

# FLICKER: A Fine-Grained Contribution-Aware Accelerator for Real-Time 3D Gaussian Splatting

Wenhui Ou<sup>1</sup>, Zhuoyu Wu<sup>2</sup>, Yipu Zhang<sup>1</sup>, Dongjun Wu<sup>1</sup>, Freddy Ziyang Hong<sup>1</sup>, C. Patrick Yue<sup>1,\*</sup>

<sup>1</sup>Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology

<sup>2</sup>School of IT, Monash University, Malaysia Campus

wouab@connect.ust.hk, zhuoyu.wu@monash.edu, {yzhangqg, dwubf, zhongad}@connect.ust.hk, eepatrick@ust.hk

**Abstract**—Recently, 3D Gaussian Splatting (3DGS) has become a mainstream rendering technique for its photorealistic quality and low latency. However, the need to process massive non-contributing Gaussian points makes it struggle on resource-limited edge computing platforms and limits its use in next-gen AR/VR devices. A contribution-based prior skipping strategy is effective in alleviating this inefficiency, but the associated contribution-testing workload becomes prohibitive when it is further applied to the edge. In this paper, we present FLICKER, a contribution-aware 3DGS accelerator that leverages a hardware–software co-design framework, including adaptive leader pixels, pixel-rectangle grouping, hierarchical Gaussian testing, and mixed-precision architecture, to achieve near-pixel-level, contribution-driven rendering with minimal overhead. Experimental results show that our design achieves up to  $1.5\times$  speedup,  $2.6\times$  energy efficiency improvement, and 14% area reduction over a state-of-the-art accelerator. Meanwhile, it also achieves  $19.8\times$  speedup and  $26.7\times$  energy efficiency compared with a common edge GPU.

**Index Terms**—3DGS, Contribution-Aware, Accelerator

## I. INTRODUCTION

Advances in photorealistic novel view synthesis (NVS) have significantly enhanced the immersive experience in augmented and virtual reality (AR/VR) applications [1]. Recently, 3D Gaussian Splatting (3DGS) [2], [3] has emerged as a leading NVS technique for its outstanding rendering speed. This makes it a promising solution for next-generation AR/VR systems. However, such real-time capability is largely limited to powerful cloud or desktop-level GPUs, while edge devices still suffer from severe performance loss and energy constraints [4].

The limitation stems from its inherent principle. Specifically, 3DGS explicitly represents a scene with a set of anisotropic Gaussians, whose number often exceeds millions in real-world scenarios [5]. To avoid excessive memory footprint, the rendering of each frame is typically split into tiles, so that only the Gaussians that might contribute to the current tile are accessed instead of the entire set. However, the **large tile size** and **over-inclusive Gaussian testing** cause pixels to process a substantial number of unnecessary Gaussians [6]. This not only wastes computational resources, but more importantly, disrupts the efficient alignment of dataflow across parallel hardware units, leading to low hardware utilization. For edge devices [7] with limited computing units, such inefficiency is particularly detrimental to both performance and energy efficiency.

To address the aforementioned challenge, several works [6], [10]–[12] try to refine the identification of contributing Gaus-

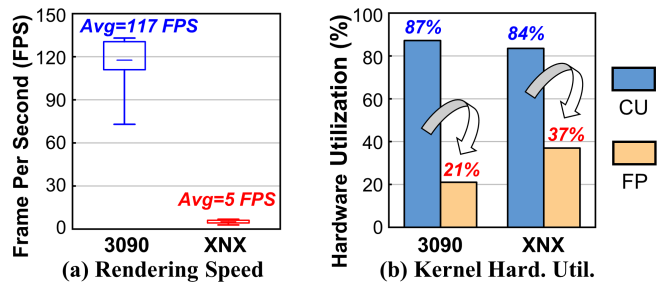


Fig. 1: (a) Rendering speed and (b) hardware utilization of the rendering kernel in vanilla 3DGS, profiled on RTX 3090 [8] and Jetson Xavier NX [9]. In (b), CU denotes the utilization of compute units (i.e., GPU SMs), reflecting overall computation activity, while FP indicates the achieved FP32 performance relative to the device peak performance.

sians by considering more Gaussian features. These features help narrow the candidate contribution region of Gaussians, thereby reducing the pixels that unnecessarily process them. Nevertheless, striving to make the candidate region exactly match the true contribution area will inevitably incur significant computational complexity. Instead, [13] adopts a GPU-based Contribution-Aware Test (CAT) by directly evaluating each Gaussian’s contribution to a leader pixel within a pixel group before rendering. If the contribution is negligible, the entire pixel group can skip processing that Gaussian. This method alone achieves a  $1.3\times$  speedup for the overall system.

Despite its effectiveness, applying this contribution-aware strategy to edge designs faces **several challenges**. **First**, the overhead of CAT is substantial. For instance, the  $2\times 2$  pixel group size adopted in [13] incurs significant computational overhead, as each four-pixel group requires a contribution test of a leader pixel, and the total number of such tests scales linearly with image resolution, making it prohibitively expensive for edge devices. **Second**, enabling independent Gaussian skipping for each pixel group requires dedicated memory allocation per group, which will lead to significant on-chip memory overhead. **Third**, the leader-pixel-based approach makes rendering quality highly sensitive to pixel group size. Simply reducing the number of leader pixels by adopting larger pixel groups will significantly increase the risk of missing contributing Gaussians, resulting in noticeable image degradation.

To address the above challenge, this work make the following contribution:

\* Corresponding author.

- We propose FLICKER, a fine-grained contribution-aware accelerator that leverages hardware-software co-design to achieve accurate skipping of non-contributing Gaussians at nearly pixel-scale granularity, facilitating real-time 3DGS rendering on the edge.
- We introduce an adaptive leader-pixel scheme that dynamically reduces the number of leader pixels based on Gaussian shapes. Furthermore, we propose a novel batch processing technique that organizes leader pixels into rectangular groups. By sharing intermediate results within each group, the overhead is nearly halved without compromising image quality (Sec. III).
- We introduce a two-stage hierarchical testing flow that effectively filters Gaussians with reduced computation and memory overhead. Moreover, we design a mixed-precision CAT engine, tightly integrated with the rendering pipeline, to minimize area overhead and effectively hide CAT latency (Sec. IV).
- Experimental results show that FLICKER achieves up to  $1.5\times$  speedup and  $2.6\times$  energy efficiency compared to the baseline design, while requiring 14% less area (Sec. V).

## II. BACKGROUND AND MOTIVATION

### A. 3DGS with Contribution-Aware Rendering

**3DGS Rendering Pipeline.** The 3DGS rendering process begins with a set of Gaussian ellipsoids defined by differentiable parameters. For a given camera pose, generating an image frame from these Gaussians involves three main steps, as shown in Fig. 2(a). In **Step(1)**, Gaussians within the view frustum are projected onto the image plane, generating 2D features such as mean ( $\mu'$ ), covariance ( $\Sigma'$ ), and color ( $c'$ ). Since rendering proceeds tile by tile, an intersection test is performed for each tile to identify Gaussians that may contribute. These Gaussians are then copied as needed to form a dedicated list for each tile. In **Step(2)**, the Gaussians in each list are then sorted by their distance from the camera (i.e., depth), arranged from near to far. With the sorted list (containing  $N$  Gaussians), all pixels within a tile are rendered in a uniform manner by iterating over the Gaussians in the list (**Step(3)**). For each Gaussian  $G_i$ , the "contribution" to pixels is first computed:

$$\alpha_i = o_i \cdot e^{-\frac{1}{2}(p-\mu'_i)^\top \Sigma_i'^{-1}(p-\mu'_i)} \quad (1)$$

where  $o_i \in [0, 1]$  is the opacity and  $p$  is the pixel coordinate. Gaussians with  $\alpha_i < \frac{1}{255}$  are considered no contribution and skipped. Otherwise,  $\alpha_i$  is further used to compute the pixel color  $c = \sum_i^N T_i c'_i \alpha_i$ . With transmittance defined as  $T_i = \prod_{j=1}^{i-1} (1 - \alpha_j)$ , rendering of the current tile can terminate early if the transmittance of all pixels falls below a predefined threshold.

**Bounding Box Intersection Test.** Over-inclusive intersection tests impose significant rendering overhead. As shown in Fig. 2(b), vanilla 3DGS employs a simple Axis-Aligned Bounding Box (AABB) test [2]. For Gaussians with anisotropic shapes, the  $3\sigma$  rule defines their effective boundaries, which are then replaced by a bounding box aligned with the coordinate

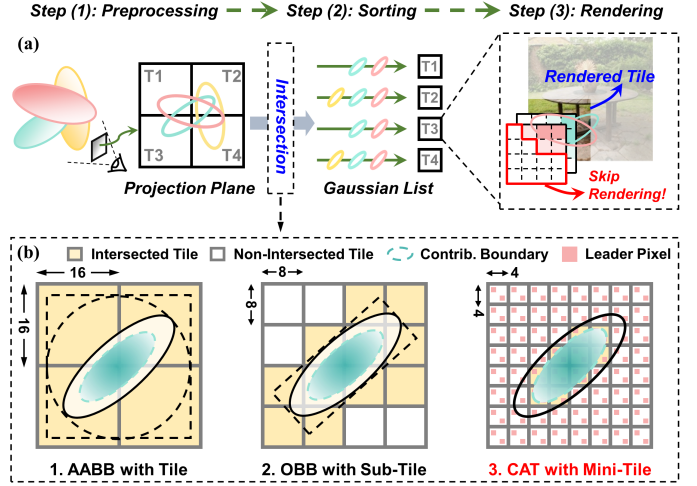


Fig. 2: (a) Overall rendering pipeline of 3DGS [2], and (b) comparison of three intersection methods—AABB in vanilla 3DGS, OBB in GScore [6] and proposed Mini-Tile CAT.

axes. In this toy example, the AABB test marks all tiles as intersected, leading to substantial redundant computation. GScore [6], an ASIC-level 3DGS accelerator, adopts an Oriented Bounding Box (OBB) technique to better fit the anisotropic shapes of Gaussians. Moreover, by dividing tiles into sub-tiles with  $8 \times 8$  pixels, the intersected region is significantly reduced. Nevertheless, the region still does not precisely match the Gaussian's actual contribution.

**Mini-Tile Contribution-Aware Test.** Building upon the CAT [13] discussed in Sec. I, we propose Mini-Tile CAT, which uses multiple leader pixels to accurately capture Gaussian contributions within a  $4 \times 4$  mini-tile. A mini-tile is marked intersected if any leader pixels is contributed by the Gaussian. Combined with customized optimization, this method can enable accurate intersection while largely reducing CAT overhead (details in Sec. III). As shown in Fig. 2(b), the intersected region closely aligns with the Gaussian's true contribution boundary, thereby ensuring higher rendering efficiency.

### B. 3DGS Profiling and Strategy Analysis

We adopt a two-stage analysis method to investigate performance bottlenecks and identify optimization opportunities: **First**, we profile the 3DGS rendering pipeline on two GPUs: a desktop-level GPU, RTX 3090 [8], and an edge GPU, XNX [9]. Since XNX shares similar hardware specifications to edge GPUs used in advanced VR headsets [14], its profiling results can better reflect the bottlenecks in practical edge scenarios. We obtain a detailed GPU performance breakdown using Nsight Compute [15] and profiling is conducted with datasets from Mip-NeRF360 [16]. **Second**, we further conduct an in-depth analysis of the introduced intersection methods to quantify the expected benefits and overheads of the Mini-Tile CAT, thereby guiding the subsequent design.

**3DGS Profiling Result.** As shown in Fig. 1(a), 3DGS achieves real-time rendering on desktop GPUs, with average FPS exceeding 100. In contrast, on edge GPUs, the FPS drops

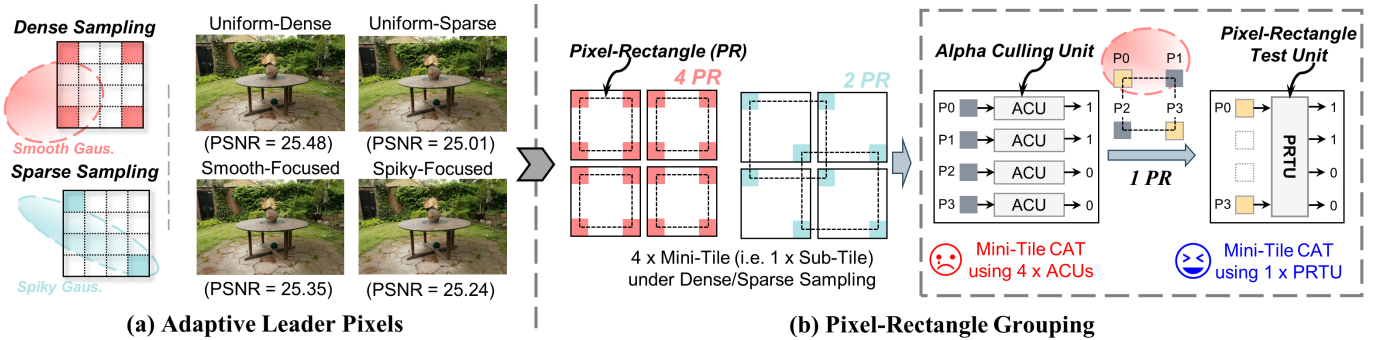


Fig. 3: Mini-Tile CAT algorithm optimization: (a) adaptive leader pixels, and (b) pixel-rectangle grouping. In (a), the PSNR of vanilla 3DGS is 25.56, while the Uniform-Dense mode shows negligible loss. Although smooth Gaussians account for 43%, the Smooth-Focused mode achieves higher PSNR, indicating that the contribution of smooth Gaussians is more significant in this case. In (b), the Mini-Tile CAT of a Pixel Rectangle (PR) can be simplified by exploiting its coordinate symmetry.

sharply to around 5 per scene, highlighting such edge devices struggle to handle the 3DGS’s workload. To understand this discrepancy, we further analyzed the hardware utilization of the rendering step in 3DGS, which accounts for a significant portion of GPU kernel execution time, often exceeding 60% [6], [17], [18]. As shown in Fig. 1(b), the overall compute unit utilization on the GPU, calculated via SM Core throughput [19], reaches an average of 85%. However, the average floating-point utilization is merely 29%, which corresponds to the core computations in the rendering step. The result indicates that the already limited computation resources on edge GPUs are underutilized, exacerbating performance degradation. This inefficiency stems from the fact that, during rendering, certain pixels within the same tile are skipped due to their negligible Gaussian contributions, which causes warp divergence. Moreover, the over-inclusive intersection test further amplifies this effect, ultimately leading to poor performance.

**Strategy Analysis.** Based on profiling, we further compare introduced intersection methods on a real-world scene to guide Mini-Tile CAT implementation on edge hardware. In this analysis, each mini-tile is initially assigned 4 leader pixels. The results, shown in Fig. 4, lead to the following observations:

**First,** Mini-Tile CAT demonstrates a clear advantage in filtering Gaussians at the fine-grained  $4 \times 4$  tile level, i.e., mini-tile. Compared to AABB on a  $16 \times 16$  tile, Mini-Tile CAT reduces the number of Gaussians each pixel must process to only 10%, the lowest among all tested methods. This underscores its potential to substantially cut rendering workload and alleviate underutilization of parallel hardware units caused by pixel skipping. **Second,** although Mini-Tile CAT reduces unnecessary Gaussian processing, it incurs significant computational overhead because leader pixels must be tested on many Gaussians. Moreover, since Mini-Tile CAT evaluates far

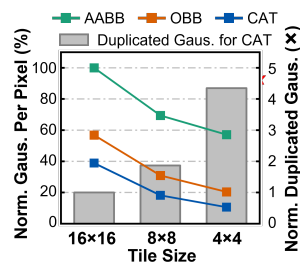


Fig. 4: Per-pixel processed Gaussians across intersection strategies and duplicate Gaussians across tile sizes.

more Gaussians than are ultimately forwarded to the rendering stage, it can easily become a pipeline bottleneck, causing downstream stalls. This highlights the need for an efficient architectural support of Mini-Tile CAT. **Third,** smaller tile sizes more effectively reduce redundant Gaussian processing but at the cost of higher memory overhead due to increased duplicated Gaussians. For instance, reducing the tile size from  $16 \times 16$  to  $4 \times 4$  pixels increases the total number of duplicated Gaussians to  $4 \times$  the original, highlighting the need for an efficient hierarchical intersection strategy and memory allocation.

### III. ALGORITHM OPTIMIZATION

In this section, we introduce two algorithmic strategies to effectively reduce the computation overhead of Mini-Tile CAT from two perspectives: reducing (1) the number of required leader pixels, and (2) the per-leader-pixel CAT overhead.

#### A. Adaptive Leader Pixels

**Basic Mode.** We begin by defining two sampling settings for Mini-Tile CAT: *Dense Sampling*, using four corner pixels per mini-tile, and *Sparse Sampling*, using two diagonal pixels. From these, we can derive two modes for Mini-Tile CAT: *Uniform-Dense*, where all Gaussians use Dense Sampling, and *Uniform-Sparse*, where all use Sparse Sampling. As shown in Fig. 3(a), Uniform-Dense captures most contributing Gaussians with only a negligible PSNR drop (0.08 dB) compared to vanilla 3DGS, while Uniform-Sparse, though reducing leader pixels by half, causes noticeable quality degradation.

**Adaptive Mode.** To combine the strengths of both uniform modes, we propose an adaptive strategy that dynamically switches the sampling based on Gaussian shape. All Gaussians are first classified as *Smooth* (axis ratio  $< 3$ ) or *Spiky* (axis ratio  $\geq 3$ ). Smooth Gaussians use Dense Sampling for broader coverage (*Smooth-Focused* mode), while spiky Gaussians use Sparse Sampling to save leader pixels. As shown in Fig. 3(a), this adaptive mode reduces PSNR loss by 73% compared to Uniform-Sparse, while retaining 57% of its leader-pixel savings. Notably, if spiky Gaussians carry more critical visual details, the strategy can be switched to a *Spiky-Focused* mode, in which Dense Sampling is applied to spiky Gaussians.

## B. Pixel-Rectangle Grouping

**CAT for Single Pixels.** As discussed in Sec. II, the contribution of a Gaussian to a leader pixel is calculated using Eq. 1. The result is then compared to a threshold (e.g.,  $\alpha < \frac{1}{255}$ ) to determine whether the Gaussian can be skipped. A straightforward way for CAT is to design a dedicated Alpha Culling Unit (ACU) [6], [17], [18] to test each pixel individually. However, this approach incurs significant computation overhead, especially when testing multiple leader pixels across dense mini-tiles.

**CAT for Pixel Rectangles.** To reduce the average CAT overhead for per leader pixel, we first simplify the Eq. 1 as follows:

$$\ln(255 \cdot o) > -\frac{1}{2}(p - \mu')^\top \Sigma'^{-1}(p - \mu') \quad (2)$$

In the inequality, the left-hand side term,  $\ln(255 \cdot o)$ , is identical for all leader pixels tested against the same Gaussian, and therefore only needs to be computed once and shared. To reduce the overhead for computing the right-hand side, i.e., the Gaussian Weight  $E$ , we propose a *Pixel-Rectangle Test Unit (PRTU)*, which evaluates the contribution of a Gaussian to a group of four leader pixels arranged in a rectangular *Pixel Rectangle (PR)*. Within each PR, the two off-diagonal corner pixels have symmetric coordinates relative to the main-diagonal pixels, allowing the intermediate results from the main-diagonal pixels to be reused for computing the off-diagonal pixels. The pseudocode for processing a PR is provided in Alg. 1.

### Algorithm 1 Pixel-Rectangle Gaussian Weight Computation

**Input:** Gaussian mean  $\mu'$ , conic entries  $\Sigma'_{xx}, \Sigma'_{yy}, \Sigma'_{xy}$ , Main diagonal pixel coordinates  $\mathbf{p}_{\text{top}}, \mathbf{p}_{\text{bot}}$  (correspond to  $p_0$  and  $p_3$  in a PR).

**Output:** Gaussian weight  $E_0, E_1, E_2, E_3$

- 1:  $\Delta_{\text{top}} \leftarrow \mathbf{p}_{\text{top}} - \mu'$      $\Delta_{\text{bot}} \leftarrow \mathbf{p}_{\text{bot}} - \mu'$
- 2:  $s_{\text{top}}^x = 0.5 \cdot \Delta_{\text{top},x}^2 \cdot \Sigma'_{xx}$      $s_{\text{top}}^y = 0.5 \cdot \Delta_{\text{top},y}^2 \cdot \Sigma'_{yy}$
- 3:  $s_{\text{bot}}^x = 0.5 \cdot \Delta_{\text{bot},x}^2 \cdot \Sigma'_{xx}$      $s_{\text{bot}}^y = 0.5 \cdot \Delta_{\text{bot},y}^2 \cdot \Sigma'_{yy}$
- 4:  $t_0 = \Delta_{\text{top},x} \cdot \Delta_{\text{top},y} \cdot \Sigma'_{xy}$      $t_1 = \Delta_{\text{bot},x} \cdot \Delta_{\text{top},y} \cdot \Sigma'_{xy}$
- 5:  $t_2 = \Delta_{\text{top},x} \cdot \Delta_{\text{bot},y} \cdot \Sigma'_{xy}$      $t_3 = \Delta_{\text{bot},x} \cdot \Delta_{\text{bot},y} \cdot \Sigma'_{xy}$
- 6:  $E_0 = s_{\text{top}}^x + s_{\text{top}}^y + t_0$      $E_1 = s_{\text{bot}}^x + s_{\text{top}}^y + t_1$
- 7:  $E_2 = s_{\text{top}}^x + s_{\text{bot}}^y + t_2$      $E_3 = s_{\text{bot}}^x + s_{\text{bot}}^y + t_3$

Compared to the ACU that tests pixels individually, our pixel-rectangle grouping method nearly halves the computation cost. Most importantly, it can be effectively combined with our adaptive leader-pixel strategy. As shown in Fig. 3(b), a sub-tile composed of four mini-tiles typically includes multiple PRs: in Dense Sampling, each mini-tile contributes one PR, resulting in four PRs per sub-tile, whereas Sparse Sampling still can form two valid PRs across mini-tiles.

## IV. HARDWARE ARCHITECTURE

We begin with an overview of the FLICKER architecture, then detail the hierarchical Gaussian testing and contribution-aware rendering pipeline, which enables high-throughput Mini-Tile CAT through dedicated architectural support. Finally, we introduce the mixed-precision contribution-aware test unit

(CTU) for applying Mini-Tile CAT with minimal hardware overhead.

### A. Overall Architecture

**Main Components.** As shown in Fig. 5, the architecture consists of four main components: preprocessing core, sorting unit, CTU, and rendering core. The preprocessing core projects 3D Gaussian features into 2D, determines whether Gaussians fall within the frustum, classifies them as spiky or smooth, and performs AABB tests for sub-tile intersections. The sorting unit fetches the converted features, sorts them by depth, and forwards them to the CTU. The CTU applies Mini-Tile CAT to filter sorted Gaussians according to their contribution. Finally, the rendering core completes the rendering step using the Gaussians that pass CTU.

**Memory Access Optimization.** Since the number of Gaussians is extremely large, most parameters must be stored off-chip. To reduce DDR traffic, we adopt a clustering method that groups multiple Gaussians into larger “big Gaussians” [17]. Frustum culling is then performed on these big Gaussians instead of on individual ones, significantly reducing the number of DDR accesses for the preprocessing core. Moreover, the bandwidth efficiency is further improved by loading only geometric features (10 parameters) during culling, while color features (45 parameters) and other parameters are fetched only for Gaussians that pass frustum culling and intersection test.

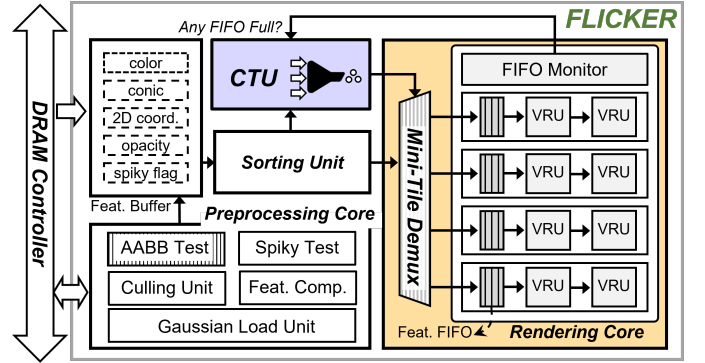


Fig. 5: Overall hardware architecture of FLICKER. The key component, the contribution-aware test unit (CTU), is highlighted in purple and will be detailed in Sec. IV-C.

### B. Hierarchical Gaussian Testing and Contribution-Aware Rendering Pipeline

**Hierarchical Testing.** As discussed in Sec. III, Mini-Tile CAT reduces per-pixel overhead, but the number of Gaussians to test remains high. To handle this, we introduce a two-stage hierarchical testing strategy (Fig. 6). **Stage 1:** In the preprocessing core, a sub-tile AABB test is performed. Gaussians are duplicated into feature buffers according to their sub-tile intersection mask, enabling efficient skipping at the sub-tile level while reducing the CTU workload (by 30%, as shown in Fig. 4). **Stage 2:** The CTU processes Gaussians that pass Stage 1 by applying Mini-Tile CAT to generate fine-grained masks. Based on these masks, the contributing Gaussians are then duplicated into the corresponding FIFOs in the rendering

core. Each FIFO drives two VRUs, which together render 16 pixels—exactly one mini-tile. Four such channels within a rendering core cover one sub-tile, while the four rendering cores in FLICKER collectively span a full tile. Since hierarchical testing reduces the Gaussian count to about 10% of the original, the required FIFO capacity is small, which in turn lowers memory overhead. Overall, this organization enables efficient and fine-grained mini-tile skipping under the tile level.

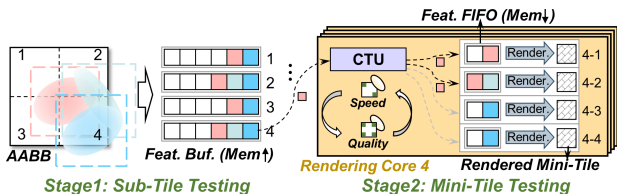


Fig. 6: Hierarchical Gaussian testing.

**Contribution-Aware Rendering Pipeline.** Beyond hierarchical testing, we further optimize the runtime pipeline to ensure smooth execution. With the dedicated CTU for Mini-Tile CAT, most of its latency is hidden by overlapping with VRU rendering. To further reduce FIFO capacity, we design a stall-resilient pipeline. When any FIFO inside the rendering core becomes full, a FIFO monitor detects the stall and notifies the CTU (Fig. 5). Upon receiving the stall signal, the CTU halts the intake of new Gaussians, while the in-flight pipeline results are safely stored in a small built-in CTU FIFO, ensuring no data loss despite its fully pipelined nature. As validated in Sec. V-B, this design allows very shallow FIFOs to achieve most of the speedup provided by mini-tile skipping. Instead, if CTU throughput falls behind the VRUs, the system can switch to Uniform-Sparse mode, boosting Mini-Tile CAT throughput.

### C. Mixed-Precision Contribution-Aware Test Unit

As shown in Fig. 7(a), the mixed-precision CTU architecture consists of two PRTUs, a Mask Merge Unit (MMU), and units for computing the shared term  $\ln(255 \cdot o)$ . The CTU is fully pipelined and can process two PRs (total 8 leader pixels) per cycle, with each PRTU handling one PR. The controller dynamically adjusts the sampling mode based on the Gaussian spiky flag. For Sparse Sampling, the two PRTUs directly generate test masks for two PRs, which are then merged by the MMU and output. For Dense Sampling, four PRs are processed in two batches: the mask from the first batch is stored in registers, and after the second batch completes, the MMU merges both batches to produce the final output, as illustrated in Fig. 7(b).

To reduce the hardware overhead of the CTU, we employ a mixed-precision PRTU: differences between pixel and Gaussian coordinates (line 1 in Alg. 1) are computed in FP16, and the results are then converted to FP8 for subsequent calculations in the Quarda Accumulation Unit (lines 2–7). We evaluate three precision schemes: Full FP16, Full FP8, and mixed precision. As shown in Fig. 7(c), the mixed precision scheme maintains high image quality, whereas Full FP8 suffers from severe PSNR degradation and noticeable blocky artifacts. The degradation primarily arises from the compression of relative

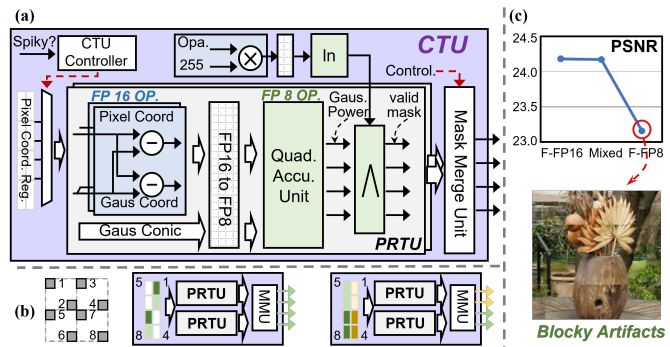


Fig. 7: Mixed-precision contribution test unit: (a) microarchitecture, (b) dataflow for the adaptive leader pixel test, and (c) comparison across different precision schemes.

TABLE I: Evaluation of rendering quality (PSNR $\uparrow$  and SSIM $\uparrow$ ) across different approaches

	Tanks & Temples [23]		MipNeRF360 (outdoor) [16]		Deep Blending [24]		Average
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
Base.	24.08	0.86	25.88	0.76	29.72	0.90	26.56
Prun.	23.61	0.84	24.71	0.73	29.64	0.90	25.99
Ours	23.51	0.84	24.52	0.73	29.62	0.90	25.88

positional information between pixels and Gaussians, which leads to the loss of interpolation details. In contrast, our mixed precision scheme preserves critical interpolation information and leverages the inherent error tolerance of Mini-Tile CAT, finally ensuring quality with low hardware cost.

## V. EVALUATION

### A. Experimental Settings

**Algorithm Setup.** We evaluate on eight real-world scenes: two outdoor scenes from Tanks & Temples, four outdoor scenes from Mip-NeRF 360, and two indoor scenes from Deep Blending. Each scene is first trained with vanilla 3DGS [2] for 30K iterations to obtain baseline models. To produce more compact models, we apply a pruning technique [20], which removes Gaussians with negligible contribution, followed by an additional 3K fine-tuning iterations. After pruning, we adopt the clustering method [17] to group Gaussians into clusters. Training is performed in FP32, and parameters are then quantized for full FP16 rendering on FLICKER.

**Hardware Setup.** The proposed accelerator comprises  $4 \times (4 \times 2)$  VRUs (4 rendering cores), 4 CTUs, 4 sorting units, and 4 preprocessing cores. The design is developed in Verilog and synthesized with Synopsys Design Compiler using the TSMC 28nm process, with SRAMs generated via the memory compiler. To evaluate performance, we build a cycle-accurate simulator of FLICKER, including an LPDDR4 memory with 51.2 GB/s bandwidth, and estimate DRAM energy following [21] [22]. For comparison, we use GSCore [6] and the Jetson XNX GPU [9] as baselines. In addition, we build a simplified version of FLICKER without the CTU to assess its impact.

### B. Critical Component Analysis

Fig. 8 presents the normalized speedup and energy savings obtained from employing the CTU. To highlight its contribution, we evaluate on the baseline model without other

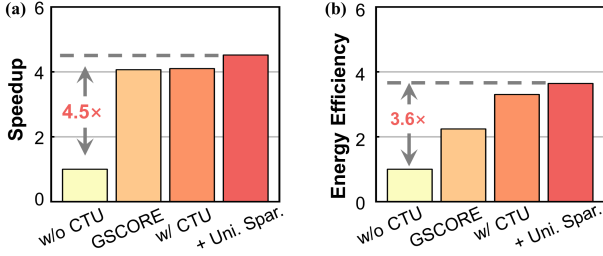


Fig. 8: Comparison of (a) speedup and (b) energy efficiency for the rendering stage. Note that GSCore [6] is configured with 64 VRUs, while ours uses 32 VRUs for smaller area. The evaluation is performed on the scene *Garden* only.

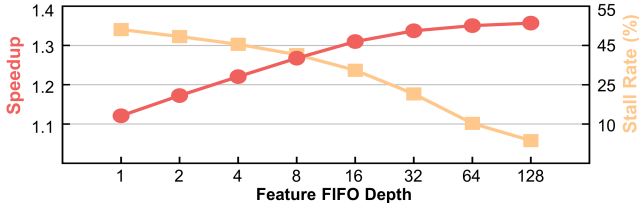


Fig. 9: The sensitivity of speedup and CTU stall rate to the depth of the feature FIFO in the rendering stage. The evaluation is performed on the scene *Garden* only.

optimizations and focus solely on the rendering stage. In terms of speedup, the simplified version of FLICKER is  $4\times$  slower than GSCore, primarily because it only adopts a basic AABB test, whereas GSCore employs an OBB test [6] and doubles the number of VRUs (32 vs. 64). Integrating the CTU improves performance to  $4\times$  over the baseline by enabling accurate mini-tile Gaussian skipping. Even with fewer VRUs, FLICKER still matches the rendering speedup of GSCore. Further configuring the CTU in Uniform-Sparse mode yields an additional  $1.1\times$  speedup, as it helps efficiently skip large groups of non-contributing Gaussians that would otherwise stall the VRUs. For energy efficiency, our design achieves up to  $1.6\times$  energy savings over GSCore, as it prevents the massive VRUs from wasting energy on non-contributing Gaussians.

To quantify the impact of FIFO depth on CTU stalls (when FIFO is full), we evaluate the rendering-stage speedup of FLICKER across depths from 1 to 128 (Fig. 9) and the corresponding CTU stall rates. Results show that increasing FIFO depth reduces stalls and improves speedup, reaching a maximum of  $1.36\times$  at depth 128. However, returns diminish beyond a depth of 16, which already achieves 96% of the maximum speedup while using only 12.5% of the memory compared to depth 128. Therefore, we select a FIFO depth of 16 for configuration. This highlights the effectiveness of hierarchical testing, which enables mini-tile skipping with shallow FIFOs rather than large sub-tile buffers and achieves most of the performance gain with minimal memory overhead.

### C. Overall System Evaluation

Tbl. I compares the rendering image quality across different methods. Ours incurs only an average PSNR loss of 0.11 dB compared to the pruning model, demonstrating that our adaptive leader pixel strategy is effective in capturing most contributing Gaussians and preserving visual quality.

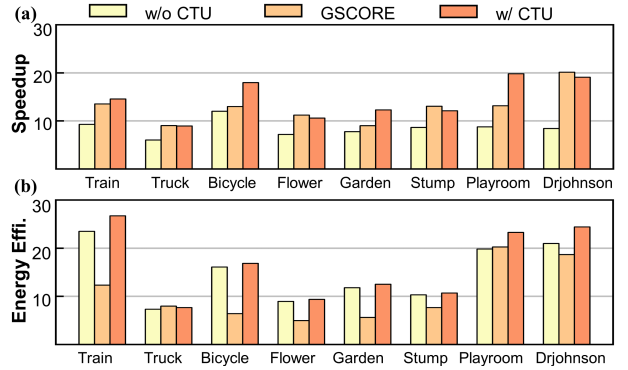


Fig. 10: Overall (a) speedup and (b) energy efficiency, normalized to the GPU baseline.

Component	Config.	Area [mm <sup>2</sup> ]
Preprocessing Core	4	0.76
Sorting Unit	4	0.16
Contri. Test Unit	4	0.09
Rendering Core	4x(4x2)	0.96
Fea. Buffer+Others	288KB	1.50
<b>Total</b>		<b>3.47</b>

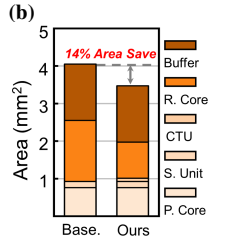


TABLE II: (a) Hardware configuration and area breakdown. (b) Area comparison between baseline and our design.

Fig. 10 shows the system performance over the baseline, with all values normalized to the XNX. Integrating CTU while adopting existing optimizations (pruning and clustering), FLICKER achieves average  $1.1\times$  speedup over GSCore and  $14.4\times$  over XNX. Furthermore, FLICKER consistently achieves the highest efficiency across all dataset, with maximum  $2.6\times$  up to GSCore and  $26.7\times$  compared to XNX. This demonstrates FLICKER’s capability to enable real-time 3DGS rendering for edge applications, and its compatibility with existing optimization techniques.

Tbl. II(a) reports the area breakdown of FLICKER. Thanks to the mixed-precision architecture and pixel-rectangle grouping, the CTU occupies less than 10% of the VRUs area (rendering core), yet delivers up to  $2.3\times$  overall speedup, which is difficult to achieve by merely adding more VRUs. We further extend the simplified version from 32 VRUs to 64 VRUs to emulate GSCore’s configuration as baseline in terms of VRU count. As shown in Tbl. II(b), although the CTU and feature FIFO introduce minor additional area, they enable a more efficient area allocation than using more VRUs, ultimately achieving 14% total area savings.

## VI. CONCLUSION

This paper introduces FLICKER, a contribution-aware accelerator that focuses on reducing unnecessary Gaussian processing over merely scaling parallel hardware. It performs prior contribution test to accurately skip Gaussians at fine-grained pixel blocks before rendering, while leveraging software–hardware co-design to alleviate its associated overheads. Experimental results show that our design outperforms a state-of-the-art 3DGS accelerator and an edge GPU device across most real-world datasets, while incurring less area overhead.

## REFERENCES

- [1] Meta, "Introducing orion, our first true augmented reality glasses." <https://about.fb.com/news/2024/09/introducing-orion-our-first-true-augmented-reality-glasses/>, 2024.
- [2] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering.," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.
- [3] T. Wu, Y.-J. Yuan, L.-X. Zhang, J. Yang, Y.-P. Cao, L.-Q. Yan, and L. Gao, "Recent advances in 3d gaussian splatting," *Computational Visual Media*, vol. 10, no. 4, pp. 613–642, 2024.
- [4] W. Lin, Y. Feng, and Y. Zhu, "Metasapiens: Real-time neural rendering with efficiency-aware pruning and accelerated foveated rendering," in *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1*, pp. 669–682, 2025.
- [5] P. Papantonakis, G. Kopanas, B. Kerbl, A. Lanvin, and G. Drettakis, "Reducing the memory footprint of 3d gaussian splatting," *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, vol. 7, no. 1, pp. 1–17, 2024.
- [6] J. Lee, S. Lee, J. Lee, J. Park, and J. Sim, "Gscore: Efficient radiance field rendering via architectural support for 3d gaussian splatting," in *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, pp. 497–511, 2024.
- [7] Meta, "Meta quest 3 mixed reality headset." <https://www.meta.com/quest/quest-3/>, 2023.
- [8] NVIDIA, "Geforce rtx 3090 family." <https://www.nvidia.com/en-gb/geforce/graphics-cards/30-series/rtx-3090-3090ti/>, 2025.
- [9] NVIDIA, "Jetson xavier series." <https://www.nvidia.com/en-gb/autonomous-machines/embedded-systems/jetson-xavier-series/>, 2025.
- [10] A. Hanson, A. Tu, G. Lin, V. Singla, M. Zwicker, and T. Goldstein, "Speedy-splat: Fast 3d gaussian splatting with sparse pixels and sparse primitives," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21537–21546, 2025.
- [11] X. Wang, R. Yi, and L. Ma, "Adr-gaussian: Accelerating gaussian splatting with adaptive radius," in *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–10, 2024.
- [12] G. Feng, S. Chen, R. Fu, Z. Liao, Y. Wang, T. Liu, B. Hu, L. Xu, Z. Pei, H. Li, *et al.*, "Flashgs: Efficient 3d gaussian splatting for large-scale and high-resolution rendering," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26652–26662, 2025.
- [13] X. Huang, H. Zhu, Z. Liu, W. Lin, X. Liu, Z. He, J. Leng, M. Guo, and Y. Feng, "Seele: A unified acceleration framework for real-time gaussian splatting," *arXiv preprint arXiv:2503.05168*, 2025.
- [14] Y. K. Zhao, S. Wu, J. Zhang, S. Li, C. Li, and Y. C. Lin, "Instant-nerf: Instant on-device neural radiance field training via algorithm-accelerator co-designed near-memory processing," in *2023 60th ACM/IEEE Design Automation Conference (DAC)*, pp. 1–6, IEEE, 2023.
- [15] NVIDIA Corporation, "Nvidia nsight compute." <https://developer.nvidia.com/nsight-compute>, 2025.
- [16] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5470–5479, 2022.
- [17] J. Jo and J. Park, "Ps-gs: Group-wise parallel rendering with stage-wise complexity reductions for real-time 3d gaussian splatting," in *2025 Design, Automation & Test in Europe Conference (DATE)*, pp. 1–7, IEEE, 2025.
- [18] Y. Wang, Y. Li, J. Chen, J. Yu, and K. Wang, "Famers: An fpga accelerator for memory-efficient edge-rendered 3d gaussian splatting," in *2025 Design, Automation & Test in Europe Conference (DATE)*, pp. 1–7, IEEE, 2025.
- [19] Z. Jia, M. Maggioni, B. Staiger, and D. P. Scarpazza, "Dissecting the nvidia volta gpu architecture via microbenchmarking," *arXiv preprint arXiv:1804.06826*, 2018.
- [20] M. S. Ali, M. Qamar, S.-H. Bae, and E. Tartaglione, "Trimming the fat: Efficient compression of 3d gaussian splats through pruning," *arXiv preprint arXiv:2406.18214*, 2024.
- [21] P. Dong, Y. Tan, X. Liu, P. Luo, Y. Liu, L. Liang, Y. Zhou, D. Pang, M.-T. Yung, D. Zhang, *et al.*, "A 28nm 0.22  $\mu$ j/token memory-compute-intensity-aware cnn-transformer accelerator with hybrid-attention-based layer-fusion and cascaded pruning for semantic-segmentation," in *2025 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 68, pp. 01–03, IEEE, 2025.
- [22] K. Song, S. Lee, D. Kim, Y. Shim, S. Park, B. Ko, D. Hong, Y. Joo, W. Lee, Y. Cho, *et al.*, "A 1.1 v 2y-nm 4.35 gb/s/pin 8 gb lppdr4 mobile device with bandwidth improvement techniques," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 8, pp. 1945–1959, 2015.
- [23] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–13, 2017.
- [24] P. Hedman, J. Philip, T. Price, J.-M. Frahm, G. Drettakis, and G. Brostow, "Deep blending for free-viewpoint image-based rendering," *ACM Transactions on Graphics (ToG)*, vol. 37, no. 6, pp. 1–15, 2018.