Recipe Diffusion: Cross-Frame Attention and Region-Aware Diffusion for Coherent Visual Recipe Instruction Generation

Weiyi Xia* and Satoru Fujita†
* Hosei University, Tokyo
E-mail: weiyi.xia.9a@stu.hosei.ac.jp Tel: +81-42-387-4545
† Hosei University, Tokyo
E-mail: fujita_s@hosei.ac.jp Tel: +81-42-387-4545

Fig. 1. Recipe Sequence Generation Results

***Abstract*—This paper presents a cross-frame attention and region-aware diffusion method for generating coherent, step-by-step visual instructions for cooking recipes. Our approach combines two complementary mechanisms: (1) cross-frame key-value sharing in attention layers to maintain global consistency across sequential frames, and (2) region-aware noise application, which preserves object identity while allowing contextual changes. Unlike conventional models that generate each image independently, our training-free framework leverages pre-trained detection and segmentation models to create region masks and modifies the attention mechanism to share visual features across frames. By integrating differential noise application with cross-frame attention consistency, our system generates recipe instruction sequences that maintain both global coherence and local object identity throughout each step.**

I. Introduction

The generation of realistic and contextually relevant images has advanced remarkably with the advent of deep learning models, particularly diffusion models[1]–[3]. These models can synthesize high-fidelity images from textual prompts with unprecedented quality and diversity. However, despite their impressive performance in single-image generation, current text-to-image models face significant limitations when applied to the generation of coherent, step-by-step visual instructions for procedural tasks—especially cooking recipes.

As illustrated in Fig.2, conventional text-to-image generation approaches encounter several critical challenges when tasked with producing visual recipe instructions. First, *object identity inconsistency* arises when the same entity (e.g., a cooking pot or ingredient) changes in visual appearance, shape, or color across different steps, even when the accompanying text remains semantically consistent. Second, *spatial relationship drift* occurs when objects appear in inconsistent locations or orientations relative to each other, disrupting the logical flow

979-8-3315-7206-8/25/$31.00

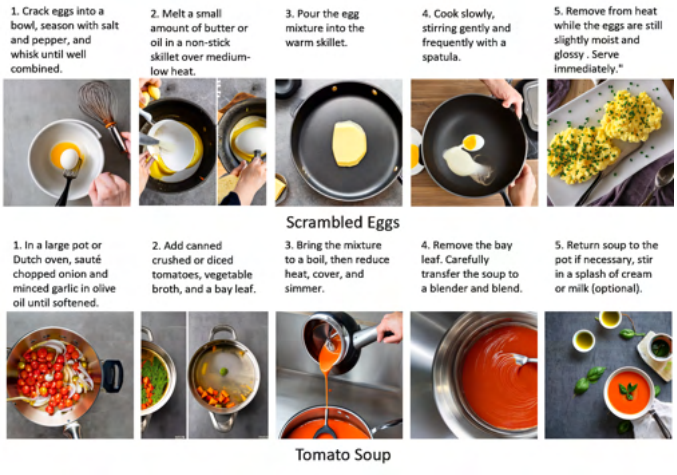

Fig. 2. Limitation of independent text-to-image generation in visual recipe instructions task

of procedural steps in the recipe.

These limitations stem from the fundamental design of current text-to-image models, which treat each generation as an independent process without mechanisms for preserving visual continuity across steps. Each image is generated from scratch using only textual conditioning, leading to stochastic variations that—while beneficial for creative diversity—are detrimental when visual consistency is essential for following recipe instructions.

While text-to-video generation methods[4], [5] offer an alternative for creating sequential visual content, they present distinct challenges and limitations when applied to procedural recipe instruction generation. Text-to-video models typically prioritize temporal smoothness and motion coherence within continuous scenes, often producing relatively short sequences (2–8 seconds) with limited capacity to handle discrete step transitions or significant scene changes. Moreover, these models are optimized for capturing motion dynamics rather than preserving object identity across drastically different contexts, making them less suitable for multi-step recipe instructions, where each frame may depict a distinct procedural state with varying backgrounds, lighting conditions, and object arrangements.

This paper proposes Recipe Diffusion, a novel framework specifically designed to address the challenge of generating visually consistent, step-by-step image sequences for cooking recipes. Our approach integrates state-of-the-art object detection (GroundingDINO[6]) and segmentation (Segment Anything Model - SAM[7]) with a powerful diffusion-based image generator (Stable Diffusion). The core innovation lies on identifying and isolating specific regions of interest (e.g., a container in a recipe) in an initial image, and then using this spatial information to guide the generation of subsequent images. By strategically applying noise to the latent space—applying less noise to regions intended to remain consistent and more noise to background or changing elements—our method encourages the preservation of visual characteristics of key objects while allowing the rest of the scene to evolve according to new prompts.

As demonstrated in Fig.1, our proposed method generates high-quality, visually consistent representations from recipe instruction texts. It overcomes the limitations of conventional text-to-image approaches while offering greater control and flexibility than existing text-to-video methods for procedural recipe content generation.

The main contributions of this work are: (1) a novel dual-mechanism framework combining cross-frame attention sharing and region-aware diffusion to generate visually consistent image sequences for recipe instructions; (2) a cross-frame key-value sharing strategy in self-attention layers that propagates visual features across sequential frames; (3) a region-aware noise application strategy that preserves object identity while allowing contextual evolution; (4) a configurable system enabling flexible control over both attention-based and region-based consistency mechanisms; and (5) demonstration of the framework's ability to generate coherent multi-step recipe instruction sequences that surpass the limitations of both conventional text-to-image and text-to-video approaches in this specific application domain.

II. Related Work

A. Image Generation with Diffusion Models

Diffusion models have emerged as the leading paradigm for high-quality image synthesis. Models like Stable Diffusion[3], DALL-E 3[2], and Imagen[8] have demonstrated impressive capabilities in generating diverse and photorealistic images from text prompts. They work by iteratively denoising a random noise vector, guided by the input prompt, to produce an image. While powerful for single image generation, maintaining consistency across sequences is not an inherent feature.

B. Controllable Image Generation and Editing

Significant research has focused on enhancing the controllability of generative models. These methods enable control over style, layout, object attributes, or incorporate additional conditioning inputs such as sketches or segmentation maps. ControlNet [9], for example, provides fine-grained spatial control over diffusion models through various conditioning mechanisms. A complementary line of work exploits precise object segmentation for controlled image editing. These approaches employ models like GroundingDINO [10] and SAM [7] to obtain accurate object masks, then confine modifications to the masked regions using diffusion-based inpainting. For instance, such methods can extract a mask for a container's contents and apply Stable Diffusion's inpainting capabilities to modify only the contents while preserving the container structure. While these mask-based inpainting approaches prove effective for targeted edits, they face limitations in maintaining nuanced visual consistency of object properties (such as texture and lighting) across image sequences with evolving contexts. This challenge arises because standard inpainting lacks explicit guidance for feature preservation. Our method addresses this

limitation by applying differential noise to the entire latent space, with the noise distribution guided by object masks to achieve better consistency control.

C. Attention Mechanisms for Temporal Consistency

Recent advances in video generation[4] and sequential image synthesis have leveraged attention mechanisms for maintaining temporal consistency. Key-value sharing in attention layers has been employed in video diffusion models to propagate features across frames. Our work adapts this concept to discrete instruction sequences, implementing selective KV sharing in self-attention layers of the diffusion U-Net to maintain visual coherence across procedural steps.

III. Recipe Diffusion

The Recipe Diffusion framework is specifically designed to generate coherent, step-by-step visual sequences for cooking recipe instructions—a domain that demands rigorous visual consistency. In recipe generation, it is critical to maintain object identity (e.g., a specific bowl or cutting board) across transformations (e.g., adding ingredients, stirring), depict plausible changes in ingredient states (e.g., raw to cooked, solid to liquid), and preserve logical spatial relationships. Our framework addresses these challenges by integrating two complementary consistency mechanisms: cross-frame attention sharing and region-aware noise application.

As illustrated in Fig.3, our framework generates recipe steps sequentially. Each recipe instruction S_i is first processed by (as detailed in Section III-A). For the initial image I_0, it is generated independently from its corresponding prompt. For subsequent images I_i in the sequence ($i > 0$), the generation process uses the previously generated image I_{i-1} as a visual reference. Specifically, a latent representation L_i for the current step is initialized by applying a controlled level of noise (e.g., 0.999 noise level) to the latent of I_{i-1} rather than starting from pure random noise. This provides a coherent starting point for the diffusion process.

During the subsequent denoising process for I_i, two key consistency mechanisms work in tandem to ensure both global coherence and local object fidelity:

1) Cross-Frame Attention Mechanism: Our framework employs a Key-Value (KV) sharing mechanism within the diffusion model's U-Net. This allows the current image's generation to reference and propagate visual features from earlier frames (e.g., I_{i-1} or a designated reference step from LLM metadata), ensuring global consistency throughout the sequence.
2) Region-Aware Noise Application: If the metadata pre-processed by the LLM indicates that the current step S_i involves a container or primary object that should remain consistent with a preceding frame, our framework utilizes pre-trained object detection (GroundingDINO) and segmentation (SAM) models to identify and create a precise mask for this container in the reference image. Then, a differential noise application strategy is employed in the latent space, where less noise is applied to the masked container region.

This integrated approach allows our framework to dynamically balance the generation of new contextual elements with the unwavering preservation of key objects, ultimately producing visually consistent step-by-step recipe sequences that are both globally coherent and locally faithful to reference objects across sequences.

A. Recipe Instruction Pre-processing with LLM

Prior to image generation, raw textual recipe instructions undergo a crucial pre-processing step using a large language model (LLM), specifically ChatGPT, to extract structured metadata essential for maintaining visual consistency. This step aims to identify key entities and their relationships across recipe steps, which directly inform our cross-frame attention and region-aware noise application mechanisms.

For each recipe, the LLM parses the step-by-step instructions and generates a comprehensive structured output. This output includes, for each step, details about the primary objects or containers involved (e.g., "colander," "salad bowl"), their relevant visual attributes or styles (e.g., "stainless steel," "wooden"), and critically, a reference to a preceding step if the same object is expected to persist visually. For objects introduced for the first time, this reference would be null. Additionally, the pre-processing tracks the initial appearance of each distinct container or key object within the overall recipe sequence. This extracted metadata allows our framework to dynamically select appropriate visual reference frames and generate precise object masks for subsequent processing, ensuring that elements requiring visual continuity are consistently represented throughout the generated sequence.

B. Cross-Frame Attention Mechanism

To maintain global visual consistency across the sequence, our framework modifies the self-attention layers of the diffusion model's U-Net architecture. Specifically, we implement a key-value (KV) sharing mechanism in self-attention layers that allows the current frame to reference visual features from previous frames. This is crucial for propagating high-level semantic features and ensuring a consistent visual style and structure throughout the generated sequence.

For each self-attention layer, when generating image I_i (the current frame), the system utilizes stored key and value tensors $(K_{\text{ref}}, V_{\text{ref}})$ from a designated reference frame I_j where $j < i$. During the attention computation for the current frame, these stored tensors are concatenated with the current frame's keys and values (denoted as K_i and V_i):

$$K^+ = [K_{\text{ref}} \oplus K_i] \tag{1}$$

$$V^+ = [V_{\text{ref}} \oplus V_i] \tag{2}$$

where $\oplus$ denotes concatenation along the sequence dimension. The attention mechanism then operates this expanded key-value space, allowing the current frame's queries (Q_i) to attend

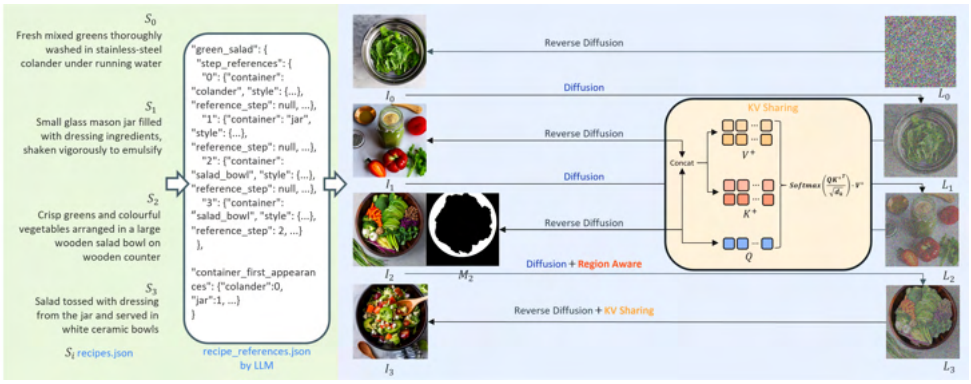

Fig. 3. Recipe Diffusion Framework

to both its own features and those from the reference frame. The attention scores are computed as:

$$A_i^+ = \text{softmax}\left(\frac{Q_i(K^+)^T}{\sqrt{d_k}}\right) \tag{3}$$

The resulting attention map $A_i^+ \in [0,1]^{P\times(P+P')}$ (where P is the number of tokens in the current frame and P' is the number of tokens in the reference frame, typically $P = P'$) allows each query from the current frame to weigh information from both the current and reference frames. Finally, the output feature map H_i for the self-attention layer is obtained by multiplying these attention scores with the combined values:

$$H_i = A_i^+ \cdot V^+ \tag{4}$$

This cross-frame attention sharing is applied selectively to important layers within the U-Net, particularly mid-block and up-sampling layers where high-level semantic features are processed and refined, as illustrated in Fig.3.

C. Region-Aware Noise Application

Complementing the global consistency provided by cross-frame attention, our framework employs a region-aware noise application mechanism to preserve specific object identities and their consistency across frames. This mechanism offers fine-grained control over local visual coherence. When generating image I_i which must remain consistent with a reference object from image I_j, the framework first identifies the object of interest in I_j using an object detection model such as GroundingDINO. It then segments it with a segmentation model like SAM (Segment Anything Model) to obtain a precise binary mask M_j.

The reference image I_j is then encoded into its latent space representation L_j. The obtained mask M_j is resized to match the dimensions of the latent space. To selectively preserve object consistency, two different noise levels are then applied to the latent representation: a specifically defined lower noise level (σ_{object}) for the object region and a standard noise level (σ_{base}) for the background. This differential noise application is defined as:

$$L_{\text{container}} = \text{AddNoise}(L_j, \sigma_{\text{object}}) \tag{5}$$

$$L_{\text{background}} = \text{AddNoise}(L_j, \sigma_{\text{base}}) \tag{6}$$

Here, σ_{object} is the pre-defined lower noise level (with $\sigma_{\text{object}} < \sigma_{\text{base}}$), ensuring that the object region ($L_{\text{container}}$) receives less noise perturbation and thus retains more of its original features. σ_{base} represents the standard noise level applied to the background. The final latent representation with noise, L_{combined} is then constructed by combining regions according to the resized mask:

$$L_{\text{combined}} = M_{j,\text{resized}} \odot L_{\text{container}} + (1 - M_{j,\text{resized}}) \odot L_{\text{background}} \tag{7}$$

where $\odot$ denotes element-wise multiplication. This combined latent L_{combined} serves as the input to the diffusion process for generating I_i, ensuring that the masked object region retains higher fidelity to its reference, while the surrounding background is diffused more freely.

Fig. 4. Examples of recipe sequences with different methods

TABLE I
COMPARISON WITH GENERAL-PURPOSE GENERATION MODELS

Method	CLIP Score↑	DreamSim↓	L2-Dinov2↓
SD2.1	0.3408	0.4623	48.8906
SC	0.4585	0.4536	47.0997
DALLE3	0.3669	0.4523	48.0874
RA w/o KV sharing (Ours)	0.3782	0.3609	38.6291
RA w/ KV sharing (Ours)	**0.3808**	**0.3534**	**36.8742**

TABLE II
COMPARISON WITH CONSISTENCY-MAINTAINING GENERATION METHODS

Method	CLIP Score↑	DreamSim↓	L2-Dinov2↓
Inpainting (Container)	0.3301	0.4325	45.7936
Inpainting (Interior)	0.3459	0.4367	45.7923
Heuristic Diffusion	0.3547	0.4303	43.2748
KV Sharing	0.2973	0.3745	40.1887
RA w/o KV Sharing (Ours)	0.3782	0.3609	38.6291
RA w/ KV sharing (Ours)	**0.3808**	**0.3534**	**36.8742**

TABLE III
ANNOTATION RESULTS FOR THE EVALUATION

Method	Best (%)	Second Best (%)	Third Best (%)
Inpainting (Container)	12.12	10.00	8.70
Inpainting (Interior)	7.07	13.75	17.39
Heuristic Diffusion	17.17	<u>20.00</u>	**20.29**
KV Sharing	4.04	7.50	17.39
RA w/o KV Sharing (Ours)	<u>25.25</u>	**36.25**	<u>18.84</u>
RA w/ KV Sharing (Ours)	**34.35**	12.50	17.39

IV. EXPERIMENTS

A. Metrics

To quantitatively assess the performance of our proposed Recipe Diffusion framework in generating visually coherent instructional sequences, we employ the following three metrics:

CLIP Score[11] quantifies semantic similarity between images and text, or between image pairs. It is used to evaluate semantic consistency across a sequence. A higher score indicates stronger semantic alignment.

DreamSim[12] measures perceptual similarity and is designed to align with human judgments. It compares the visual appearance of key objects across sequence frames. A lower value indicates higher perceptual similarity.

L2-DINOv2[13] quantifies deep feature similarity between image regions by computing the L2 distance of DINOv2 embeddings. This metric assesses the preservation of visual attributes for the same objects across sequence frames. A lower distance indicates greater feature-level consistency.

B. Implementation

Our framework is implemented using Stable Diffusion 2.1 as the base diffusion model, with GroundingDINO (SwinT-OGC variant) for object detection and SAM (ViT-H) for precise segmentation. The system runs on a single CUDA-enabled GPU due to the computational demands of both diffusion inference and segmentation processing. The cross-frame attention mechanism is realized by modifying the self-attention processors within the U-Net architecture. Key-value sharing is selectively applied to self-attention layers, particularly those in the mid-block and up-sampling stages, where high-level semantic features are most prominent. For region-aware noise application, we employ differential noise levels: a base noise level of 0.999 is used for background regions, while a reduced noise level of 0.9 for object regions that require consistency.

C. Automatic Metric Comparison

To objectively evaluate the quality and consistency of the generated recipe sequences, we employed automatic evaluation metrics. These metrics provide quantitative insights into different aspects of image quality and perceptual similarity. The evaluation was conducted on the average generation results across 8 distinct recipe types.

As shown in Table I and Table II, our proposed methods consistently demonstrate superior performance across all evaluation metrics. Region Aware with KV sharing(Ours) achieved the highest CLIP Score (0.3808) and the lowest DreamSim and L2-Dinov2 scores, indicating the best overall text-image alignment, perceptual quality, and feature-level similarity. Region Aware variant without KV sharing (Ours) also performed exceptionally well, outperforming other general-purpose models (SD2.1, SC[14], DALLE3) as well as consistency-maintaining techniques (Inpainting, Heuristic Diffusion, KV Sharing) on most metrics. While Heuristic Diffusion yielded moderate results, the Inpainting and standalone KV Sharing methods generally exhibited lower performance in comparison to our region-aware approaches. These objective findings consistently support the effectiveness of our proposed methods in generating high-quality and consistent recipe sequences.

D. Human Annotation

To assess the perceived quality and logical consistency of the generated recipe sequences, we conducted a double-blind human annotation study. A total of 27 university students participated in evaluation, assessing 6 different recipe types. For each recipe, participants selected the "Best," "Second Best," and "Third Best" sequences from 6 different generation methods, based on criteria such as accuracy, practical utility, and logical flow.

The results indicate a strong preference for our proposed methods. Region Aware with KV Sharing (Ours) achieved the highest percentage for "Best Selected" at 34.35%. Meanwhile, Region Aware without KV Sharing (Ours) excelled in the "Second Best Selected" category with 36.25%. These findings suggest that our region-aware approaches, especially with KV sharing, significantly improve the perceived quality and coherence of generated recipe sequences. While baseline methods like Heuristic Diffusion showed moderate performance, the inpainting and standalone KV Sharing methods generally received lower rankings. This human evaluation provides strong support for the effectiveness of our proposed framework in generating high-quality visual instructions.

E. Qualitative Examples

To provide a more intuitive understanding of the performance of different sequence generation methods, we present qualitative examples of generated recipe sequences. Fig.4 illustrates the step-by-step visual outputs for a sample recipe, comparing our proposed methods with representative baselines. As observed, our methods—particularly Region Aware with KV sharing— demonstrate superior visual coherence and logical progression across frames. These qualitative results align with the trends identified in both the automatic metric comparisons and the human annotation study, further validating the effectiveness of our approach.

V. Future Work

While Recipe Diffusion demonstrates promising results in generating coherent visual recipe instructions, several avenues for future research remain. We plan to explore more sophisticated methods for dynamic object interactions and complex state changes—such as modeling nuanced ingredient transformations (e.g., chopping or melting)—while preserving object identity. Extending the framework to support more complex recipe structures, including branching logic and multi-object interactions, also represents a critical next step. Additionally, we intend to investigate the benefits of fine-tuning the base diffusion model on larger procedural image sequence datasets to improve coherence, realism, and scalability. Expanding the framework's application to other domains, such as DIY guides or scientific protocols, will require adapting consistency mechanisms to new task-specific challenges.

VI. Conclusion

This paper introduced Recipe Diffusion, a novel framework designed to generate visually coherent image sequences for procedural instructions. We addressed the key limitations of conventional text-to-image models, which often struggle with object identity inconsistency and spatial relationship drift across sequential frames.

This work represents a significant advancement in the field of visual instruction generation, introducing a robust and configurable system capable of producing high-quality, step-by-step visual guides that effectively preserve both global coherence and local object identity. Our findings lay the groundwork for more reliable and impactful applications in automated content creation for educational, culinary, and technical domains.

Acknowledgment

This work was conducted as a "collaborative study under the NIJL Project ¡Research No.K452052428¿."

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[2] Z. Shi, X. Zhou, X. Qiu, and X. Zhu, "Improving image captioning with better use of captions," *arXiv preprint arXiv:2006.11807*, 2020.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[4] L. Khachatryan, A. Movsisyan, V. Tadevosyan, *et al.*, "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 954–15 964.

[5] J. Z. Wu, Y. Ge, X. Wang, *et al.*, "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7623–7633.

[6] S. Liu, Z. Zeng, T. Ren, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *European Conference on Computer Vision*, Springer, 2024, pp. 38–55.

[7] A. Kirillov, E. Mintun, N. Ravi, *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[8] C. Saharia, W. Chan, S. Saxena, *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[9] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3836–3847.

[10] S. Liu, Z. Zeng, T. Ren, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *European Conference on Computer Vision*, Springer, 2024, pp. 38–55.

[11] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," *CoRR*, vol. abs/2103.00020, 2021. arXiv: 2103.00020. [Online]. Available: https://arxiv.org/abs/2103.00020.

[12] S. Fu, N. Tamir, S. Sundaram, *et al.*, "Dreamsim: Learning new dimensions of human visual similarity using synthetic data," *arXiv preprint arXiv:2306.09344*, 2023.

[13] M. Oquab, T. Darcet, T. Moutakanni, *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[14] P. Pernias, D. Rampas, M. L. Richter, C. J. Pal, and M. Aubreville, "Würstchen: An efficient architecture for large-scale text-to-image diffusion models," *arXiv preprint arXiv:2306.00637*, 2023.