotations of 30 categories (such as door, window, bed, etc.). Annotations may also derive from other input types rather than being direct annotations of floorplans; for instance, the Zillow Indoor Dataset [7] generates floorplans with user assistance from 360° panoramas, yielding plans for 1,524 homes after over 1.5K hours of manual annotation labor. To bypass such manual procedures, other works generate synthetic floorplans using predefined constraints [8]. By contrast, *WAFFLE* contains diverse Internet imagery of floorplans, including both original digital images and scans captured in the wild, and is curated with a fully automatic pipeline. See Table 1 for a comparison of the most related datasets with our proposed *WAFFLE* dataset.

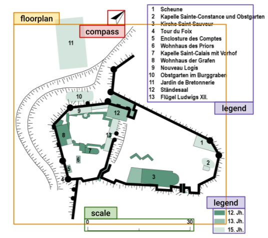
Finally, there are also large-scale datasets of landmark-centric image collections, such as Google Landmarks [26, 39] and WikiScenes [41]. Along with photographs and similar imagery of these landmarks, such collections may include schematic data such as floorplans. While prior works focus on the natural imagery in these collections for tasks such as image recognition, retrieval, and 3D reconstruction, we specifically leverage the schematic diagrams found in such collections for layout generation and understanding.

3. *WAFFLE*: Internet Floorplans Dataset

In this section, we introduce *WAFFLE* (***W**ikipedi**A**-**F**ueled **FL**oorplan **E**nsemble*), a new dataset of 18,556 floorplans, derived from Wikimedia Commons* and associated textual descriptions available on Wikipedia. *WAFFLE* contains floorplan images with paired structured metadata containing overall semantic information and spatially-grounded legends. Samples from our dataset are provided in



Building name: St. Paul's Episcopal Church
Building type: church
Country: United States of America
Grounded architectural features: {organ room rector, chapel, vestry, altar, tower, nave, apse}



Building name: Château de Blois
Building type: castle
Country: France
Grounded legend: {1: barn, 2:chapel, 3: saint savior church, 6: prior's house, 8: residence of counts, 9: new home, 10: orchard in moat, 11: bretonry garden, 12: booth hall}

Figure 2. **Samples from *WAFFLE***. Above, we show images paired with their structured data, including the building name and type, country of origin, and their grounded architectural features. We also visualize the detected layout components (floorplan, legend, compass, and scale, as relevant) overlaid on top of the images.

Figure 2. We provide an interactive viewer of samples from the *WAFFLE* dataset, and additional details and statistics of our dataset, in the supplementary material. We proceed to describe the curation process and contents of *WAFFLE*.

3.1. Data Collection

Images and metadata in Wikimedia Commons data are ordered by hierarchical categories (*WikiCategories*). To find relevant data, we recursively scrape the WikiCategories `Floor plans` and `Architectural drawings`, extracting images and metadata from Wikimedia Commons and the text of linked Wikipedia articles. As many images contain valuable textual information (e.g. hints to the location of origin, legend labels, etc.), we also extract text from the images using the Google Vision API* for optical character recognition (OCR). Finally, we decompose images into constituent items by fine-tuning the detection model DETR [3] on a small subset of labeled examples to predict bounding boxes for common layout components (floorplans, legend boxes, compass, and scale icons).

The raw data includes a significant amount of noise along with floorplans, including similar topics such as maps and cross-sectional blueprints as well as other unrelated data. Therefore, we filter this data as follows:

Text-based filtering (LLM). We perform an initial text-only filtering stage by processing our images' textual metadata with an LLM to extract structured information. We provide the LLM with a prompt containing image metadata and ask it to categorize the image in multiple-choice format, providing it with a closed set of possible categories. These include positive categories such as *floorplan* and *building* as well as some negative categories (not floorplans) such as *map* and *city*.

*https://commons.wikimedia.org

*https://cloud.google.com/vision?hl=en

Image-based filtering (CLIP). We use CLIP [28] image embeddings to filter for images likely to be floorplans. Firstly, as the WikiCategory `Architectural drawings` contains many non-floorplan images, we train a linear classifier on a balanced sample of items from the two WikiCategories and select images that are closer to those in the `Floor plans` WikiCategory. Moreover, we filter all images by comparing them with CLIP text prompt embeddings, following the use of CLIP for zero-shot classification. We compare to multiple prompts such as *A map*, *A picture of people*, and *A floorplan*, aggregating scores for positive and negative classes and filtering out images with low scores. Finally, we train a binary classifier using high-scoring images and negative examples to adjust the zero-shot CLIP classifications for increased recall.

This step results in a final dataset of nearly 20K images. Each image is accompanied by the following raw data extracted from its Wikimedia Commons page and linked pages: the image file name, its Wikimedia Commons page content (including a textual description), a list of linked WikiCategories, the contents of linked Wikipedia pages (if present), OCR detections in the image, and bounding boxes of constituent layout components.

3.2. LLM-Driven Structured pGT Generation

Our raw data contains significant grounded information about each image in diverse formats, which we wish to systematically organize and structure for use in downstream tasks. To this aim, we harness the capabilities of large language models (LLMs) for distilling essential information from diverse textual data. In particular, we extract the following information (also illustrated in Figure 2) by prompting Llama-2 [33] with an instruction and relevant metadata fields: building name, building type (i.e. *church*, *hotel*, *museum* etc.), location information (country, state, city), and a list of architectural features that are grounded in the image.

In general, the raw metadata contains considerable and diverse noise, involving multilingual content and multiple written representations of identical entities (e.g. *Notre Dame Cathedral* vs. *Notre-Dame de Paris*). To control for the source language, we employ prompts that instruct the LLM to respond in English and request translations when necessary. For linking representations of identical entities (also known as *record linkage*), we employ LinkTransformer [1] clustering along with various textual heuristics. We provide additional details, including prompts used, in the supplementary material, and proceed to describe our method for grounding architectural features in floorplans.

Architectural Feature Extraction and Grounding. Many floorplan images indicate architectural information either directly with text on the relevant region, or indirectly using a legend. To identify legends and architectural information

Figure 3. We automatically extract legends and architectural features from the image raw data (illustrated on the left, either the image metadata or OCR detections) by prompting LLMs. We associate the keys with text detected in the image, yielding grounded regions associated with semantics.

marked directly on the floorplan, we examine the bounding boxes of floorplan and legend detections (using the model described in Section 3.1) and select OCR detections within these areas. We also extract additional legend information from image metadata by prompting the LLM with an instruction including page content from the image's Wikimedia Commons page or the code surrounding the image in its linked Wikipedia pages (as legends often appear in these locations). We further structure the legend outputs using regular expressions to identify key-value pairs. Finally, we link the legend keys and architectural features to the regions in the floorplan images coinciding with OCR detections, thus providing grounding for the semantic values of the image. See Figure 3 for an example.

3.3. Dataset Statistics

Our dataset contains nearly 20K images with accompanying metadata, in a range of formats. In particular, we note that our dataset contains over 1K vectorized floorplans. Additionally, our dataset contains more than 1K building types spread over more than 100 countries across the world, and over 11K different Grounded Architectural Features (GAFs) across almost 3K grounded images. We split into train and test sets (18,259 and 297 images respectively) by selecting according to country (train: 50 countries; test: 57 countries), thus ensuring disjointedness with regards to buildings and preventing data leakage.

Data Quality Validation. We manually inspect the test set images, removing images that do not contain a valid floorplan. Based on this validation, we find that 89% are indeed relevant floorplan images. We find this level of noise acceptable for training models on in-the-wild data, while the manual filtering assures a clean test set for evaluation. In addition, we manually inspect the quality of our generated pGTs. We find that 89% of the building names, 85% of the building types and 96% of the countries of origin are accurately labeled (considering 100 random data samples).

4. Experiments

In this section, we perform several experiments applying our dataset to both discriminative and generative building

	R@1	R@5	R@8	R@16	MRR
CLIP	1.5%	7.6%	10.3%	19.7%	0.07
CLIP$_{FT}$	**11.8%**	**34.1%**	**40.0%**	**52.9%**	**0.23**

Table 2. Results on CLIP retrieval of building types, for CLIP before and after fine-tuning on our dataset. We report Recall@k (R@k) for $k \in \{1, 5, 8, 16\}$ and Mean Reciprocal Rank (MRR) for these models, evaluated on our test set. As seen above, fine-tuning on *WAFFLE* significantly improves retrieval metrics.

	CC5K*	CLIPSeg	Ours
AP	0.138	0.157	**0.226**
mIoU	0.057	0.066	**0.131**

Table 3. Open-Vocabulary Floorplan Segmentation Evaluation. We compare against a pretrained CLIPSeg model and against a closed-vocabulary segmentation model (CC5K). As illustrated above, our method improves localization across all evaluation metrics. *Evaluated only over a subset of residential buildings.

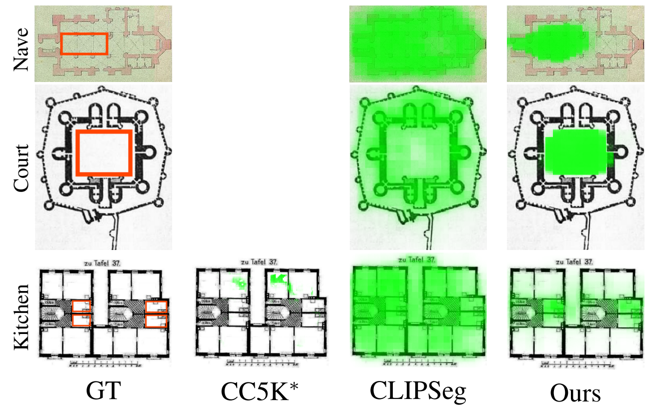


Figure 4. Comparison of open-vocabulary segmentation probability map results. We show the input images in the first column, with the corresponding GT regions in red. *Note that CC5K is a closed-vocabulary model designed for residential floorplan understanding, and therefore we cannot compare to it over additional building types (such as castles and cathedrals illustrated above). In addition to improving on the base CLIPSeg segmentation model, we outperform the strongly-supervised CC5K, suggesting that this model cannot generalize well beyond its training set distribution.

understanding tasks. For all tasks, we use the the train-test split outlined in Section 3.3. Please refer to the supplementary material for further training details.

4.1. Building Type Understanding

Task description. We test the ability to predict building-level semantics from a floorplan, similarly to a human who might look at a floorplan and make an educated guess as to what type of building it depicts. To learn this understanding, we fine-tune CLIP with a contrastive objective on paired images and building type pseudo-labels from *WAFFLE*. Our fine-tuned model (CLIP$_{FT}$) is expected to adjust CLIP to assign floorplan image embeddings close to those of relevant building types, allowing for subsequent retrieval or classification with floorplan images as input. We test the extent to which this understanding has been learned in practice with standard retrieval metrics, evaluating Recall@k for $k \in \{1, 5, 8, 16\}$ and Mean Reciprocal Rank (MRR).

Results. Results for fine-tuning CLIP for building type understanding are shown in Table 2. As is seen there, CLIP$_{FT}$ significantly outperforms the base model in retrieving the correct building type pseudo-labels, hence showing a better understanding of their global semantics.

4.2. Open-Vocabulary Floorplan Segmentation

Task description. To model localized semantics within floorplans, we use the GAFs in *WAFFLE* to fine-tune a text-driven segmentation model. We adopt the open-vocabulary text-guided segmentation model CLIPSeg [20] and perform fine-tuning on the subset of these grounded images.

To provide supervision, we use the values of the GAFs as input text prompts for the segmentation model and the OCR bounding box regions of the associated grounded values as segmentation targets. This yields partial ground truth supervision; for a text query, we use OCR bounding box regions corresponding to text labels that semantically match the query (implemented via text embedding similarity) as positive targets and the remaining bounding box regions as negative targets. To prevent leakage from the written text in the images, we perform inpainting with Stable Diffusion [29] to replace the contents of the OCR bounding boxes. As our inpainting process may cause artifacts, for evaluation purposes we manually select images that do not contain GAFs. We follow prior work [20] and report mean Intersection over Union (mIoU) and Average Precision (AP). The mIoU metric requires a threshold, which we empirically set to 0.25. AP is a threshold-agnostic metric that measures the area under the recall-precision curve, quantifying to what extent it can discriminate between correct and erroneous matches. In addition to comparing against the pretrained CLIPSeg model, we compare against the closed-vocabulary segmentation model provided by CubiCasa5K (CC5K) [16] over a subset of residential buildings in our test set (evaluating semantic regions which this model was trained on).

Results. Quantitative results are reported in Table 3, showing a clear boost in performance across both metrics. This is further reflected in our qualitative results in Figures 4. In addition, the results on residential buildings of the strongly-supervised residential floorplan understanding model [16] yields inferior performance, likely because the latter model uses supervision from a specific geographical region and style alone (a limitation of existing datasets, as we describe in Section 2). Overall, both metrics show that there is much room for improvements with future techniques leveraging

	Walls	Doors	Windows	Interior	BG
Precision	0.737	0.201	0.339	0.799	0.697
Recall	0.590	0.163	0.334	0.521	0.912
IoU	0.488	0.099	0.202	0.461	0.653

Table 4. Benchmark for Semantic Segmentation Evaluation. We benchmark prior work, reporting performance over the CubiCasa-5k [16] segmentation model, on common grounded categories. Note that background is denoted as BG above. As illustrated, *WAFFLE* serves as a challenging benchmark for existing work.

	FID ↓	KMMD ↓	CLIP Sim. ↑
SD	194.8	0.10	24.9
SD_{FT}	**145.3**	**0.07**	**25.6**

Table 5. Results on generated images, using a base and fine-tuned Stable Diffusion (SD) model. We compare the quality of the generated images (FID, KMMD) and the similarity to the given prompt (CLIP Sim.). As illustrated above, SD fine-tuning improves both realism and semantic correctness of image generations.

our data for segmentation-related tasks.

4.3. Benchmark for Semantic Segmentation

Following prior work [16, 19, 40], we consider segmentation of rasterized floorplan images into fine-grained localized categories, as locating elements such as walls has applications to various downstream tasks. To provide a new benchmark for performance on the diverse floorplans in *WAFFLE*, we manually annotate pixel-level segmentation maps for more than a hundred images over categories applicable to most building types: *wall*, *door*, *window*, *interior* and *background*. As our dataset contains a variety of data types, we annotate SVG-formatted images, which can be easily manually annotated by region.

We illustrate the utility of this benchmark by evaluating a standard existing model, namely the supervised segmentation model provided by CC5K [16]. We also evaluate a modern diffusion-based architecture trained with the same supervised data to predict wall locations as black-and-white images, to explore whether architectural modifications can yield improved performance. Further details of these models are provided in the supplementary material.

Results. Table 4 includes a quantitative evaluation of the existing model provided by CC5K on our benchmark. As illustrated in the table, our dataset provides a challenging benchmark for existing models, yielding low performance, particularly for more fine-grained categories, such as doors and windows. In addition to these results, we find that the modern diffusion architecture shows significantly better performance at localization of walls, generating binary maps with higher quantitative metric values (+1.2% in precision, +36.4% in recall and +29.5% in IoU, in comparison to the values obtained on the wall category in Table 4). This additional experiment shows promise in using stronger architectures for improving localized knowledge on weakly supervised in-the-wild data to ultimately approach the goal of pixel-level localization within diverse floorplans. Qualitative results from both models, along with ground truth segmentations, are provided in the supplementary material.

4.4. Text-Conditioned Floorplan Generation

Task description. Inspired by the rich literature on automatic floorplan generation, we fine-tune a text-to-image generation model on paired images and pGT textual data from *WAFFLE* for text-guided generation of floorplan images. We adopt the latent diffusion model Stable Diffusion [29] (SD), using prompts of the form “A floor plan of a <building_type>” which use the building type pseudo-labels from our LLM-extracted data. We balance training samples across building names and types to avoid overfitting on common categories. We evaluate the realism of these generations using Fréchet Inception Distance (FID) [12] as well as Kernel Maximum Mean Discrepancy (KMMD), since FID can be unstable on small datasets [5]. Similar to prior work [4, 25, 37], we measure KMMD on Inception features [32]. To measure semantic correctness, we measure CLIP similarity (using pretrained CLIP) between generations and prompts. All metrics were calculated per building type, averaging over the most common 15 types.

Results. Table 5 summarizes quantitative metrics, comparing floorplan generation using the base SD model with our fine-tuned version. These provide evidence that our model generates more realistic floorplan images that better adhere to the given prompts. Supporting this, we provide examples of such generated floorplans in Figure 5, observing the diversity and semantic layouts predicted by our model for various prompts. We note that our model correctly predicts the distinctive elements of each building type, such as the peripheral towers of castles and numerous side rooms for patient examinations in hospitals. Such distinctive elements are mostly not observed in the pretrained SD model, which generally struggles at generating floorplans. To further illustrate that our generated floorplans better convey the building type specified within the target text prompt, we conducted a user study. Given a pair of images, one generated with the pretrained model and one with our finetuned model, users were asked to select the image that best conveys the target text prompt. We find that 70.42% of the time users prefer our generated images, in comparison to the generations of the pretrained model. Additional details regarding this study are provided in the supplementary.

4.5. Structure-Conditioned Floorplan Generation

Task description. Structural conditions for floorplan generation have attracted particular interest, as architects may wish to design floorplans given a fixed building shape or

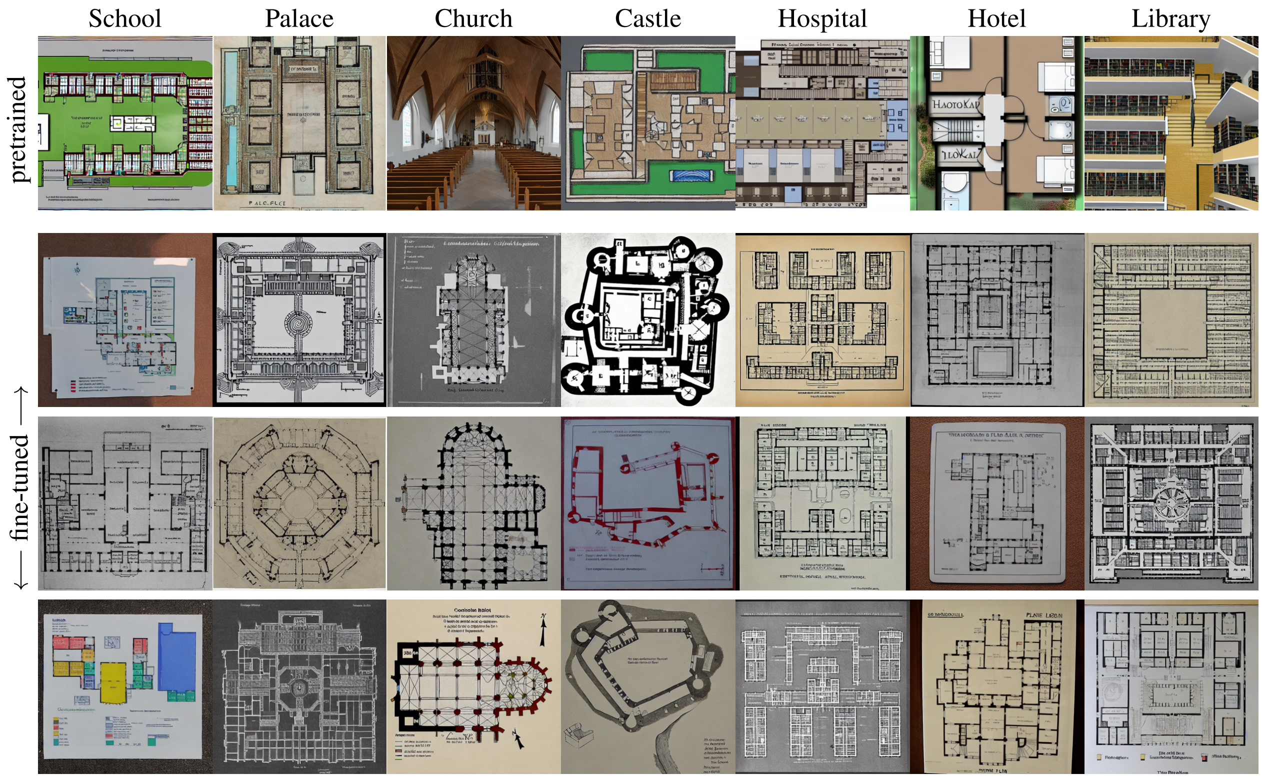


Figure 5. Examples for generated floorplans for various building types, using the prompt “A floor plan of a <building_type>” (corresponding types are shown on top). The first row shows samples from the pretrained SD model, and the bottom three show results from the model fine-tuned on *WAFFLE*. As seen above, pretrained SD struggles at generating floorplans in general and often yields results that do not structurally resemble real floorplans. By contrast, our fine-tuned model can correctly generate fine-grained architectural structures, such as towers in castles or long corridors in libraries.

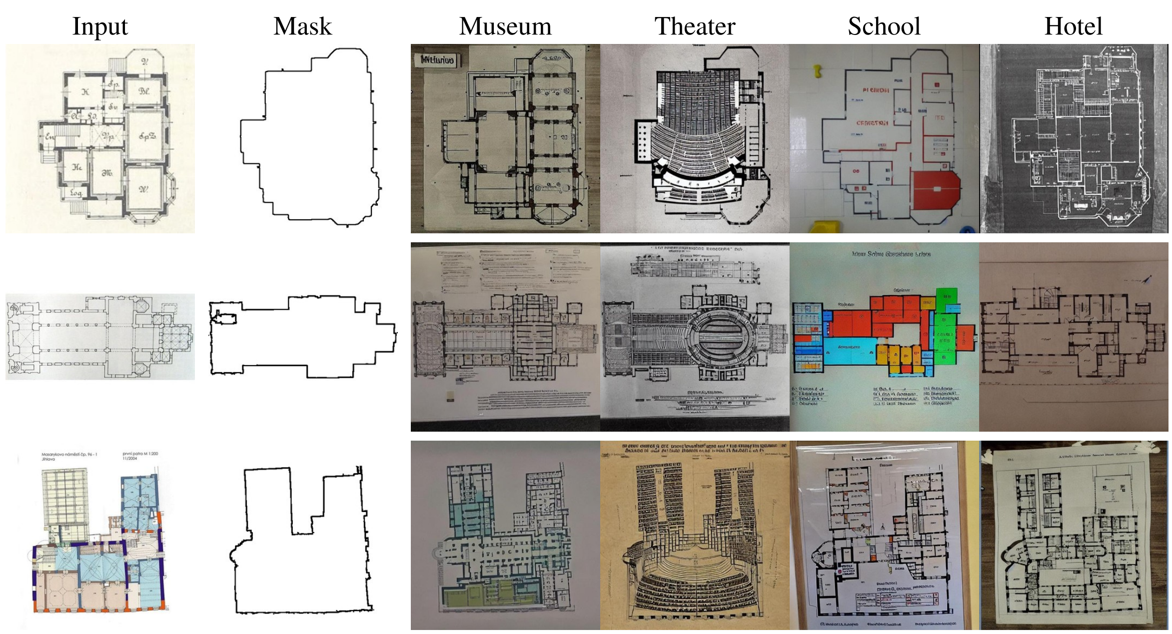


Figure 6. Boundary-conditioned generation. The first column shows images in *WAFFLE*, the second column shows automatically-extracted boundary masks, and the following columns show floorplan image generations conditioned on this generation with diverse building types provided as prompts.

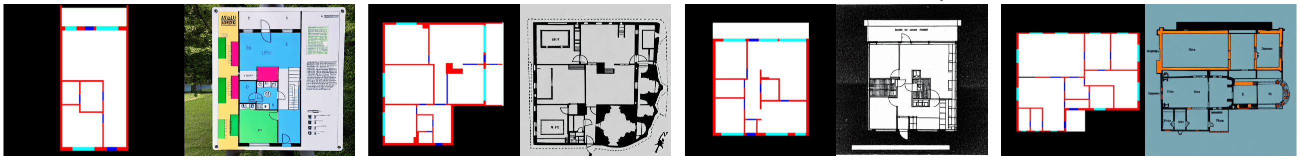


Figure 7. Structure-conditioned generation. For each image pair, the first image displays a building layout condition, taken from the existing CubiCasa5K dataset, which defines foreground (white) and background (black) regions, walls (red), doors (blue), and windows (cyan). The second image shows a generation conditioned on this layout, using the ControlNet-based model described in Section 4.5. Our image data and metadata enable the generation of diverse building types with structural constraints, without requiring any pixel-level annotations of images in *WAFFLE*. Notably, this succeeds even when the constraint is highly unusual for the corresponding building type, such as the condition above for *cathedral* (as cathedrals are usually constructed in a cross shape).

desired room configuration [14, 15, 23, 40]. Unlike existing works that consider residential buildings exclusively, we operate on the diverse set of building types and configurations found in *WAFFLE*, providing conditioning to our generative model SD_{FT} by fine-tuning ControlNet [46] for conditional generation, combined with applying text prompts reflecting various building types.

We are challenged by the fact that data in *WAFFLE*, captured in-the-wild, does not contain localized structural annotations (locations of walls, doors, windows or other features) such as those painstakingly annotated in some existing datasets. Therefore, we leverage our data in an unsupervised manner to achieve conditioning. To condition on the desired building boundary, we approximate the outer shape for all images in the training set via an edge detection algorithm. To condition on more complex internal structures, we instead train ControlNet on image-condition pairs derived from the existing annotated CubiCasa5K dataset [16]. By using SD_{FT} as the backbone for this ControlNet, reducing the conditioning scale (assigning less weight to the conditioning input during inference) and using a relatively high classifier-free guidance scale factor (assigning higher weight to the prompt condition), we fuse the ability of SD_{FT} to generate diverse building types while incorporating the structural constraints derived from external annotated data of residential buildings.

Results. We provide structurally-conditioned generations in Figures 6–7 for various building types. For boundary conditioning, the condition shape is extracted from existing images in our dataset. For structure conditioning, the conditions are derived from the annotations in the external CubiCasa5K dataset, using categories relevant to the diverse buildings in *WAFFLE*. These examples illustrate that our model is able to control the contents and style of the building according to the text prompt while adhering to the overall layout of the condition. This again demonstrates that the model has learned the distinct characteristics of each building type. In addition, we note that this succeeds even when the structural constraint is highly unusual for the paired building type, such as the cathedral in Figure 7 which deviates from the typical layout of a cathedral (usually constructed in a cross shape) in order to obey the condition.

5. Conclusion

We have presented the *WAFFLE* dataset of diverse floorplans in the wild, curated from Internet data spanning diverse building types, geographic locations, and architectural features. To construct a large dataset of images with rich, structured metadata, we leverage SOTA LLMs and multimodal representations for filtering and extracting structure from noisy metadata, reflecting both the global semantics of buildings and localized semantics grounded in regions within floorplans. We show that this data can be used to train models for building understanding tasks, enabling progress on both discriminative and generative tasks which were previously not feasible.

While our dataset expands the scope of floorplan understanding to new unexplored tasks, it still has limitations. As we collect diverse images in-the-wild, our data naturally contains noise (mitigated by our data collection and cleaning pipeline) which could affect downstream performance. In addition, while our dataset covers a diverse set of building types, it leans towards historic and religious buildings, possibly introducing bias towards these semantic domains. We focus on 2D floorplan images, though we see promise in our approach and data for spurring further research in adjacent domains, such as 3D building generation and such as architectural diagram understanding in general. In particular, although our work does not consider the 3D structure of buildings, we see promise in the use of our floorplans for aligning in-the-wild 3D point clouds or producing 3D-consistent room and building layouts. Finally, our work could provide a basis for navigation tasks which require indoor spatial understanding, such as indoor and household robotics. We envision future architectural understanding models that are enabled by datasets such as *WAFFLE* will explore new challenging tasks such as visual question answering for floorplans, which could be enabled by our textual metadata and open-vocabulary architectural features.

References

[1] Abhishek Arora and Melissa Dell. Linktransformer: A unified package for record linkage with transformer language models. *arXiv preprint arXiv:2309.00789*, 2023. 4

[2] Ricardo Cabral and Yasutaka Furukawa. Piecewise planar and compact floorplan reconstruction from images. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 628–635. IEEE, 2014. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[4] Eric Ming Chen, Jin Sun, Apoorv Khandelwal, Dani Lischinski, Noah Snavely, and Hadar Averbuch-Elor. What's in a decade? transforming faces through time. In *Computer Graphics Forum*, volume 42, pages 281–291. Wiley Online Library, 2023. 6

[5] Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them, 2020. 6

[6] Hang Chu, Dong Ki Kim, and Tsuhan Chen. You are here: Mimicking the human thinking process in reading floorplans. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2210–2218, 2015. 2

[7] Steve Cruz, Will Hutchcroft, Yuguang Li, Naji Khosravan, Ivaylo Boyadzhiev, and Sing Bing Kang. Zillow indoor dataset: Annotated floor plans with 360deg panoramas and 3d room layouts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2133–2143, 2021. 3

[8] Mathieu Delalandre, Ernest Valveny, Tony Pridmore, and Dimosthenis Karatzas. Generation of synthetic documents for performance evaluation of symbol recognition & spotting systems. *International Journal on Document Analysis and Recognition (IJDAR)*, 13(3):187–207, 2010. 3

[9] Samuel Dodge, Jiu Xu, and Björn Stenger. Parsing floor plan images. In *2017 Fifteenth IAPR international conference on machine vision applications (MVA)*, pages 358–361. IEEE, 2017. 2

[10] Zhiwen Fan, Lingjie Zhu, Honghua Li, Siyu Zhu, and Ping Tan. Floorplancad: A large-scale cad drawing dataset for panoptic symbol. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021. 3

[11] Arnaud Gueze, Matthieu Ospici, Damien Rohmer, and Marie-Paule Cani. Floor plan reconstruction from sparse views: Combining graph neural network with constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1583–1592, 2023. 2

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. 6

[13] Sepidehsadat Hosseini, Mohammad Amin Shabani, Saghar Irandoust, and Yasutaka Furukawa. Puzzlefusion: Unleashing the power of diffusion models for spatial puzzle solving. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[14] Ruizhen Hu, Zeyu Huang, Yuhan Tang, Oliver Van Kaick, Hao Zhang, and Hui Huang. Graph2plan: Learning floorplan generation from layout graphs. *ACM Transactions on Graphics (TOG)*, 39(4):118–1, 2020. 2, 8

[15] Hao Hua. Irregular architectural layout synthesis with graphical inputs. *Automation in construction*, 72:388–396, 2016. 2, 8

[16] Ahti Kalervo, Juha Ylioinas, Markus Häikiö, Antti Karhu, and Juho Kannala. Cubicasa5k: A dataset and an improved multi-task model for floorplan image analysis. In *Image Analysis: 21st Scandinavian Conference, SCIA 2019, Norrköping, Sweden, June 11–13, 2019, Proceedings 21*, pages 28–40. Springer, 2019. 2, 3, 5, 6, 8

[17] Ryan S Kaminsky, Noah Snavely, Steven M Seitz, and Richard Szeliski. Alignment of 3d point clouds to overhead images. In *2009 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 63–70. IEEE, 2009. 2

[18] Chen Liu, Jiaye Wu, and Yasutaka Furukawa. Floornet: A unified framework for floorplan reconstruction from 3d scans. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–217, 2018. 2

[19] Chen Liu, Jiajun Wu, Pushmeet Kohli, and Yasutaka Furukawa. Raster-to-vector: Revisiting floorplan transformation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2195–2203, 2017. 2, 3, 6

[20] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022. 5

[21] Xiaolei Lv, Shengchu Zhao, Xinyang Yu, and Binqiang Zhao. Residential floor plan recognition and reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16717–16726, 2021. 2

[22] Ricardo Martin-Brualla, Yanling He, Bryan C Russell, and Steven M Seitz. The 3d jigsaw puzzle: Mapping large indoor spaces. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, pages 1–16. Springer, 2014. 2

[23] Paul Merrell, Eric Schkufza, and Vladlen Koltun. Computer-generated residential building layouts. In *ACM SIGGRAPH Asia 2010 papers*, pages 1–12. 2010. 2, 8

[24] Medhini Narasimhan, Erik Wijmans, Xinlei Chen, Trevor Darrell, Dhruv Batra, Devi Parikh, and Amanpreet Singh. Seeing the un-scene: Learning amodal semantic maps for room navigation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 513–529. Springer, 2020. 2

[25] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation, 2019. 6

[26] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3465, 2017. 3

[27] Akshay Gadi Patil, Omri Ben-Eliezer, Or Perel, and Hadar Averbuch-Elor. Read: Recursive autoencoders for document layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 544–545, 2020. 2

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 5, 6

[30] Mohammad Amin Shabani, Weilian Song, Makoto Odamaki, Hirochika Fujiki, and Yasutaka Furukawa. Extreme structure from motion for indoor panoramas without visual overlaps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5703–5711, 2021. 2

[31] Matthias Standfest, Michael Franzen, Yvonne Schröder, Luis Gonzales Medina, Yarilo Villanueva Hernandez, Jan Hendrik Buck, Yen-Ling Tan, Milena Niedzwiecka, and Rachele Colmegna. Swiss dwellings: a large dataset of apartment models including aggregated geolocation-based simulation results covering viewshed, natural light, traffic noise, centrality and geometric analysis, 2022. 3

[32] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 6

[33] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 4

[34] Casper van Engelenburg, Seyran Khademi, Fatemeh Mostafavi, Matthias Standfest, and Michael Franzen. Modified swiss dwellings: a machine learning-ready dataset for floor plan auto-completion at scale, 2023. 3

[35] Madhawa Vidanapathirana, Qirui Wu, Yasutaka Furukawa, Angel X Chang, and Manolis Savva. Plan2scene: Converting floorplans to 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10733–10742, 2021. 3

[36] Shenlong Wang, Sanja Fidler, and Raquel Urtasun. Lost shopping! monocular localization in large indoor spaces. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2695–2703, 2015. 2

[37] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9332–9341, 2020. 6

[38] Ramon Elias Weber, Caitlin Mueller, and Christoph Reinhart. Automated floorplan generation in architectural design: A review of methods and applications. *Automation in Construction*, 140:104385, 2022. 2

[39] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584, 2020. 3

[40] Wenming Wu, Xiao-Ming Fu, Rui Tang, Yuhan Wang, Yu-Hao Qi, and Ligang Liu. Data-driven interior plan generation for residential buildings. *ACM Transactions on Graphics (TOG)*, 38(6):1–12, 2019. 2, 3, 6, 8

[41] Xiaoshi Wu, Hadar Averbuch-Elor, Jin Sun, and Noah Snavely. Towers of babel: Combining images, language, and 3d geometry for learning multimodal vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 428–437, 2021. 3

[42] Bingchen Yang, Haiyong Jiang, Hao Pan, and Jun Xiao. Vectorfloorseg: Two-stream graph attention network for vectorized roughcast floorplan segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1358–1367, 2023. 2

[43] Shang-Ta Yang, Fu-En Wang, Chi-Han Peng, Peter Wonka, Min Sun, and Hung-Kuo Chu. Dula-net: A dual-projection network for estimating room layouts from a single rgb panorama. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3363–3372, 2019. 2

[44] Yuanwen Yue, Theodora Kontogianni, Konrad Schindler, and Francis Engelmann. Connecting the dots: Floorplan reconstruction using two-level queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 845–854, 2023. 2

[45] Zhiliang Zeng, Xianzhi Li, Ying Kin Yu, and Chi-Wing Fu. Deep floor plan recognition using a multi-task network with room-boundary-guided attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9096–9104, 2019. 2

[46] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 8

[47] Xinru Zheng, Xiaotian Qiao, Ying Cao, and Rynson WH Lau. Content-aware generative modeling of graphic design layouts. *ACM Transactions on Graphics (TOG)*, 38(4):1–15, 2019. 2