# T2LF: LLM-Guided Multimodal Diffusion for Text-to-Light Field Synthesis - Supplementary Material -

Soyoung Yoon, Namhyuk Ahn, and In Kyu Park
Department of Electrical and Computer Engineering, Inha University
Incheon 22212, Korea
{thdud679@gmail.com, nhahn@inha.ac.kr, pik@inha.ac.kr}

## A. LLM Prompt Design

In this section, we describe the prompt design strategies that allow the large language model (LLM) to generate structured light field (LF) layouts and motion trajectories. Our approach leverages structured task instructions and in-context learning to guide the LLM in generating spatially coherent object layouts and depth-aware motion for LF synthesis.

### A.1. LF Layout Generation

The first stage of our method involves using the LLM to generate LF layouts from input text. To achieve this, we design structured prompts that guide the LLM to generate bounding boxes that define the image layout. The prompt format, as shown in Figure 1, consists of three components: a basic prompt summarizing the scene, detailed descriptions of sub-objects, and their positions (bounding boxes).

The effectiveness of this prompt design is further enhanced through in-context learning. The in-context examples used for layout generation correspond to the captions and bounding box annotations in Figure 3, which are curated from segmentation-based motion tracking A.3 performed on real-world LF data. By providing structured examples, LLM can learn realistic spatial relationships that are essential for LF scene synthesis.

The LLM-generated layouts are then used in the layout-guided diffusion model for LF scene generation. Through contextual reasoning with structured instructions and examples, LLM leverages the ability to understand the scene to maintain consistency between the described scene and the generated image, enabling accurate spatial alignment and object representation.

### A.2. LF Motion Generation

In the second stage of our method, LLM also functions as a motion generator for LF SAI synthesis. The inputs to this process include the layouts and the image generated in the previous step. Despite being layout-guided, images generated in previous stages may contain additional elements or may exhibit some structural inconsistencies due to model limitations.

To address these, the multimodal LLM refines the initial layouts based on the generated images and subsequently produces motion trajectories for each object in the scene. These motion trajectories are determined using structured task instructions, such as those presented in Figure 2, which include descriptions of the scene, object positions (bound-

```
Generate a structured representation of a
scene based on the given text prompt. Identify
 and list the key objects present in the scene
, providing their names and bounding box
coordinates. The bounding boxes should
indicate the position and size of each object
in the scene, normalized to a range of [0, 1],
 using the format [xmin, ymin, xmax, ymax].

Ensure that the description of the scene is
concise and provides a clear summary of the
layout and arrangement of objects. Include at
least one more background elements (e.g. walls
, roads, etc.) induced by text prompts.

The output format should follow this structure
:

{
  "basic_prompt": "A one-sentence description
  of the scene",
  "sub_objects": [
    {
      "object": "Name of the object",
      "bounding_box": [xmin, ymin, xmax, ymax]
    },
    ...
  ]
}
```

Figure 1. Task instruction for LF layout generation. The prompt includes a structured scene description, object details, and bounding box coordinates to guide the LLM in generating layouts.

```
You are an advanced vision assistant specializing in refining layouts and generating structured
motion trajectories. Your task is as follows:

  1. Refine and update the layout:
    - You are given a 512x512 image and a set of pre-existing layouts with objects and bounding
    boxes in normalized coordinates `[x_min, y_min, x_max, y_max]` (0 to 1 range).
    - Carefully analyze the entire image to detect every object.
    - Add object with its appropriate bounding box, if an object is present in the image but missing
     from the provided layout.
    - If a bounding box needs correction, modify it to fully enclose the object.

  2. Calculate Motion (Disparity per Frame):
    - For each object (both refined and newly identified):
      - Calculate the `disparity_per_frame` based on depth and camera perspective.
      - Use light field camera principles:
        - Closer objects move in the opposite direction of the camera (-).
        - Farther objects move in the same direction as the camera (+).
        - Zero disparity means no movement (0).
      - Provide clear, realistic reasoning for each object's disparity value.

  3. Output Format:
    - The result should be formatted as:
      {
        "basic_prompt": "A one-sentence description of the scene based on the provided image",
        "sub_objects": [
          {
            "object": "Name of the object",
            "bounding_box": [x_min, y_min, x_max, y_max],
            "disparity_per_frame": "Movement per frame reflecting depth"
          },
          ...
        ]
      }

[In-context examples]

User: {Image URL, Image Layout}
```

Figure 2. Task instruction for LF motion generation. The structured instructions guide the LLM in refining layouts and generating object motion trajectories based on LF principles.

ing boxes), and the object's movement per frame.

The LLM's ability to generate realistic motion trajectories is further enhanced through in-context learning shown in Figure 3. By providing examples of LF camera behavior and object interactions, the LLM learns to generalize these concepts and produce motion trajectories for new scenes. This approach ensures that the generated trajectories are consistent with camera perspectives, object depths, and real-world physics.

The LLM-generated motion contains frame-by-frame disparity that determines depth-based motion. These are added to the center of bounding box coordinates to create motion trajectories that match the LF's perspective. It ensures that the motion of objects across the frame is consistent with the spatial dynamics of the LF. The motion trajectory is then used as input for motion-guided video diffusion models. This integration ensures that the final LF is geometrically consistent and realistic, highlighting LLM's ability as an independent motion generator for LF applications.

### A.3. Segmentation-based Motion Tracking

To use LLM as a layout and motion generator for LF synthesis, it is important to provide high-quality in-context examples. These examples, derived from real LF data, help LLM determine the correlations between object movements and LF trajectories.

Segmentation-based motion tracking is applied to real LF data to construct these in-context examples. This process involves using segmentation masks to identify key objects in the scene and tracking their motion across frames relative to the center point of the mask in the first frame. By averaging the tracked motion frame-by-frame, it generates depth-aware disparity data for modeling LF trajectories. This approach serves as an effective method for estimating disparity in LF data that lack ground truth (GT) disparity annotations.

In addition to segmentation-based tracking, in-context

```
Example analysis:
  Image Description: A bowl on a table with various plants, oranges, and herbs in a vase.

  Reasoning:
    - The wooden table is closest to the camera and has the largest negative disparity(-1.375),
    moving significantly in the opposite direction.
    - The Green leafy plant 1 and orange bowl is moderately close with a smaller negative disparity
    (-0.75, -0.95), moving less than the table.
    - Green leafy plant 3 remains stationary with zero disparity (0.0), being at the focal plane.
    - Green leafy plant 2 is far from the camera and has positive disparity (1.5), showing moderate
    movement in the same direction.
    - The wall is the farthest object, moving the most in the same direction as the camera (2.875).

Output:
  {
    "basic_prompt": "A bowl on a table with various plants, oranges, and herbs in a vase.",
    "sub_objects": [
      {
        "object": "wooden table",
        "bounding_box": [0.0, 0.7812, 0.2167, 1.0],
        "disparity_per_frame": -1.375
      },
      {
        "object": "green leafy plant 1",
        "bounding_box": [0.0859, 0.4804, 0.332, 0.7265],
        "disparity_per_frame": -0.75
      },
      {
        "object": "wall",
        "bounding_box": [0.0, 0.0, 0.3457, 0.3457],
        "disparity_per_frame": 2.875
      },
      {
        "object": "green leafy plant 2",
        "bounding_box": [0.2969, 0.1836, 0.4882, 0.375],
        "disparity_per_frame": 1.5
      },
      {
        "object": "green leafy plant 3",
        "bounding_box": [0.4961, 0.3105, 0.7188, 0.5332],
        "disparity_per_frame": 0.0
      },
      {
        "object": "orange bowl",
        "bounding_box": [0.4492, 0.5332, 0.7773, 0.8613],
        "disparity_per_frame": -0.95
      }
    ]
  }
```

Figure 3. In-context examples for LF motion generation. This example illustrates how object movements are guided by depth-based disparity shifts.

examples include reasoning about LF camera motion. These examples illustrate how LF cameras capture disparity between different viewpoints, clarify the relationship between object depth and motion (e.g., negative disparity for close objects, positive disparity for far objects, and zero disparity for focal objects), and explain how camera motion interacts with objects in the scene.

By integrating segmentation-based motion tracking with camera motion reasoning, LLM obtains a comprehensive understanding of LF data. These curated examples improve LLM's ability to extract accurate motion trajectories and enable verification with reasoning of expected motion. As a result, LLM effectively supports LF synthesis by consistently generating motion vectors that align with LF trajectories.

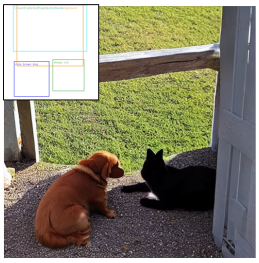
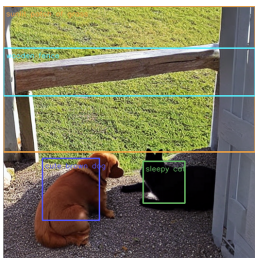
| Contextual Inputs (Initial Layout & Image) | VLM Reasoning and Refined Outputs |
|---|---|
|  |  The red tractor is close to the camera, hence a negative disparity (-0.75) indicating it moves in the opposite direction as the camera. The white farmhouse is farther back with a positive disparity (0.25), moving slightly in the same direction. The trees in the distant background is a larger positive disparity (0.75), indicating relatively far distance. The green lawn is the closest to the camera's perspective, exhibiting a more significant negative disparity (-0.5). |
|  |  The soft pink pillow has a negative disparity (-0.5), indicating it is fairly close to the camera and is moving slightly in the opposite direction. The round wooden stool also has a negative disparity (-0.75), suggesting it is positioned similarly but slightly closer to the camera. The curtain is even closer to the camera with a disparity (-1.0). The carpeted floor, with a disparity (-1.5), is the closest among the objects, following convention that closer positions indicate larger negative disparities. The wooden wall is farther back in the image, resulting in a positive disparity (1.0). |
|  |  The cute brown dog is closest to the camera, as indicated by its larger negative disparity (-1.125), meaning it moves significantly in the opposite direction of the camera. The sleepy cat, lying beside the dog, is slightly farther back, resulting in a smaller negative disparity (-0.75) as it also moves in the opposite direction. The sunlit patch of grass is farther from the camera, indicated by a larger positive disparity (0.625), meaning it moves in the same direction as the camera. The wooden fence is in the background but is slightly closer to the camera than the grass, reflected in its positive disparity value (0.25). |

Figure 4. Visualization of the VLM-guided layout and motion generation process for LF synthesis. The left column presents the contextual inputs provided to the VLM, including the initial layout and synthesized image. The right column shows the VLM's refined layout with updated bounding boxes and disparity values, alongside its reasoning statements. These reasoning statements enhance the interpretability of the VLM's disparity estimation process.

# B. Additional Results

## B.1. LF Layout and Motion Generation

To enhance transparency and interpretability, we visualize the intermediate outputs produced by the VLM during LF synthesis. Figure 4 shows the initial layout and synthesized image from the LF scene generation stage, along with the refined layout and motion generated during the LF motion generation. In parallel, we present the VLM's structured reasoning statements, which describe object-wise depth relationships and expected motions used to predict disparities.

For example, the VLM adds or refines objects such as a curtain, carpet, and fence, updating their positions with disparities consistent with the LF principles. The carpet, which is close to the camera, is assigned a large negative disparity (-1.5), while the wooden wall in the background receives a positive disparity (1.0), reflecting its farther distance. These visualizations demonstrate that the structured reasoning of the VLM provides transparency in how disparity values are estimated according to the principles of the LF camera, allowing verification that the outputs align with the depth geometry of the synthesized images and supporting the reliability of the VLM in achieving geometrically consistent, depth-aware LF motion synthesis.

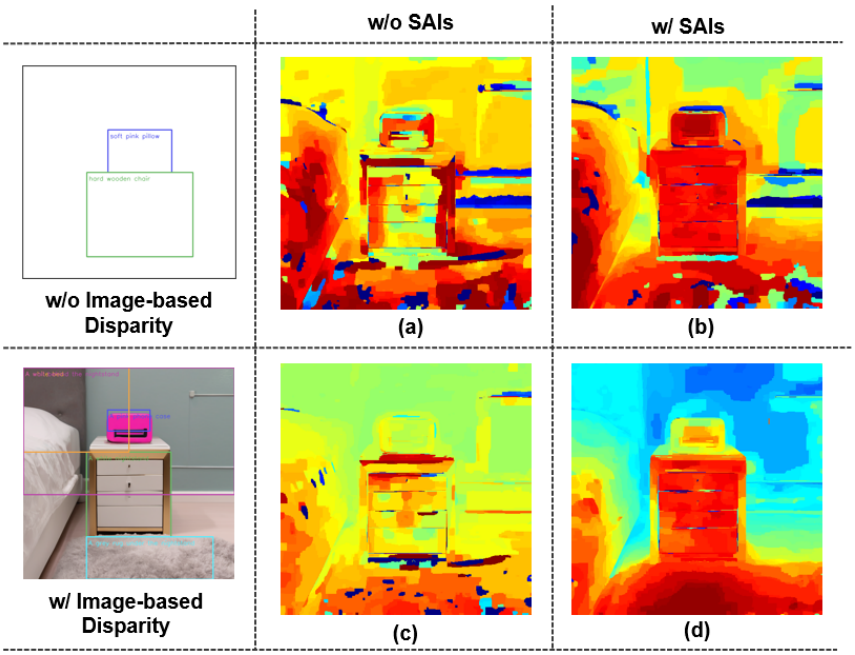



Figure 5. Qualitative analysis of ablation study on LF synthesis. (a) Removing both SAIs and image-guided disparity estimation. (b) Removing image-guided disparity, (c) Removing SAIs (Disparity-only setting), and (d) Our full model (T2LF), which integrates both components. The full model demonstrates the best geometric consistency by depth estimation results from synthesized LF.

## B.2. Qualitative Analysis of Ablation Study

To further analyze the impact of key components in our framework, we present qualitative comparisons in Figure 5. we analyze the depth estimation results from different ablation settings. This analysis evaluates how sub-aperture images (SAIs) and image-guided disparity contribute to geometric consistency in our T2LF framework.

**Effect of Image-Guided Disparity Estimation.** Image-guided disparity estimation improves depth by incorporating images generated from LF scene layout. Although the initial layout defines object placement, it lacks accurate depth information, which can be derived after image generation. Additionally, newly generated objects that were not originally present in the layout require additional depth constraints derived from the image itself. Without image-guided disparity estimation (Figure 5 (b)), inaccurate depth-aware motion occurs. For example, the pink phone case is incorrectly perceived as closer than the white nightstand. Furthermore, some objects present in the image, such as the bed and floor carpet, are not tracked, resulting in incomplete disparity supervision. In contrast, our full model (Figure 5 (d)) effectively integrates real image depth. As a result, the nightstand is correctly estimated to be closer than the pink phone case. Additionally, newly generated objects, such as the bed and carpet, receive appropriate depth constraints, aligning them with real-world depth relationships.

**Effect of SAI-Guided Motion Synthesis.** SAIs constraint LF synthesis by additional multiview guidance. In the absence of SAIs, the motion-guided video diffusion model relies solely on disparity information to generate motion, which may lead to unstable or misaligned object movements. As shown in Figure 5 (a), removing both SAIs and image-guided disparity estimation results in the most severe degradation, where even pre-difined objects, such as the nightstand and the pink phone case, fail to maintain stable motion. Similarly, removing only SAIs (Figure 5 (c)) leads to more accurate depth compared to (a), but the motion remains noisy and less stable due to the lack of additional geometric constraints. By incorporating both image-guided disparity estimation and SAI guidance, our full model (Figure 5 (d)) achieves the most depth-aware and stable motion synthesis. The nightstand and pink phone case exhibit accurate depth, while newly generated objects such as the bed and carpet are properly aligned with their real-world depth relationships.

Our analysis demonstrates that image-guided disparity estimation and SAI guidance are critical for achieving high-quality LF synthesis. Image-guided disparity estimation ensures that depth relationships are correctly inferred, preventing errors such as incorrect perception of object distances. Meanwhile, SAIs improve motion stability by constraining per-object shifts, reducing misalignment and noise. By integrating both components, our model achieves the most depth-aware and stable LF synthesis.

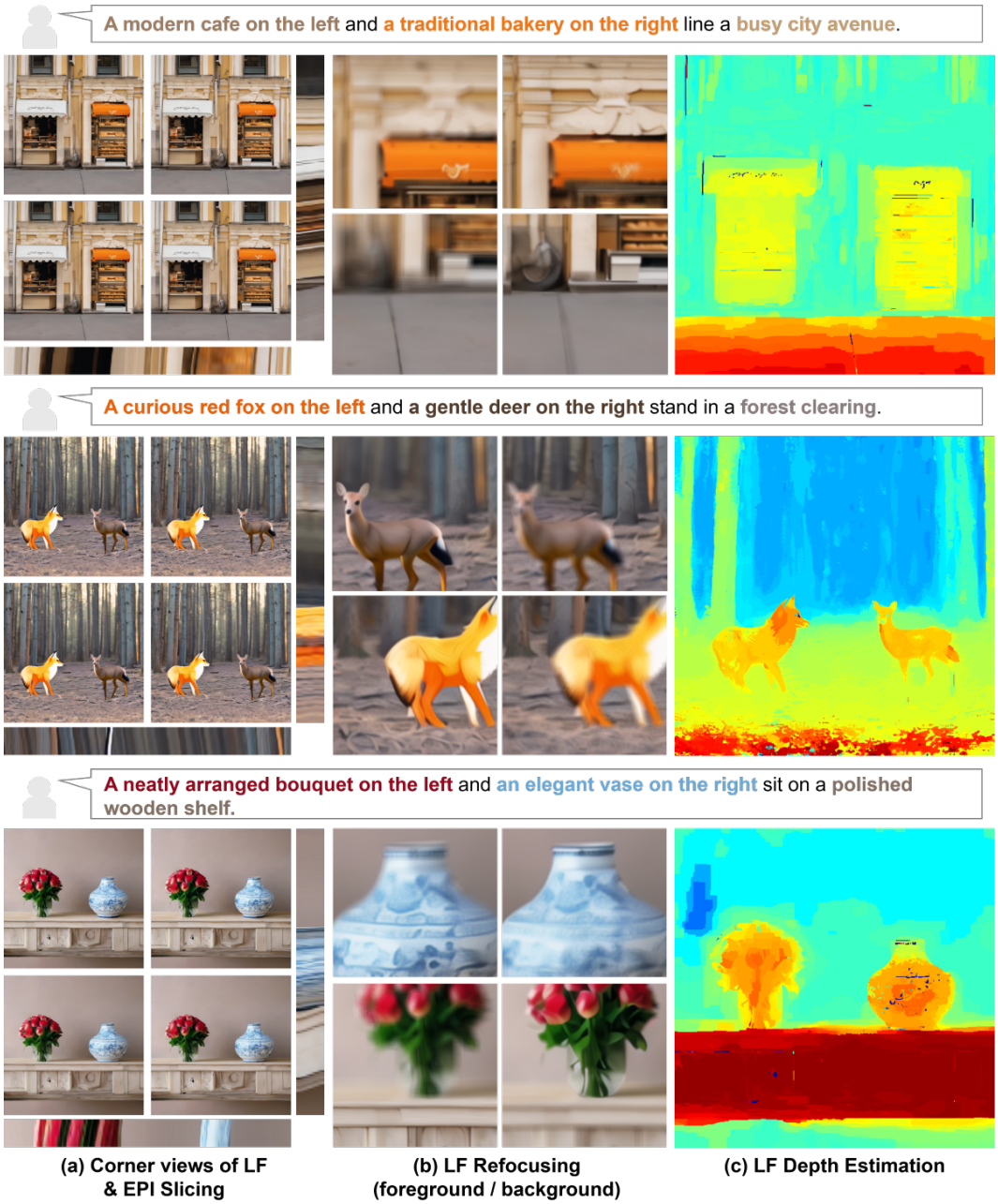



Figure 6. Additional LF synthesis results with complex structural prompts. (a) Corner views and EPI slicing, (b) Refocusing on the foreground and background, (c) Depth estimation from the synthesized LF.

### B.3. Additional Qualitative Results

To demonstrate the diversity and robustness of our method, we provide additional results for different text inputs in Figure 6 and Figure 7. Figure 6 presents results for complex structural prompts, where textual descriptions primarily define the placement and relationships of objects. Figure 7 contains both structural and stylized prompts, including artistic elements. The synthesized LFs effectively capture the scene described by the text prompt, as confirmed through the corner views of the LF.

The epipolar plane images (EPI) illustrate the consistency of the viewpoint shifts across the synthesized LF. Furthermore, we evaluated the refocusing and depth estimation from the synthesized LF, demonstrating that our method accurately preserves depth relationships. These results show that our model can synthesize LFs that accurately reflect

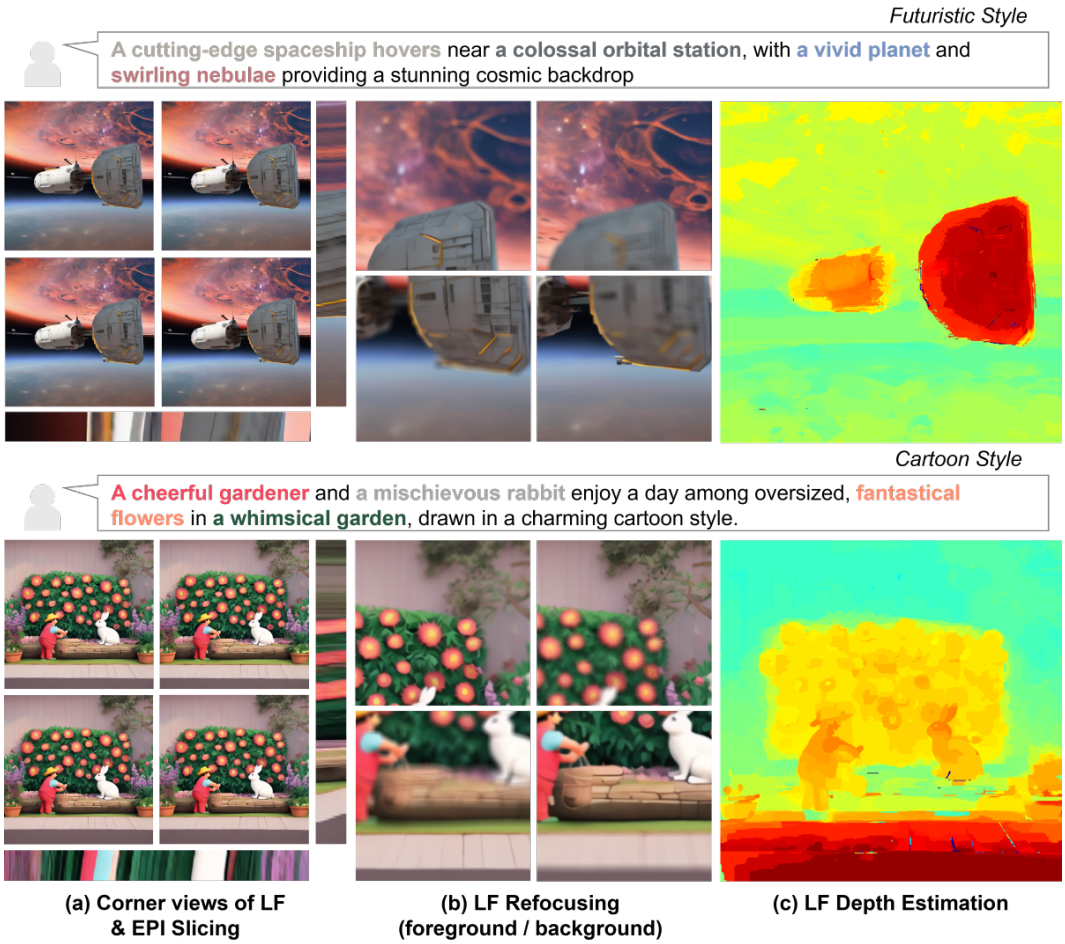



Figure 7. Additional LF synthesis results with structural and stylized prompts. (a) Corner views and EPI slicing, (b) Refocusing on the foreground and background, (c) Depth estimation from the synthesized LF.

spatial composition and artistic stylization while maintaining geometric consistency and depth-aware motion from multiple perspectives.

### B.4. Effect of LLM Choice

We further evaluate the impact of different multimodal LLMs used for layout and disparity inference in the T2LF pipeline. Table 1 compares GPT-4o with two open-source alternatives, LLaMA-3.2-Vision and Qwen2.5-VL. In general, GPT-4o achieves the strongest results in most metrics, particularly in multiview quality and LF consistency. Nevertheless, the open-source backbone also provides competitive performance. In particular, Qwen2.5-VL achieves the best BLIP-BLEU score (0.2309), suggesting stronger text-to-scene alignment, and performs on par with GPT-4o in LF consistency (NR-LFQA 2.995 and PPLC 0.1307). LLaMA-3.2-Vision shows similar overall performance. This shows that open-source models can extract sufficiently accurate layout and disparity cues from text or text + image inputs, resulting in LFs with stable quality and geometry.

While GPT-4o provides performance advantages, the results of LLaMA-3.2-Vision and Qwen2.5-VL clearly outperform the existing text-to-video and camera controllable baselines reported in Table 1 of the main paper. This highlights that the overall pipeline design and refinement steps play a central role in ensuring high LF quality, reducing sensitivity to LLM backbone selection. These findings demonstrate that our framework does not strictly rely on closed-source models, remains effective even on open-source alternatives, and provides reproducibility and flexibility while maintaining clear advantages over previous approaches.

## C. LoRA Fine-tuning and Inference Details

**LoRA fine-tuning setup.** We fine-tune DragAnything, a motion-guided video model based on Stable Video Diffusion, using LoRA on the spatio-temporal UNet. Training data consists of 20 scenes from the HCI-new dataset [1], 9

| Multimodal LLM Backbone | Multiview Quality | | Text-Video Alignment | Light Field Consistency | |
|---|---|---|---|---|---|
| | FID-VID ↓ | FVD ↓ | BLIP-BLEU ↑ | NR-LFQA ↑ | PPLC ↓ |
| LLaMA-3.2-Vision | 102.44 | <u>1403.51</u> | 0.2191 | **2.995** | 0.1343 |
| Qwen2.5-VL | <u>99.26</u> | 1614.89 | **0.2309** | **2.995** | <u>0.1306</u> |
| GPT-4o | **95.80** | **1362.58** | <u>0.2205</u> | **2.995** | **0.1267** |

Table 1. Ablation on multimodal LLM backbones for layout and disparity inference. The best results are shown in **bold** and the second-best are <u>underlined</u>.

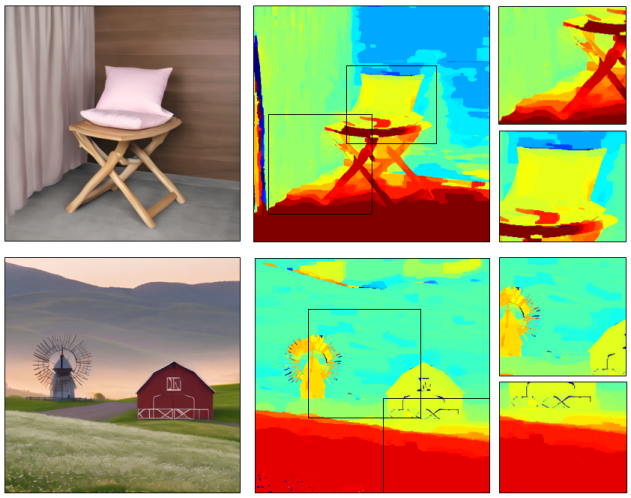


Figure 8. Limitation of per-object depth estimation of the multimodal LLM. Depth maps estimated from synthesized LFs reveal errors caused by assigning a single disparity per object and occasional disparity misestimation.

scenes from the real STFGantry dataset [2], and a subset of RealEstate10K [3] to diversify camera motions. Due to the relatively small HCI-new and STFGantry LF datasets, we increase data diversity by sampling different LF intervals and extracting 25 frame clips. This effectively augments the LF training set. The LoRA adapters are inserted into the attention projection layers (`to_q`, `to_k`, `to_v`, `to_out.0`) with rank $r = 4$ and $\alpha = 4$. We train for 1000 epochs using AdamW with learning rate $1 \times 10^{-4}$, weight decay $10^{-2}$, a constant scheduler with 500 warm-up steps, and batch size 1 per GPU. All experiments are performed on NVIDIA RTX A6000 GPUs.

**Inference setup.** During inference, we synthesize a $5 \times 5$ LF by generating 25 sub-aperture images at $512 \times 512$ resolution from each text prompt. We extract structured scene layouts and disparity information using GPT-4o and generate layout-guided images with GLIGEN. These images, together with the disparity cues, are then used for LF motion synthesis with the LoRA-tuned DragAnything model, where approximate SAIs are additionally produced to guide disparity-aware refinement. Inference is performed with 25 sampling steps. RePaint refinement is applied for 70% of the steps to ensure reliable disparity-based motion, while the remaining 30% are performed without RePaint, relying only on disparity cues to control the motion-guided model and produce natural viewpoint transitions.

## D. Limitations and Future Work

The proposed T2LF framework shows promising results for LF synthesis but still has several limitations. Currently, disparity reasoning is performed at the object level, which may oversimplify the depth representation of large or complex objects with spatially varying disparities. As illustrated in Figure 8, a single per-object disparity inference can flatten parallaxes in wide structures that span near and far depths such as curtains and fields. In addition, we sometimes observe near-far inversion between different objects, such as windmills and houses. Although motion-guided video diffusion helps alleviate this issue, the limitation remains and could be addressed in future work with more fine-grained depth representations. In addition, scalability is constrained by increasing runtime and memory requirements as LF resolution and the number of synthesized SAIs increase. The current architecture based on Stable Video Diffusion with motion adapters is memory intensive. Thus, scaling to higher resolutions or larger angular extents will require further optimization.

## References


[1] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A Dataset and Evaluation Methodology for Depth Estimation on 4D Light Fields. In *Proc. Asian Conference on Computer Vision (ACCV)*, pages 19–34, 2016. 7

[2] Vaibhav Vaish and Andrew Adams. The (new) Stanford Light Field Archive. `http://lightfield.stanford.edu/papers.html`, 2008. 8

[3] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo Magnification: Learning View Synthesis using Multiplane Images. *ACM Trans. on Graphics*, 37 (4):1–12, 2018. 8