Holographic Parallax Improves 3D Perceptual Realism - Supplementary Material

DONGYEON KIM* and SEUNG-WOO NAM*, Seoul National University, Republic of Korea
SUYEON CHOI*, Stanford University, USA
JONG-MO SEO, Seoul National University, Republic of Korea
GORDON WETZSTEIN, Stanford University, USA
YOONCHAN JEONG, Seoul National University, Republic of Korea

This is a supplementary material for 'Holographic Parallax Improves 3D Perceptual Realism'.

S1 SYSTEM

In this section, we describe the scheme of the holographic near-eye display system and the setup built for the experiment.

S1.1 Scheme

The holographic near-eye display is briefly explained in Figure S1(A). By using a spatial light modulator (SLM) with a pitch of p and illuminating it with a coherent source of light with a wavelength of λ, a wave field can be generated within the diffraction angle of $\theta_{diff} = 2\sin^{-1}(\lambda/2p)$. The SLM field then propagates and reconstructs a wave field at a certain distance, with a width similar to the SLM's width (W) and an angle within the diffraction angle. The aim is to reconstruct the intensity profile of $I(x, y, z)$, $z \in [z_{FCP}\, z_{NCP}]$, where x, y represents the horizontal, vertical position, respectively, and z denotes the axial distance from the SLM, within the axial distance between the far clipping plane (FCP, $z = z_{FCP}$) and the near clipping plane (NCP, $z = z_{NCP}$). Additionally, the wavefront recording plane (WRP, $z = z_{WRP}$), which is equivalent to the reference plane of orthographic light fields, is located at the middle of FCP ($z_{FCP} = z_{WRP} - z_o$) and NCP ($z_{NCP} = z_{WRP} + z_o$). We locate the FCP of the rendered volume at the focal length (f) of the eyepiece lens (EL). Then, the FCP will be virtually floated at the optical infinity and NCP will be located at the dioptric distance of $D_{NCP} = 1/(f - 2z_o) - 1/f$.

The beam with the limited diverging angle will form an eyebox, which is an exit pupil of the system. In Fig. S1(B), the relationship between the WRP domain and the pupil domain is demonstrated. The three beams that propagate in different directions with a small angular bandwidth in the WRP plane are remapped in the pupil domain. It shows the inversion of spatial and angular dimensions as the beams pass the lens. If the WRP domain is filled with the light field having the spatio-angular size of (W, θ_{diff}), the pupil domain will contain spatio-angular size of $(f\theta_{diff}, W/f)$. The spatial dimension corresponds to the eyebox of the near-eye display, and the field of view (FoV) is proportional to the angular dimension. Note that the product of eyebox and field of view is proportional to the display resolution and wavelength of the beam.

If the WRP is not placed at the focal length of EL, the projected light field is tilted resulting in the FoV difference depending on the pupil location inside the eyebox. However, placing the WRP plane in the middle is advantageous as the resolution degradation of LF-based hologram is proportional to the distance between the WRP and the depth of an object.

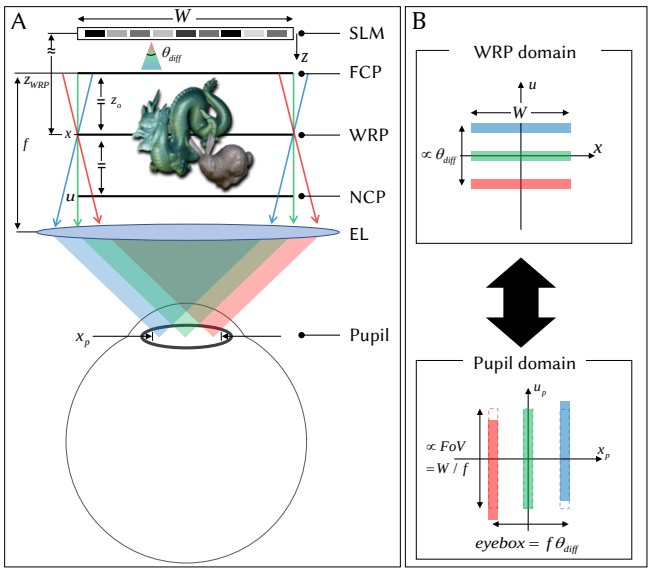

Fig. S1. Illustration that describes the (A) schematic of holographic near-eye display. (x, u) represents the spatial dimension of WRP domain and NCP, respectively. Note that u corresponds to the angular dimension of WRP. Likewise, (x_p, u_p) is the spatial, angular dimension of the pupil domain, respectively. (B) The light field of the WRP domain and pupil domain shows the relationship between the two domains. The colored light field corresponds to the beam shown in (A). The dashed line shows the relationship of two domains when the WRP is placed at the focal length of EL. We deduced the entire dimension to two for simplicity.

*Authors contributed equally to this research.

Authors' addresses: Dongyeon Kim; Seung-Woo Nam, Seoul National University, Republic of Korea; Suyeon Choi, Stanford University, USA; Jong-Mo Seo, Seoul National University, Republic of Korea; Gordon Wetzstein, Stanford University, USA; Yoonchan Jeong, Seoul National University, Republic of Korea.

0730-0301/2024/7-ART68
https://doi.org/10.1145/3658168

S1.2 Setup

Figure S2 demonstrates an overview of the holographic near-eye display prototype. A fiber-coupled laser diode of Wikioptics emanates a full-color beam with a central wavelength of 638 nm, 520 nm, and 450 nm. The beam is collimated with a lens (AC-508-200-A, Thorlabs) and the beam is linearly polarized with a series of linear polarizer (LPVISE200-A, Thorlabs) and an achromatic half-wave plate (AHWP10M-600, Thorlabs). We additionally placed a half wave plate to maintain the color balance as there are difference in polarization states by color. The beam is redirected with a 1-inch beam splitter and modulated with a reflective-type spatial light modulator.

We use a binary ferroelectric liquid crystal on silicon spatial light modulator (FLCoS SLM) to modulate the incident coherent beam. This SLM (QXGA-R10, a product of Forth Dimension Display) operates 1920 × 1200 pixels with a pitch of 8.2 μm at a speed of 3600 Hz to serve 24 full-color binary frames within 1/50 seconds. Placing an analyzer in the beam path allows the operation of SLM in amplitude mode. The field at the SLM plane is relayed with a 4-f system built with two identical camera lenses (AF Nikkor 50mm f/1.4D, Nikkon) facing opposite each other. A filter is placed in the Fourier domain to filter out the high-order signals arising from diffraction and the conjugate noise from complex representation with an amplitude SLM. The filter is fabricated with a rectangle aperture in an aspect ratio of 2:1 with a size determined by the SLM's diffraction angle in blue and the focal length of the 2-f lens. The relayed field is virtually floated by a 2-inch eyepiece lens (AL5040M, Thorlabs) having a focal length of 40 mm to guarantee a wide field of view. Thus, the eyebox size of the near-eye display is 2.2 mm × 1.1 mm.

We made an additional beam path by placing a beam splitter after the 4-f system to capture the experimental results and monitor the user experiment. For this arm, a lens (AC508-100-A-ML, Thorlabs) having a focal length of 100 mm is used as an eyepiece lens, and the scenes are captured with a c-mount lens with a 25 mm focal length and charge-coupled device (CCD) camera (BFS-U3-51S5C-C, FLIR) having a resolution of 2448×2048 and a pitch of 3.45μm. The CCD camera is placed on the two single-axis motorized stages (M-112.1DG1, a product of PI) to capture the image in distinct viewpoints with high accuracy. Note that the eyebox size of this arm is 2.5 times larger than the actual user experiment settings. Thus, we translated the CCD with the converted geometry. Additional spatial filters are placed at the relayed WRPs to eliminate the noise present in the peripheral region. Additional components shown in Fig. S2 but not described in this section will be explained in Section S4 with the description of user study implementation.

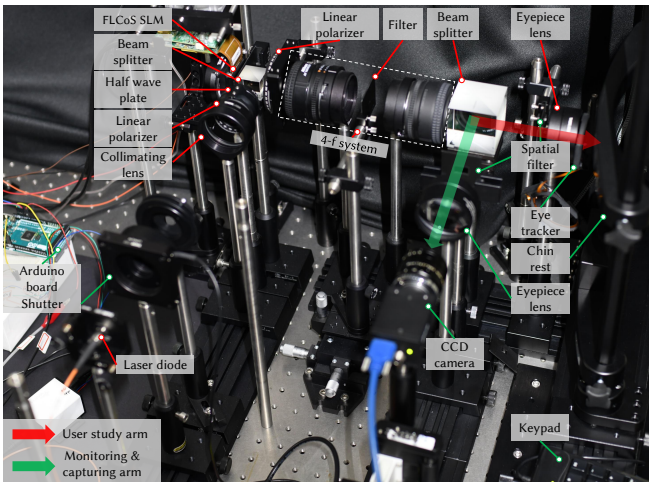

Fig. S2. The photograph depicts the testbed of a holographic near-eye display prototype used for user validation. The various components of the testbed are connected to a box. The components highlighted by red lines represent the essential equipment necessary for the holographic near-eye display. Conversely, the components connected by green lines are specifically implemented for the user experiment. The beam path is divided into two paths: the user study arm (indicated by the red arrow) and an additional arm for monitoring and image capturing (indicated by the green arrow).

S2 SOFTWARE IMPLEMENTATION

While Choi et al. [2022] previously demonstrated the effectiveness of this surrogate gradient method using the Gumbel-Softmax for phase SLMs, our work represents the first application of this technique to binary amplitude SLMs. Figure S3 demonstrates that this Gumbel-Softmax-based optimization outperforms the previous state-of-the-art binary CGH [Lee et al., 2022]. This approach offers a promising new direction for optimizing binary amplitude SLMs. For the content generation speed and quantitative comparison, please refer to Sec. S7.2

S3 LIGHT FIELD DATASET

We utilized a total of five different scenes in our paper, and these scenes were rendered using Unity. Fig. S4 presents the rendered light field maps and RGB-D images of each scene, along with the corresponding epipolar plane images (EPIs). For the light field maps, we rendered 25×25 orthographic views per color channel. However, in the figure, we have provided a subset of only 9×9 sampled views due to space limitations.

The EPIs provide insights into the angular distribution of the scenes. As depicted in the EPIs shown in Fig. S4, the individual slices exhibit distinct spatial information along the angular dimension. Notably, certain objects are only visible within specific angular ranges, while they disappear in other angular regions. This disparity in information across angles signifies the presence of "valid parallax," and it emphasizes that parallax containing meaningful information can be obtained when working with data formats that have four or more dimensions.

S4 USER STUDY IMPLEMENTATION

We implemented the overall setup for the user study. We additionally equipped an eye tracker, and a keypad for data acquisition. In addition, a chin-and-head rest, a shutter, an Arduino board, and an optical power meter are equipped to improve the study's accuracy and guarantee the subjects' eye safety.

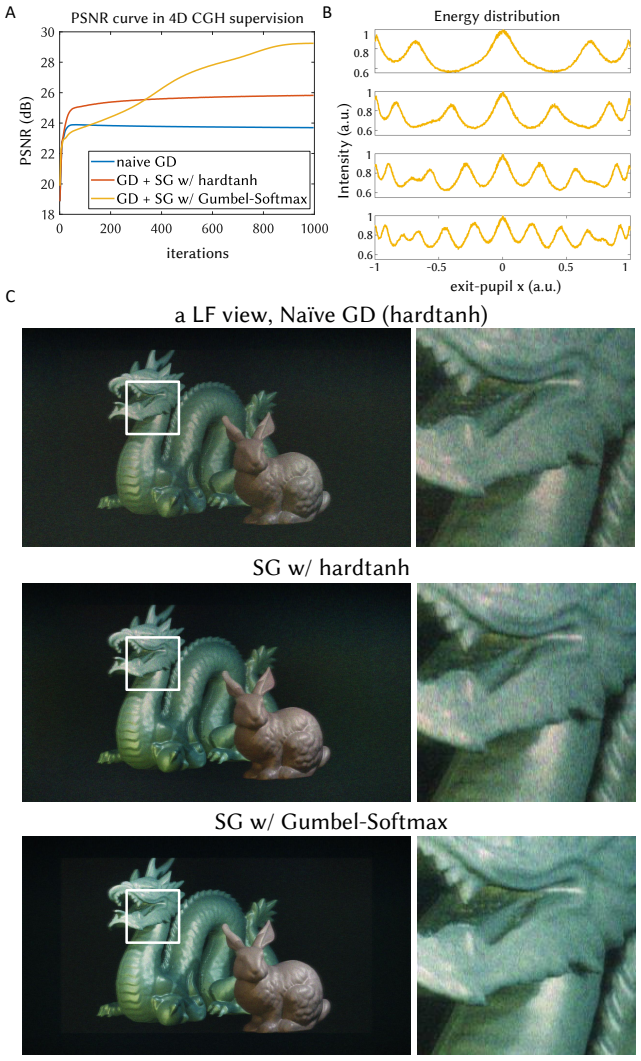

Fig. S3. A direct comparison of the surrogate gradient approaches for 4D supervision. We present (A) a convergence graph for binary amplitude SLMs using unit gradient and Gumbel-Softmax gradient methods. On the right, we present the (B) one-dimensional energy distributions across the exit pupil by combining and summing up the intensities at the Fourier plane, with a different number of light field views supervised ($3 \times 3, 5 \times 3, 7 \times 3,$ and 9×3, respectively). We also show (C) a sampled view from reconstructed light field for a qualitative comparison.

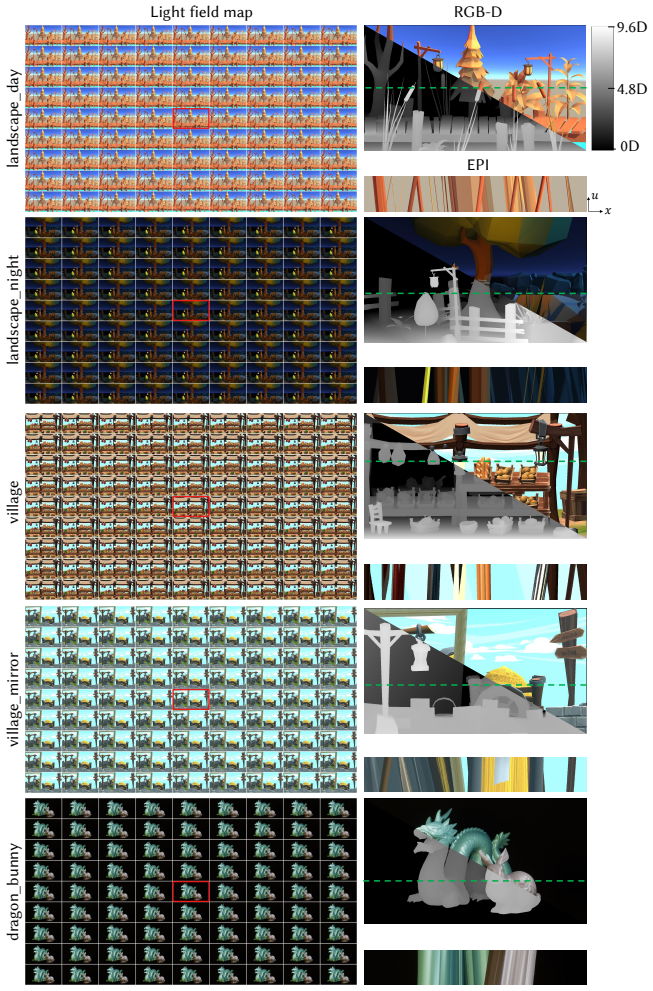

Fig. S4. Light field map, RGB-D image, and epipolar plane image (EPI) of the scenes (landscape_day, landscape_night, village, village_mirror, dragon_bunny) used in the paper are demonstrated. 9×9 orthographic images are provided as the light field map. The intensity and depth profile of the corresponding scenes' center view image (red box) is shown. The depth profiles of the orthographic scenes in a metric unit are converted to a unit of diopter considering the optical configuration of the near-eye display system. The EPI of the horizontal section depicted with a green dashed line is provided. The EPIs are drawn based on orthographic light fields and the upright slope implies the object is placed at the WRP. (Low Poly Series: Landscape, Fantastic-Village Pack: purchased unity asset, and Dragon, Bunny: credit to Stanford Computer Graphics Laboratory)

We utilized the Add-on eye tracker, developed by Pupil Labs, to measure the displacement of the subject's eye while they viewed the stimulus. Since we recorded the pupil displacement of a single eye, we couldn't utilize the built-in calibration functions designed for tracking both eyes simultaneously. Instead, we calibrated the measured data, which represented the center of the detected pupil, using a scale factor obtained through a pre-calibration procedure. This pre-calibration involved an eye figure with a black pupil that was moved laterally at the eye relief of the near-eye display. The collected data was obtained using the Pupil Labs Network API and saved in comma-separated value (csv) file format. Each trial involved a 2-second recording session, capturing the data at a speed of approximately 120 frames per second. We only included data points with a confidence value higher than 0.85 for further analysis. The response for each pair of options from the participants was received using a keypad and saved in csv file format, with values indicating the options that were compared.

To ensure the absolute position of the subject's head, a chin-and-head rest was employed to restrict head movement. The chin-and-head rest was positioned on a stage that allowed lateral movement. Subjects were instructed to adjust the initial position of their pupil by translating the stage. To address differences in intensity levels between holographic images created with various CGH supervision targets, as well as to ensure safety, we incorporated an Arduino board (Arduino Mega 2560) for two purposes. Firstly, it helped balance the intensity levels by adjusting the pulse width of the light source. The reconstructed images displayed different intensity levels due to variations in the scale factor for CGH supervision. By modulating the width of the rectangular pulse generated by the Arduino board, we could standardize the intensity levels across the images. Secondly, the Arduino board controlled a shutter placed in front of the light source, preventing uncontrolled light emission during the initialization process.

Luminance measurement. We ensured eye safety by measuring the luminance of the scenes. Directly measuring the luminance of each scene using the luminance meter proved challenging due to the small exit pupil of the holographic near-eye display system, which led to the underfilling of the entrance pupil of the measurement device. Instead, we opted to measure the power of the scene at the eyebox using a power meter (Newport, 818-SL/DB) and converted this data to luminance based on the geometry of the near-eye display system. The luminance (L_v) is measured in candela per meter square (nit) and can be calculated using the following equation:

$$L_v = \frac{683}{S \cdot \Omega} \int \Phi(\lambda) V(\lambda) d\lambda. \tag{S1}$$

Here, S represents the two-dimensional area where the image is displayed, Ω denotes the solid angle of the display source, $\Phi(\lambda)$ signifies the measured power at different wavelengths, and $V(\lambda)$ stands for the luminosity function of photopic vision.

The luminosity function varies depending on the light condition as cone cells are nonfunctional in low-light conditions [Wandell, 1995], but we used the photopic luminosity function as the standard of the scotopic (dark-adapted) vision level ranges below 0.001 nits. The holographic display utilizes a narrow-band source, thus wavelength-dependent power is provided with the power of color primaries. The measured power ranges in hundreds of picowatts, thus the average luminance is approximately converted to 2 nits. This value is significantly lower than the permissible level of laser exposure stated in the cited reference [on Non-Ionizing Radiation Protection et al., 1996].

S5 ADDITIONAL SIMULATION RESULTS

In this section, we provide the additional simulation results mostly consisting of the reconstructed results depending on the CGH supervision targets, and the number of views used in 4D CGH supervision with the analysis based on LF sampling theorem.

S5.1 Various 3D CGH supervision target formats

S5.1.1 Comparison by quality metric. We demonstrate the reconstructed holographic images with various 3D CGH supervision targets as Fig. S5. In Fig. S5(A), near-depth (8th out of 9) images of landscape_day scene acquired with different 3D CGH supervision assets are provided. In Fig. S5, the images are reconstructed with different pupil states and they are evaluated with three image metrics: Peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and FovVideoVDP quality metric. To calculate the image metrics, we utilize the amplitude of the reconstructed image and the corresponding ground truth images. The ground truth images are cropped to 80% of the entire FoV to eliminate the effects on the image boundaries from additional propagation. The FovVideoVDP metric is obtained under the identical conditions described in the paper, ensuring consistency in the evaluation process.

In reconstructed holographic images with a 2.5D supervision target, the artifacts get noticeable as the pupil displacement gets larger with defects in color. This is because out-of-focus regions are not penalized for 2.5D supervision. And this effect gets noticeable when a large depth difference between the object and its surroundings is present. Likewise, the reconstructed images with 3D w/ RGB-D target exhibit a similar problem without color artifacts. But, the overall contrast gets dimmer since the occlusion handling in the boundaries affects the contrast of the contents. In the reconstructed images of 3D w/ LF, the parallax is noticeable as the pupil gets decentered, and the overall image quality is far better in terms of metric. This is because the ground truth images are CGH supervision targets in this case. However, the FovVideoVDP metric deteriorates as the pupil size decreases and becomes decentered, deviating from the image formation model in CGH supervision. For cases supervised by 4D content, the image metric worsens as the reconstruction model is based on a plane-to-plane model, while the optimization model relies on a plane-to-perspective model, resulting in differences. Nevertheless, the FovVideoVDP metric remains consistent across various pupil states, and the discrepancy between the simulation results and the user experiment suggests the need for a perceptual quality metric for 3D content as a future research endeavor.

Fig. S5(C) provides the reconstructed images obtained using sampled pupil states corresponding to different focal depths. When the eyebox is fully sampled by a large pupil positioned at the center, the 3D w/ LF outperforms other 3D CGH supervision approaches, and this trend remains consistent regardless of the reconstructed depth. However, when the pupil becomes smaller and decentered, the cases supervised by 4D assets demonstrate better perceptual metrics.

S5.1.2 Content dependency. The FovVideoVDP metric extracts data from the image at the center depth across 15 distinct pupil states within the eyebox domain. Comparing the JOD value of 4D CGH supervision with 2.5D, 3D w/ RGB-D, and 3D w/ LF involves subtracting values, depicted in Fig. S6, alongside an illustration of the eyebox domain. These plots showcase three scenes (landscape_day in red, landscape_night in orange, and village in blue).

While some variations occur based on the content, the comparison consistently highlights the superior performance of 4D supervision over other approaches in overfilled pupil conditions. However, the increment diminishes as the pupil becomes underfilled, even showing JOD differences below zero in the case of overfilled pupils when comparing 4D vs. 3D w/ LF.

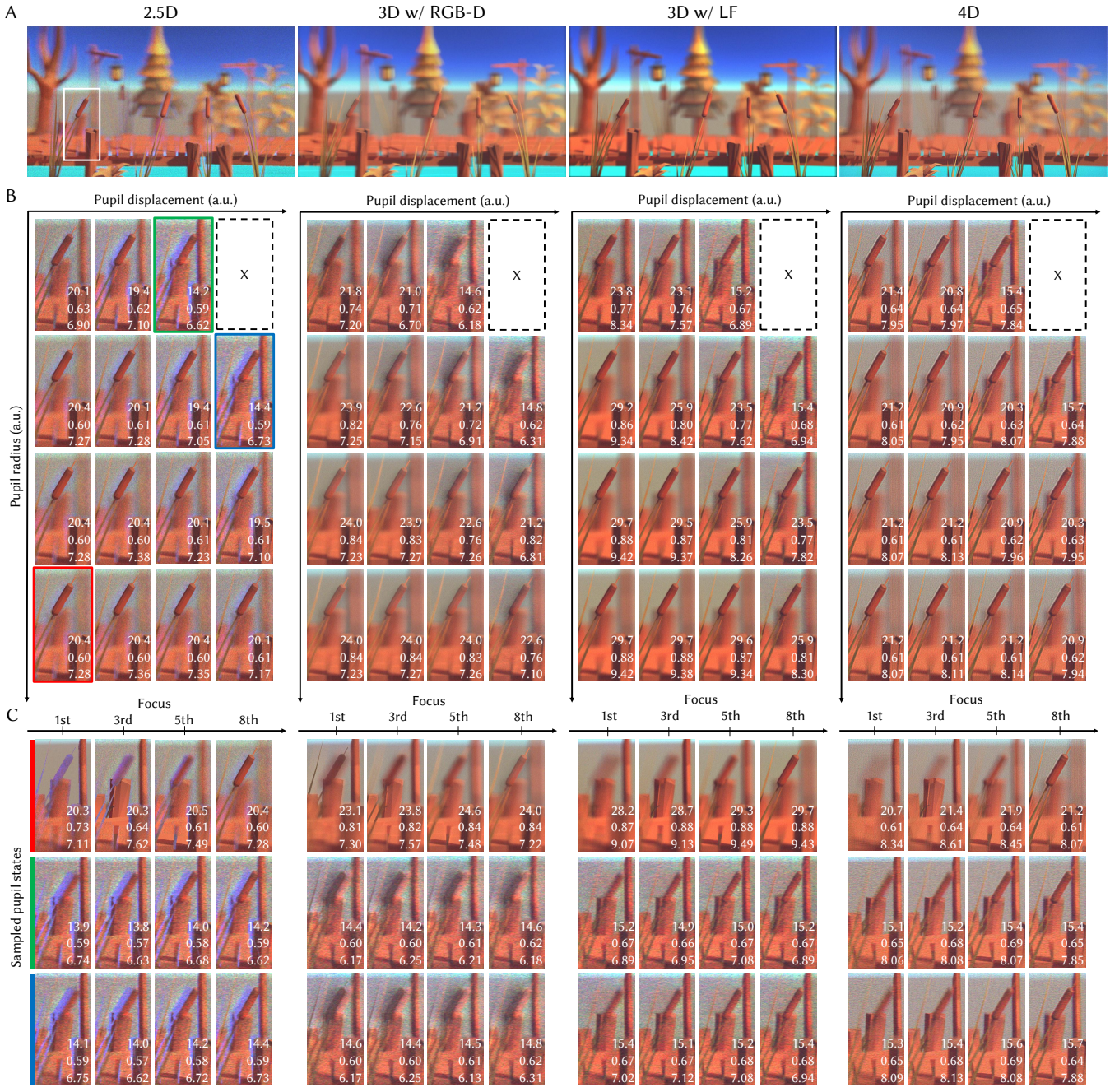

Fig. S5. Reconstructed holographic images with various 3D CGH supervision targets: (A) Near-depth reconstructed image of landscape_day (2.5D (1st column), 3D w/ RGB-D (2nd column), 3D w/ LF (3rd column), and 4D (4th column) supervision targets) scene. (B) The white box sections of the images reconstructed with various pupil states (pupil displacement, and pupil radius) are shown. The blank section with a dashed boundary in the figure indicates the that visualization is not capable as the state is fully vignetted. Peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and FovVideoVDP quality metric in a unit of JOD are consecutively provided on the bottom of every inset. For FovVideoVDP metric, the ground truth image is referenced and demonstrates 10 JOD and the difference of 1 JOD corresponds to a 50 percent preference over the ground truth image. (C) The identical parts of images reconstructed with three different pupil states are provided depending on the focal depths. Four out of the nine depths equally sampled in the diopter are provided for simplicity. The images reconstructed with various pupil shifts, pupil sizes, and focal depths can help understand the impact of those aspects in the realized holographic scenes.

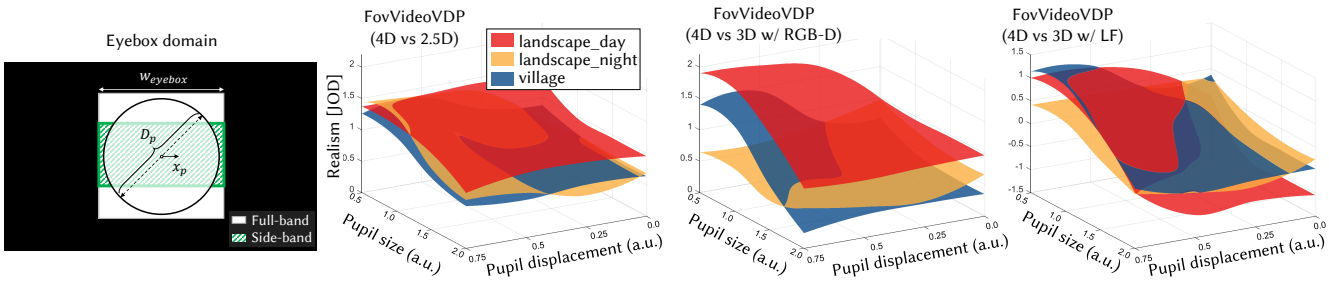

Fig. S6. Illustration of the eyebox domain (1st col) of holographic near-eye display with the width of w_{eyebox} and the circular pupil with a diameter of D_p and displacement of $(x_p, 0)$. The side-band eyebox (green) is vertically halved in size relative to the full-band eyebox (white) for complex modulation with a single amplitude SLM. Comparison of 4D CGH supervision with 2.5D (2nd col), 3D w/ RGB-D (3rd col), 3D w/ LF (4th col) CGH supervision is conducted with the image reconstructed with various normalized pupil displacement ($x_{p,norm} = x_p / w_{eyebox}$) and normalized pupil size ($D_{p,norm} = D_p / w_{eyebox}$). Note that the scale of the grid differs for the last figure.

S5.1.3 Eyebox evaluation. The accurate provision of a 4D light field across the eyebox requires attention to two critical aspects. First, it is necessary to ensure that the entire energy is distributed across the eyebox. Unlike other types of near-eye displays, holographic near-eye displays can adjust the effective size of the eyebox by controlling the phase randomness of the reconstructed field. It is important to note that manipulating the randomness of the phase profile of the reconstructed field impacts the effective size of the eyebox, the depth of field of the display, and the gain of dynamic accommodation response [Kim et al., 2022]. Therefore, the size of the eyebox must be maximized, as determined by the product of the maximum angle and the focal length of the eyepiece lens. The eyeboxes created by the different CGH supervision approaches are measured and the captured results are provided in Fig. S18. Additionally, the perspective images with designated carrier frequency should provide scenes with desired energy distribution.

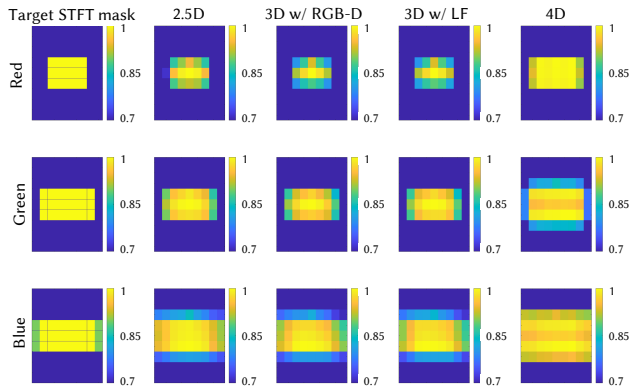

Fig. S7. Energy distribution of the 9×9-STFT-reconstructed images of village scene. The 9×9 masks are tiled based on the index of the carrier frequency and provided depending on the various 3D CGH supervision approaches (2nd col: 2.5D, 3rd col: 3D w/ RGB-D, 4th col: 3D w/ LF, 5th col: 4D) with the target STFT weight mask (1st col). These energy distributions are provided based on each color channel (1st row: red, 2nd row: green, 3rd row: blue). The overall intensity is normalized and the scale bar is clipped with a minimum value of 0.7.

For an accurate analysis of the energy carried by each individual localized beam, each of which carries the signal of the discrete light field, it is essential to examine the energy of each orthogonal view. The energy distribution of the reconstructed view images is illustrated in Fig. S7. This figure demonstrates the energy distribution of 9×9 STFT-reconstructed images of the village scene using different CGH supervision approaches. Each 9×9 tile represents the averaged intensity of the STFT-reconstructed image with a specific direction.

From the figure, it becomes evident that both 2.5D supervision and 3D supervision struggle to accurately reconstruct the light field with a discrete carrier frequency, as the energy of the views near the boundary decreases. In contrast, 4D supervision uniformly generates the light field with a discrete carrier frequency. The target STFT mask varies in color to match the physical eyebox of the system. It's worth noting that the weight of the target STFT mask is adjusted at the boundary to mitigate the impacts of undesirable diffraction at the edge.

S5.1.4 Eye movement. To simulate how scenes are perceived as the eye moves within the eyebox, we present simulated results using different 3D CGH supervision approaches based on various eye positions and diameters, as depicted in Fig. S8. When the eye's pupil is sufficiently small, it samples only a portion of the eyebox signal, making 4D supervision highly tolerant, regardless of the eye's position. In detail, the object near to the eye is placed relatively leftward compared to the object placed far when the pupil is placed in the left area of the eyebox. Furthermore, parallax information is well-preserved in cases of 4D supervision, as discussed in Fig. S5.

If the eye's pupil enlarges due to practical conditions such as changes in light conditions, the view-dependent parallax effect naturally diminishes. Moreover, when the eye's pupil becomes large enough to sample the entire eyebox, no differences are observed as the eye moves. This is because we assume that the eye pupil is a diffraction-limited pupil without any fluctuation in transmittance, in contrast to real-world conditions. This assumption downsizes the effect of displacement-dependent parallax, which has been discussed in the main paper.

S5.1.5 Ocular parallax detection analysis. Figure S9 provides the amount of the ocular parallax realized by two objects placed at

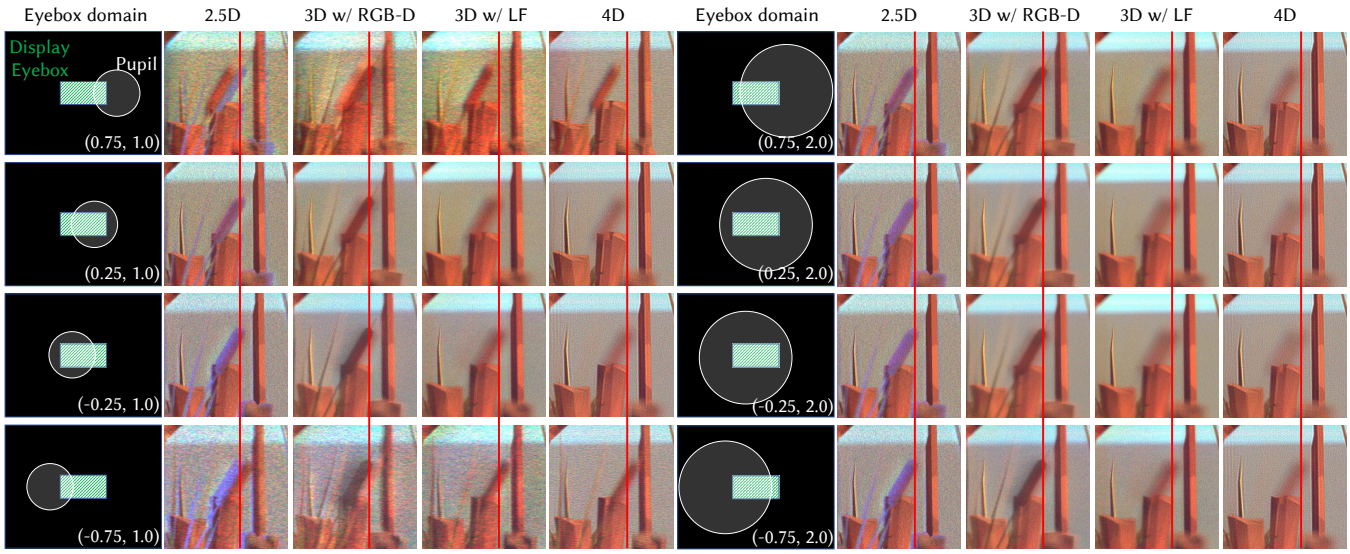

Fig. S8. Reconstructed scenes depending on the state of the pupil inside the eyebox domain when the human eye pupil is simulated as diffraction-limited. The green section represents the eyebox of the near-eye display system and the pupil is represented with a white circle. For each state of the eye position and diameter $(x_{p,norm}, D_{p,norm})$, parts of landscape_day scene with focal depth of 2nd depth are provided. The scenes assume the pupil functions as a diffraction-limited aperture. Red lines are placed on purpose to emphasize the disparity of the objects.

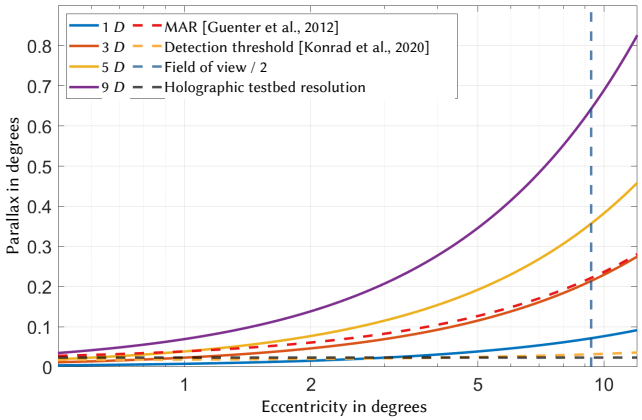

Fig. S9. The amount of ocular parallax, estimated in a unit of degrees of visual angles, depending on the retinal eccentricity and the dioptric disparity of two objects placed at different depths (solid lines). Minimum angle of resolution (MAR) in the work of Guenter et al. [2012] (red, dashed line) and the ocular parallax detection threshold investigated in the work of Konrad et al. [2020] (orange, dashed line) are present along with the half of the field of view (blue, dashed line), and minimum resolution supported by holographic near-eye display testbed (black, dashed line). Note that the ocular parallax detection threshold is far smaller than the MAR.

different depths. The depth of the reference object is assumed as 0.5 D for simplicity. We refer to the work of Konrad et al. [2020] for the details on the ocular parallax simulation.

Along with the parallax of two objects, we demonstrate the minimum angle of resolution (MAR, ω) [Guenter et al., 2012] depending on the eccentricity (e) as $\omega = 0.022e + \omega_0$, where, ω_0 denotes the minimum angular resolution as 1/60 (20/20 vision) in a unit of degrees. In addition, the detection threshold of ocular parallax was measured as 0.36 D in the eccentricity of 15 degrees in the user evaluation of the work [Konrad et al., 2020]. Here we assume the threshold also follows the linear model of the eccentricity-dependent resolution falloff and the slope is fitted as 0.0016. Other than these thresholds, we additionally provide the halved horizontal FoV and the minimum angular resolution supported by the testbed.

As observed in Fig. S9, the detection threshold is far below the MAR measured with the static stimuli as the detection threshold involves motion-based perception, which may result in detection in the periphery. Although the depth range of the stimuli used for the user study spans 9.6 D, the ocular parallax can still be detected even with the objects having a narrower depth range. This ocular parallax estimation leaves a question; Does the images perceived with the eye movement present parallax larger than the detection threshold?

To answer the raised concern, we roughly quantify ocular parallax detection using two frames processed with specified pupil displacement and pupil diameter. We employ a technique commonly used in finding 3D stereo pairs - feature point matching [Liu et al., 2010] to calculate disparities in these frames.

In detail, if a set of feature points ($\mathbf{X}$) is extracted from a single image (I), it can be formulated as $\mathbf{X}(I) = \{\mathbf{x} | \mathbf{x} \in F(I)\}$, where, $\mathbf{x} \in \mathbb{R}^{2\times 1}$ is a vector presented the two-dimensional angular displacement from the center axis and $F(\cdot)$ is the feature point extraction operator. Then, a set ($\mathbf{P}$) of pairs $(\mathbf{X}_1, \mathbf{X}_2)$ extracted from two different frames (I_1, I_2) can be presented depending on the threshold

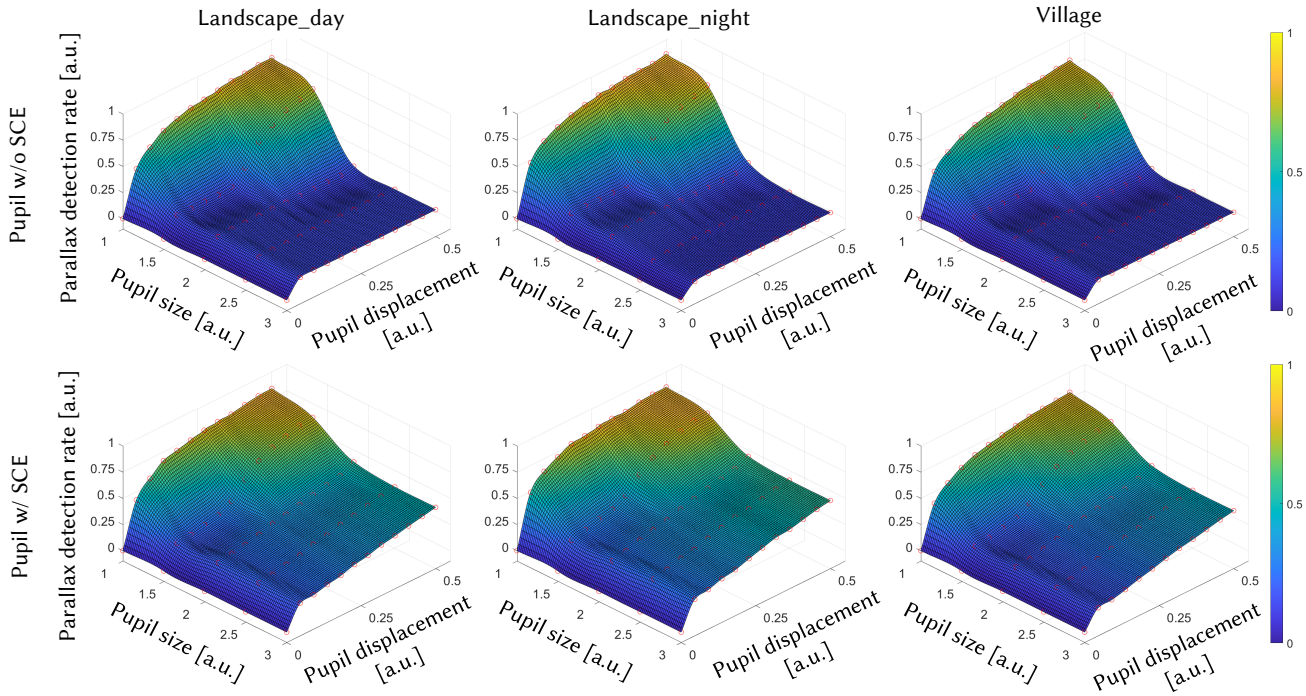

Fig. S10. The ocular parallax detection rate is estimated using three different scenes (landscape_day, landscape_night, and village) processed with 60 pupil states (depicted by red circles) and interpolated across the pupil state domain, considering pupil size and pupil displacement. The reconstructions are performed under two human eye pupil apodization profiles: (*top*) pupil w/o Stiles-Crawford effect (SCE) and (*bottom*) pupil w/ SCE.

value (θ_{th}) as follows:

$$\mathbf{P}_\theta(\mathbf{X}_1, \mathbf{X}_2) = \{\mathbf{x} | \mathbf{x}_1 \in \mathbf{X}_1, \mathbf{x}_2 \in \mathbf{X}_2, \|\mathbf{x}_1 - \mathbf{x}_2\| \geq \theta_{th}\}. \tag{S2}$$

Here, the subscript in the single-frame image denotes the pupil state (p) comprised of pupil displacement and diameter and the focal state (j) of the eye. Here, each pair of feature points is evaluated whether the l-2 norm ($\|\cdot\|$) of the difference in the angle space exceeds the detection threshold.

We normalized the count of feature point pairs meeting the specified condition by dividing it by the total number of extracted pairs. This normalization process entailed traversing through the sampled focal states and dividing by the total extracted feature pairs. The resulting value represents the parallax detection rate and can be presented as

$$R_\theta(p_1, p_2) = \frac{\bigcup_{j=1}^{J} \mathbf{P}_\theta\left(\mathbf{X}_{p_1,j}, \mathbf{X}_{p_2,j}\right)}{\bigcup_{j=1}^{J} \mathbf{P}\left(\mathbf{X}_{p_1,j}, \mathbf{X}_{p_2,j}\right)}. \tag{S3}$$

We simulate the ocular parallax detection rate under the conditions emulating an ideal 3D display capable of supporting 25×25 perspective images, matching the viewing conditions of our experimental setup. This step helps eliminate potential errors in the feature point extraction process, particularly in holographic images containing speckle noise. Furthermore, we extend this evaluation to encompass an ideal 3D display scenario that does not experience angular resolution degradation, achieved through a dense distribution of view images. This quantitative analysis broadens the scope of validity of our findings.

In our analysis, we utilized the SIFT flow [Liu et al., 2010] method for feature point extraction. We gathered paired images representing nine distinct focal states for our investigation. To simplify our computations, we reduced the dimensions of the parallax detection rate function described in Eq. S3 from six to two. We assumed that both pupils were horizontally aligned. Furthermore, we maintained consistent pupil diameters in both states ($p_1 = (x_{p,norm}, D_{p,norm}), p_2 = (0, D_{p,norm})$), given that the pupil size is primarily influenced by the scene's luminance.

The parallax detection rate is calculated with 5 (pupil diameter) × 12 (pupil displacement) discrete pupil states and interpolated across the pupil state domain as provided in Fig. S10. In the figure, the pupil diameter and pupil displacement are normalized with the width of the eyebox, respectively.

The parallax detection rate is simulated based on two different pupil apodization modes (pupil w/o Stiles-Crawford effect (diffraction-limited pupil) and pupil w/ Stiles-Crawford effect). The figures demonstrate how the parallax detection rate would be affected by the human eye apodization profile. For the detection threshold value, we chose the eccentricity-dependent model, which is approximated as a linear function of eccentricity with the given parameter of the work of Konrad et al. [2020].

In Fig. S10, it is clear that the parallax is hardly detected when the pupil displacement is minimal regardless of the pupil diameter. However, there are large differences when the eye is in motion. Especially, if the pupil is assumed to a diffraction-limited, the parallax detection rate does not vary depending on the pupil displacement when the normalized pupil size exceeds 2.0. On the other hand, the parallax detection rate begins to saturate at a certain value when the pupil is apodized. It can be observed in the bottom figures of Fig. S10 plotting the projected line showing the relationship between pupil size and parallax detection rate. Note that the average of the measured subjects' pupil diameters exceeds 4.4 mm ($D_{p,norm} = 2.0$) in the actual experiment.

Here, we do not claim that the given ocular parallax detection rate is built upon an accurate model with measurements nor it represents the absolute detection probability, but we evaluate with the given simulation to show the difference in terms of ocular parallax. Interestingly, there is limited exploration into the detection and discrimination threshold of ocular parallax. Accurate modeling of the ocular parallax and quantifying the impact on 3D realism presents an intriguing avenue for future research.

S5.1.6 Stiles-Crawford effect depending on CGH supervisions. We present the reconstructed results of different CGH algorithms assuming the pupil with Stiles-Crawford Effect in Fig. S11. They are provided with different pupil states, where two of the pupil states (($x_{p,norm}$,$D_{p,norm}$) = (0.75, 3.0), (-0.75, 3.0)) show fully underfilled pupil cases. The disparity of the objects that corresponds to the depth difference can be seen in the results, while LF-based supervision targets (3D w/ LF and 4D) presented images with robust defocus blur. In addition, it can be observed that the additional adoption of SC pupil reconstructs different perspectives even in the case when the pupil is completely underfilled. The 3D w/ LF and 4D cases show similar results while the parallax is slightly more noticeable in the results of 4D case. Note that the difference among the algorithms can be found in Fig. S8.

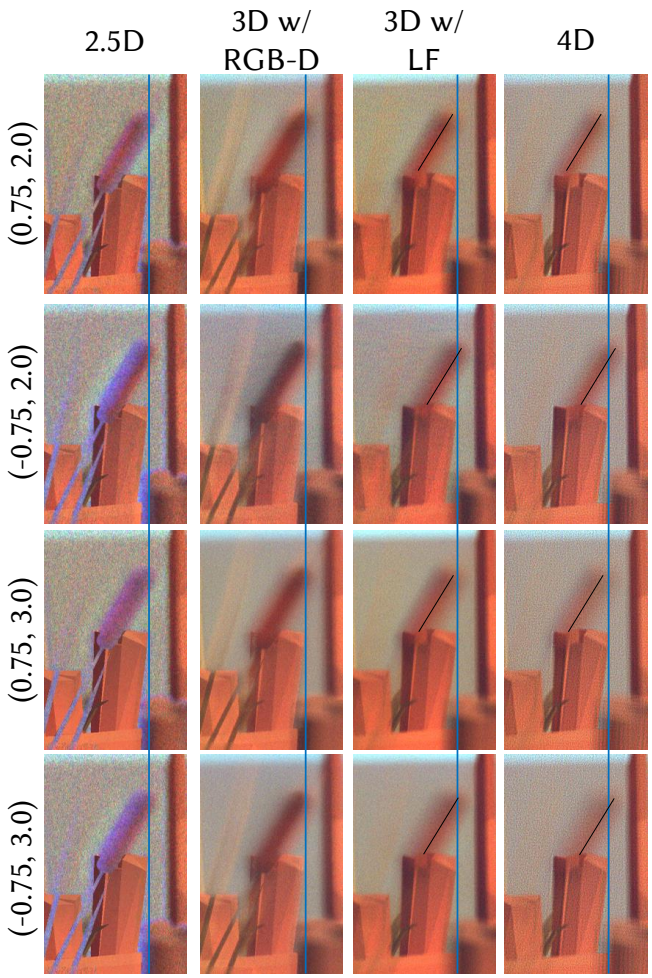

Fig. S11. Enlargements of reconstructed results with apodized pupil having Stiles-Crawford effect depending on CGH algorithms (1st col: 2.5D, 2nd: 3D w/ RGB-D, 3rd: 3D w/ LF, 4th: 4D). They are reconstructed in different pupil states ($x_{p,norm}$,$D_{p,norm}$). For visibility, the blue, and black lines are additionally drawn to emphasize the disparity of the objects.

S5.2 Number of views for 4D CGH supervision

In section 5, we investigate the number of views required for 4D CGH supervision through both camera-incorporated experiments and a user study. Each condition is easily understood with the simulated images of various scenes provided in Fig. S12-S13.

By examining these images, it becomes apparent that the overall quality of the reconstructed 3D scene improves as more views are incorporated into the 4D CGH supervision. However, the objects placed near the WRP are reconstructed with a high resolution (see the enlarged images of tree, baguette in Fig. S12, and handle of basket, head of dragon in Fig. S13) and do not demonstrate the noticeable difference when the scenes are rendered with denser views. On the other hand, the image quality suffers for the objects lying at planes that deviated from the WRP (see the enlarged images of street light, fruit in Fig. S12 and wagon reflected by a mirror, head of bunny in Fig. S13) especially when the views are sparsely sampled. Therefore, the reconstructed results visually demonstrate that the placement of objects relative to the WRP and the number of views directly affect the resolution of the reconstructed scene.

The resolution of CGHs exhibits a depth-dependent characteristic due to the reconstruction of individual view images in 4D-supervised CGHs. However, when CGHs are generated using a plane-to-plane model, the highest resolution remains consistent regardless of changes in depth. Surprisingly, the results obtained from the user study present contrasting findings, even though the experiments are conducted with scenes featuring a large distribution of depth. These outcomes indirectly suggest that the assessment of 3D perceptual realism cannot be solely determined by the resolution of objects located at varying distances from the user.

S5.3 Number of views in LF sampling analysis

We analyze the number of views required for 4D CGH supervision based on light field sampling theorem [Park, 2017, Zhang and Levoy, 2009]. In the theorem, the depth range covered by the light field

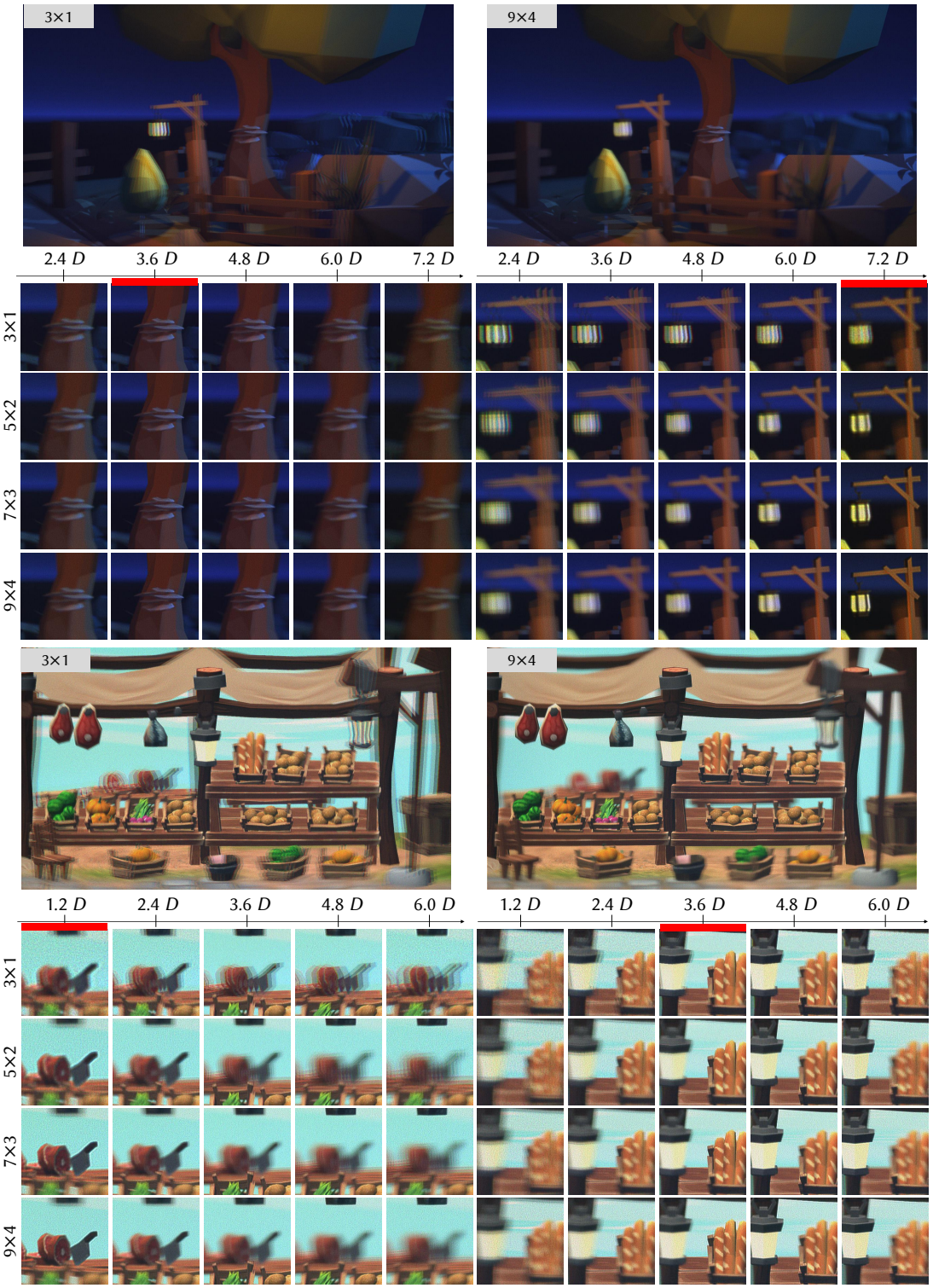

Fig. S12. Reconstructed holographic images of landscape_night (*top*) and village (*bottom*) scene supervised with sparse light field (*left*) and dense light field (*left*). Two different sections of the holographic images supervised with different view numbers (3×1, 5×2, 7×3, 9×4) are provided with five different focal states. The red bar placed at the top of the column indicates that the enlarged object is best focused at the depth.

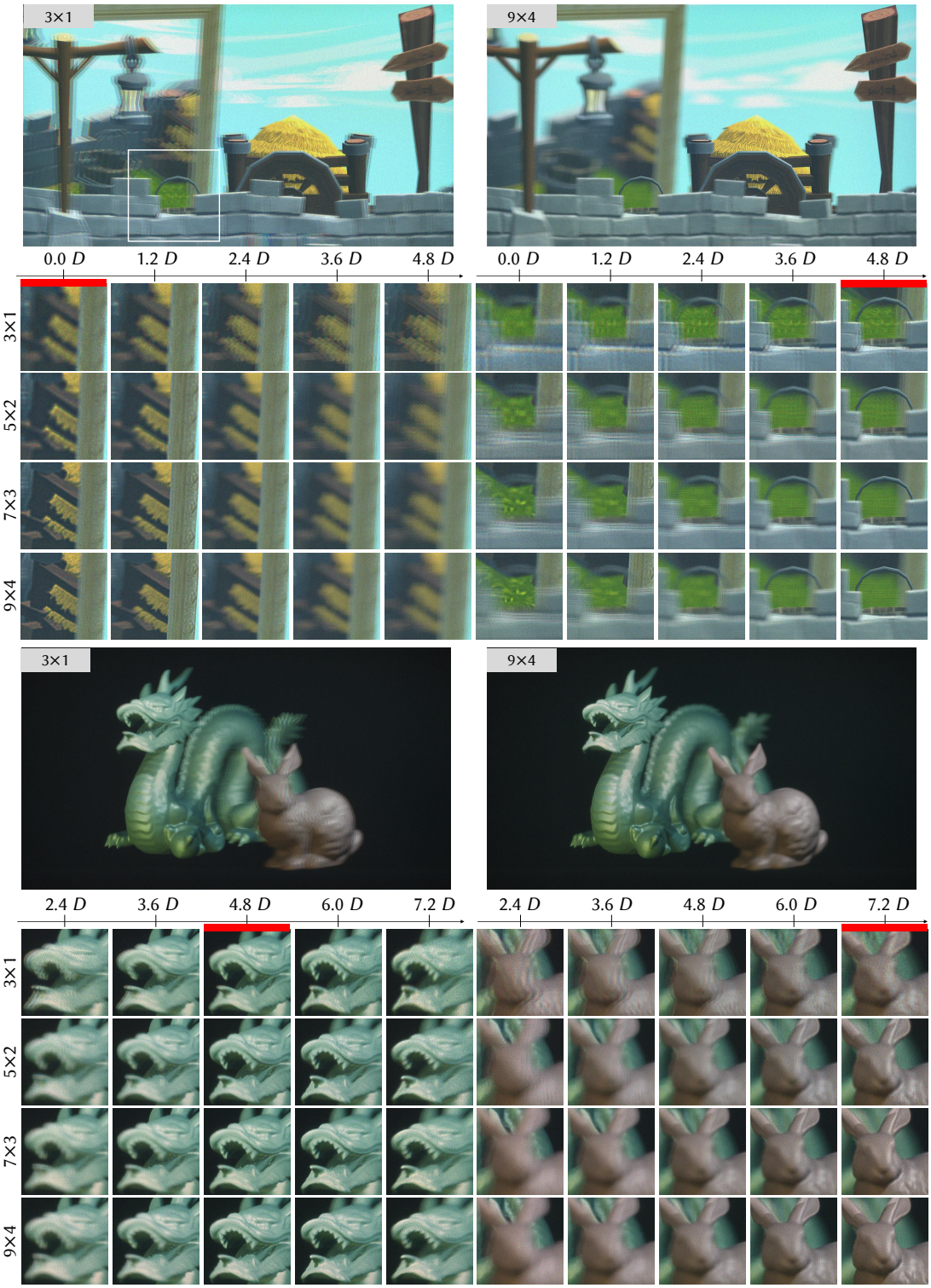

Fig. S13. Reconstructed holographic images of village_mirror (*top*) and dragon_bunny (*bottom*) scene supervised with sparse light field (*left*) and dense light field (*left*). Two different sections of the 4D-supervised holographic images with different numbers of views (3×1, 5×2, 7×3, 9×4) are provided with five different focal states. The red bar placed at the top of the column indicates that the enlarged object is best focused at the depth.

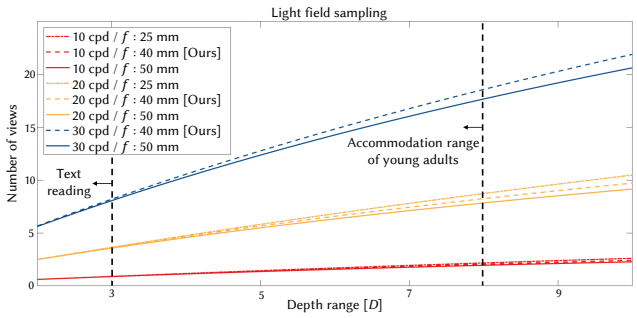

Fig. S14. Required number of views in horizontal for 4D CGH supervision based on light field sampling theorem. It differs by the spatial bandwidth of the scene, and the depth range of the 3D scene. The graph is plotted with three different spatial bandwidths (red: 10 cpd, orange: 20 cpd, blue: 30 cpd) and the optical configurations having eyepiece lenses with different focal lengths (dotdash line: f : 25 mm, dashed line: f : 40 mm, and solid line: f : 50 mm). The wavelength is assumed as 532 nm for the simulation. The simulation is conducted with an SLM having a pixel pitch of 8.2 μm and a horizontal resolution of 1920, which is identical to the experimental setup. We additionally placed black dashed lines in 3 D and 8 D, each of which indicates the dioptric range of text reading and accommodation range [Duane, 1912], respectively. Note that the case of 30 cpd / f : 50 mm is not plotted as the cut-off frequency of the condition is below 30 cpd.

is proportional to the angular resolution. We deviate the analysis with the optical configuration of near-eye displays. Let's assume the situation when the WRP is placed at a certain distance, and the FCP is located at the focal length of the eyepiece lens as Fig. S1. The metric distance between FCP and NCP is $2z_o$ which makes the depth coverage from 0 D to D_{max} and the relationship between the two variables is as follows:

$$z_o = \frac{f}{2} - \frac{1}{2\left(D_{\max} + \frac{1}{f}\right)} = \frac{f^2 D_{max}}{2\left(f D_{\max} + 1\right)}. \tag{S4}$$

Here, the angular resolution (N_u) of the light field required to reconstruct the image with spatial bandwidth of (B_x) in the distance of z_o can be obtained as:

$$N_u \geq \lambda z_o B_x^2 \tag{S5}$$

Integration of the equations (Eq. S4 and Eq. S5) allows us to calculate the maximum depth range supported by the given angular resolution of the light field and the spatial bandwidth of the scene. If the spatial bandwidth is bounded to a certain range, the low-pass filtered spatial bandwidth ($B_{x,v}$) can be simply acquired with the ratio of spatial frequency (v) and cut-off frequency (v_{cutoff}) as $B_{x,v} = B_x \frac{v}{v_{cutoff}}$. The cut-off spatial frequency can be acquired with the optical configuration of the near-eye display.

Therefore, the graph labeled as Fig. S14 illustrates the number of horizontal perspectives required for 4D-supervised CGH based on LF sampling analysis, depending on the depth range. We sampled three different spatial bandwidth regions, and the results are depicted using three different eyepiece lenses. We sampled three different spatial bandwidth regions, and the results are drawn with three different eyepiece lenses. Although the eye relief of the near-eye display, which we assume that the focal length of the eyepiece lens is identical to the eye relief, is down to the conventional range, the estimated value for the required number of views is similar to the parameter obtained with our system's configuration.

Furthermore, if the entire depth range is reduced to 3 D, which corresponds to the depth at which we often position books for reading, the required number of horizontal views is approximately 8, supporting a resolution of 30 cpd. However, if the 3D content aims to cover the full depth range of 8 D with high resolution, more than 15 views are necessary. It is important to note that this analysis does not consider other potential factors such as diffraction from the eye pupil or aberration in an individual's eye, which can affect the point spread functions and ultimately impact the results.

S5.4 VDP simulation with matched display model

The comparisons of VDP in Fig. 3, Fig. S4-S5 are performed based on the luminance and contrast conditions of conventional VR headsets, while the experimental conditions differ from the simulated conditions. First, the experiment was conducted under low-luminance lighting conditions due to safety concerns regarding eye safety. Additionally, the contrast may be lower for holographic displays due to imperfect black level expression. These mismatches in the display model between the simulation and experimental conditions may result in different outcomes.

Fig. S15 demonstrates the simulated FovVideoVDP results of various CGH algorithms conducted under three different conditions: the conventional VR condition, with a luminance of 100 cd/m^2 and a contrast level of 1000:1; the luminance-lowered condition to match the peak luminance with the experiment, corresponding to 2 cd/m^2 and a contrast level of 1000:1; and the contrast-lowered condition, with a luminance level of 100 cd/m^2 and a contrast level of 20:1. These are simulated with seven different pupil displacements and the results are plotted based on pupil sizes ($D_{p,norm}$ = 1.0, 2.0, 3.0).

It can be observed that as the luminance level decreases, the absolute values of JOD tend to increase. This is due to suppressed contrast sensitivity at low luminance levels, which allows the simulation to neglect grainy noise that is prevalent in holographic images, especially when the contrast level is lower. However, the overall trend does not change, nor does it reverse. This implies that the minute mismatch of luminance or contrast specifications is not the major reason for the superiority of the 4D case over the others.

S6 ADDITIONAL EXPERIMENTAL RESULTS

In this section, we additionally demonstrate the experimental results that are not provided in the paper.

S6.1 Additional captured results

Figure S16 additionally presents the images captured while changing the camera positions. For the village_mirror scene, the depth difference of the objects reflected by the mirror is observed when comparing RGB-D-based approaches and LF-based approaches. This also presents the limitation of the RGB-D-based presentation of volumetric scene.

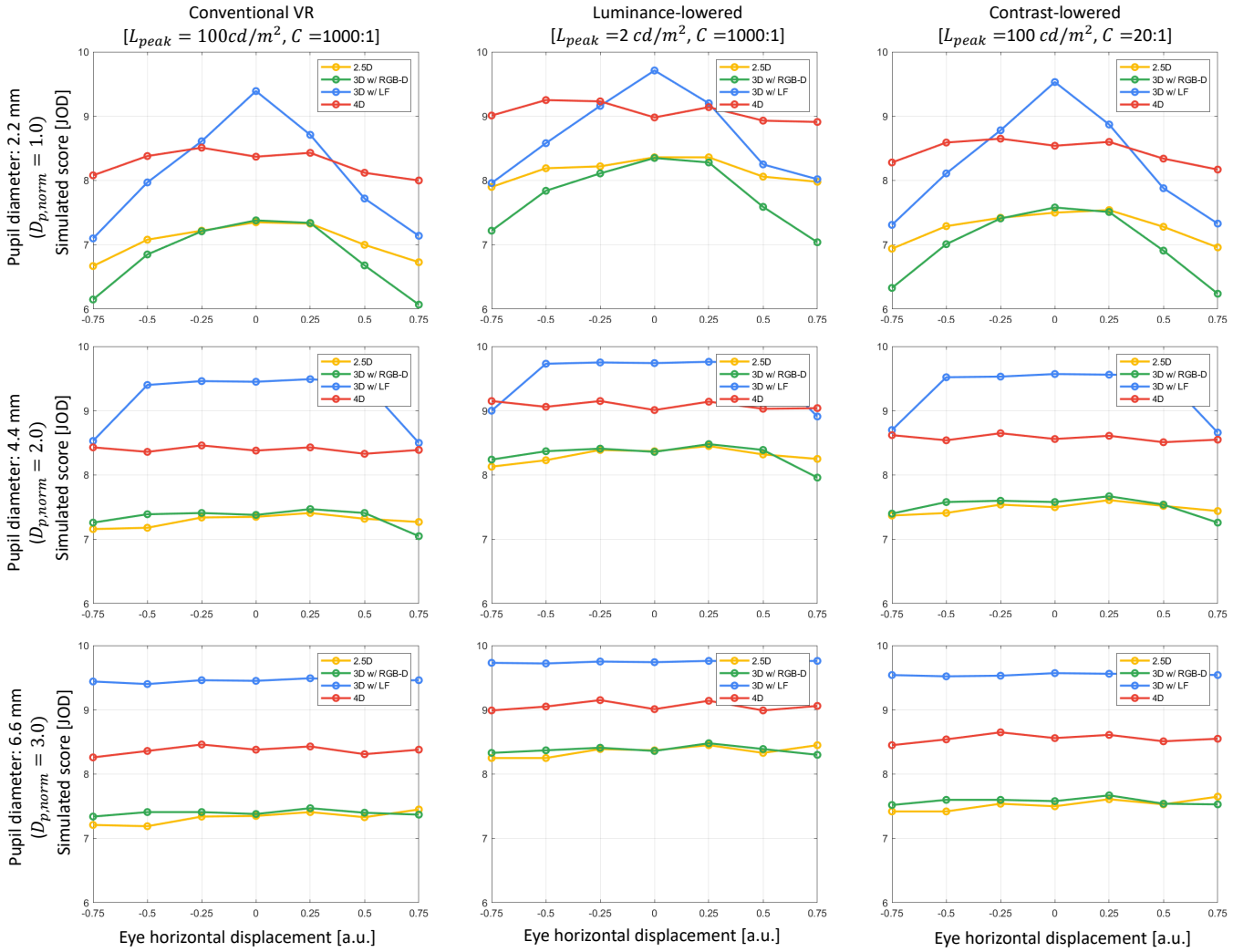

Fig. S15. FovVideoVDP simulation of different CGH algorithms (yellow: 2.5D, green: 3D w/ RGB-D, blue: 3D w/ LF, and red: 4D) under different luminance (L_{peak}) and contrast ($C = L_{peak} : L_{min}$) conditions. The first column denotes the conventional VR model having luminance of 100 cd/m^2 and contrast level of 1000. The second column corresponds to the low luminance condition matched with the experimental condition. The third column demonstrates the condition when the contrast level is worse. They are demonstrated with three different pupil diameter conditions. The values are extracted with the reconstructed image of fifth focal stack of the landscape_day scene assuming diffraction-limited pupil condition. The range of -0.5 to 0.5 in the eye horizontal displacement corresponds to the width of the eyebox.

S6.2 Captured results with pupil movement contour

We present the captured frames of two different pupil states of the camera at the eyebox domain as shown in Fig. S17. In detail, the camera is placed at the rightmost ($\theta_k = 96°$) and leftmost ($\theta_k = 264°$). For better visualization, we provide enlarged images of various 3D CGH supervision approaches. Refer to Fig. 2 in the paper, 2.5D and 3D have occlusion boundary problems, and 3D w/ LF demonstrates images with averaged intensity across the view images, which inherently suppresses the visualization of view-dependent effects.

S6.3 Captured eyebox

We additionally provide the captured eyebox of the holographic near-eye display system with various CGH supervision targets as Fig. S18. At the same time, the ruler with the millimeter scale is placed at the eyebox plane to roughly measure the physical size of the eyebox.

Some previous works on holographic displays showed some results limiting the eyebox size while improving the 2D quality of holographic scene at the sampled depth. However, we employed binary SLM that results in complex-valued field with random phase

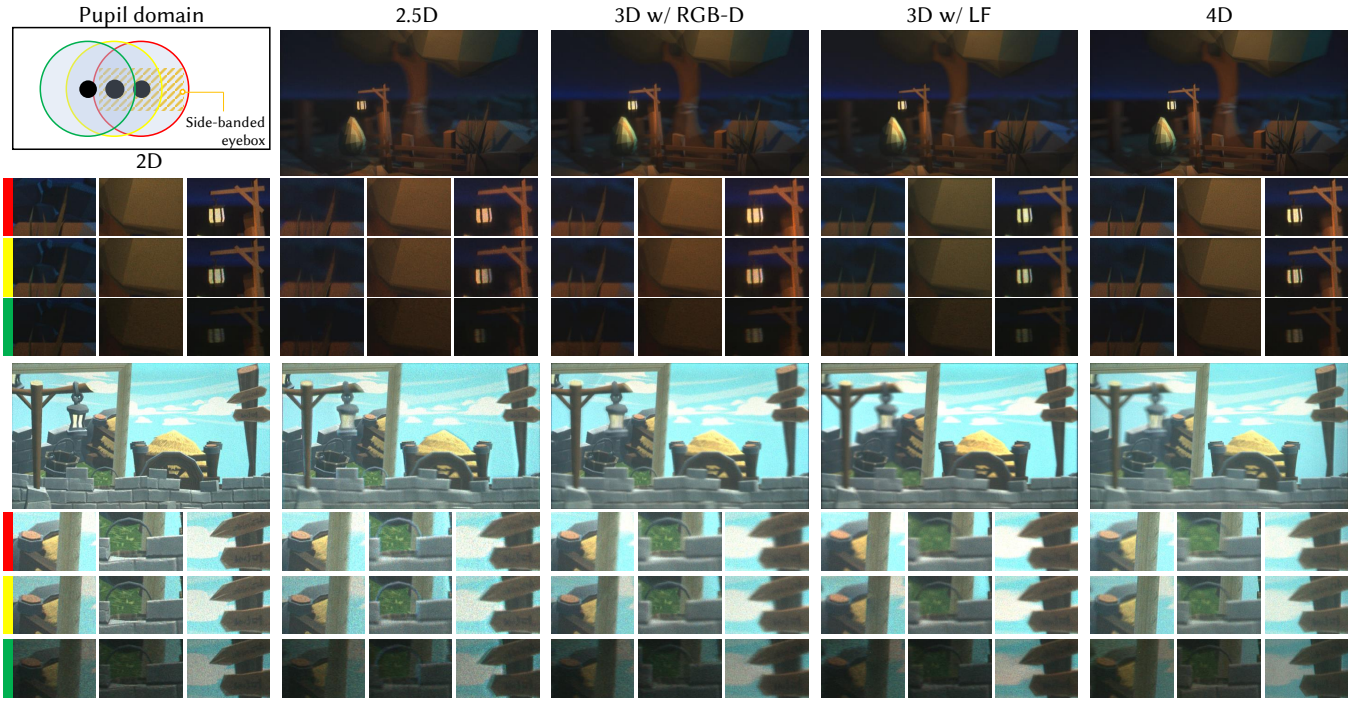

Fig. S16. Additional experimental results with different pupil positions. Holographic scenes supervised with 2D (1st col), 2.5D (2nd), 3D w/ RGB-D (3rd), 3D w/ LF (4th), 4D (5th) targets are captured with different pupil positions (red: $(x_{p,norm}, D_{p,norm}) = (0, 1.1)$, yellow: $(-0.68, 1.1)$ and green: $(-1.36, 1.1)$). The scenes are photographed with different focal states (landscape_night: 7th, village_mirror: 3rd) out of 9 distinct focal states equally sampled in diopter. Enlargements are provided with the image focused on the magnified object. We intentionally provide the results without modifying the brightness to show the energy across the eyebox. Note that it is hard to discriminate 3D w/ LF case and 4D case with the captured results.

distribution that eventually supports full eyebox irrespective of the CGH supervision targets.

S6.4 3D realism depending on the CGH supervision targets (user experiment 1)

Following the pairwise comparison, the responses from two subjects out of twenty-eight were identified as outliers and subsequently excluded from the analysis. To guide the outlier analysis, we referred to the work of Perez-Ortiz and Mantiuk [2017]. After removing the outliers, we estimated the confidence interval using the bootstrapping method. The statistical test was conducted using a two-tailed z-test on the JOD scores obtained for each viewing condition.

In detail, in the viewing condition of *Center*, statistically significant differences in JOD scores were found in most of the paired conditions (p<0.001: *4D* vs the other cases, *3D w/ LF* vs *3D w/ RGB-D*, *3D w/ LF* vs *2.5D*, p<0.05: *3D w/ LF* vs *2.5D*) except the *2.5D* vs *3D w/ RGB-D* (p= 0.37). In the viewing condition of *Decentered*, the significant results were found in the paired conditions (p <0.001: *4D* vs the other cases, p<0.01: *3D w/ LF* vs *3D w/ RGB-D*, p<0.05: *3D w/ LF* vs *2.5D*) except the *2.5D* vs *3D w/ RGB-D* (p=0.36). In case of *Vignetted*, significant differences were present in the paired conditions (p<0.001: *4D* vs the other cases, *3D w/ LF* vs *3D w/ RGB-D*, p<0.05: *3D w/ LF* vs *2.5D*) except the paired condition of *2.5D* vs *3D w/ RGB-D* (p=0.38). Lastly, in the case of *w/ head movement*, the JOD scores of paired conditions were significantly different (p<0.001: *4D* vs the other cases, *3D w/ LF* vs *3D w/ RGB-D*, *3D w/ LF* vs *2.5D*) except the paired condition of *2.5D* vs *3D w/ RGB-D* (p=0.85).

Additional raw data of pairwise comparison. In the paper, we initially presented scaled JOD results in Fig. 4. Here, we provide the raw data of vote counts obtained from pairwise comparisons, illustrated in Fig. S19. These counts depict the preference for the column option over the row option across four distinct viewing conditions accumulated with three evaluation scenes. The diagonal elements of the matrices are zero since the comparisons are only performed between different CGH supervision targets.

Additionally, we performed one-tailed Wilcoxon signed rank tests using the vote counts obtained from twenty-six subjects. It is important to note that these statistical results may slightly differ from those in Sec. 3, which were based on scaled JOD rather than raw vote counts. Even preferences below 0.75, equivalent to 1 JOD in the scaled unit, exhibit strong statistical significance in the non-parametric significance test. These findings further highlight the superior performance of 4D CGH supervision, surpassing even the performance of 3D w/ LF case under the *Center* viewing condition.

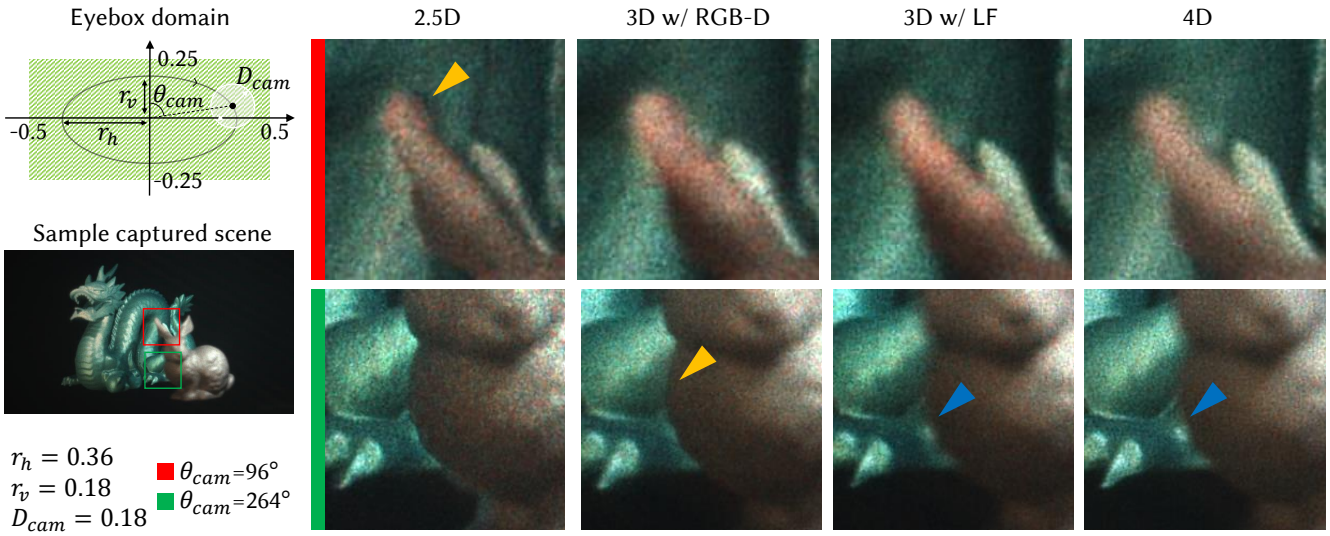

Fig. S17. Additional captured results showcasing different 3D CGH supervision approaches under distinct acquisition conditions (red and green) are presented below the sample captured scene. To maintain consistency in the specifications regarding pupil states, we have provided them with the normalized coordinates. Enlargements near the bunny's ear are provided for images captured under the red acquisition condition, while those near the bunny's body are provided for images captured under the green acquisition condition. These captured results offer a clearer insight into the issues within each CGH supervision method. Yellow arrows highlight the occlusion boundary problem, with the 2.5D case displaying discontinuity in the occlusion boundary, while the 3D w/ RGB-D still reconstructs a sharp occlusion boundary. Blue arrows in the 3D w/ LF and 4D cases indicate the issue with focal-stack-based targets even if they are processed with dense LF. This averaging in the focal stack generation procedure limits the reconstruction of view-dependent visual effects in the case of 3D w/ LF, particularly in sections with view-dependent visual effects.

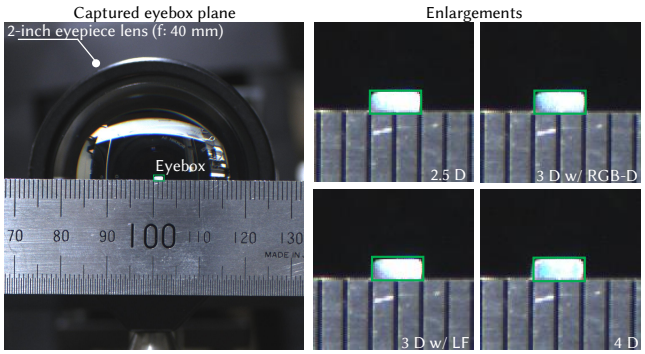

Fig. S18. Captured eyebox plane with camera is shown. The enlargements of captured eyebox (green rectangle) are provided depending on the CGH supervision approaches (2.5D, 3D w/ RGB-D, 3D w/ LF, and 4D). Note that the physical size of the entire eyebox is 2.2 mm × 1.1 mm for the illumination of 450 nm (blue channel) as the work exploits side-band filtering for complex modulation with an amplitude-only SLM. Slight misalignment between the near-eye display system and the camera that captured eyebox presents asymmetrical energy distribution, which is not the main focus of the figure.

S6.5 Eye tracking trajectory

To validate our assumptions regarding the continuous eye movements during the experience of holographic near-eye displays, we utilized an eye tracker to record the participants' pupil movements while they performed visual tasks. Fig. S20 illustrates the recorded eye movement trajectories of fourteen subjects during a single session, corresponding to each viewing condition.

The presented figure clearly demonstrates that the experiments were carried out in distinct sections of the eyebox as intended. Prior to the experiment, the center of the eyebox was determined in the global coordinate system and served as the reference point. The initial position of the pupil for each session was adjusted to achieve the desired deviations of 1.25 mm for the *Decentered* condition and 2.5 mm for the *w/ head movement* condition from the center position. However, there were slight variations in the average measurements among individuals. It is important to note that the average displacement for each viewing position is less than 1 mm away from the desired placement. Due to potential factors such as occlusion by the eyelid and blinking, we have only provided the measured pupil displacement of a few subjects for better visualization.

S6.6 3D realism depending on the number of views for 4D CGH supervision (user experiment 2)

In Sec. 4, the 3D realism of the holographic scene is evaluated depending on the number of views used in CGH supervision, and the JOD value is provided depending on the target stimuli. The landscape_day scene shows -1.88, -0.05, 0.70, and 1.22 JOD. The landscape_night scene demonstrates -1.31, -0.20, 0.51, and 1.00 JOD. Lastly, the village scene resulted in -0.88, -0.18, 0.47, and 0.59 JOD as shown in Fig. 5(A).

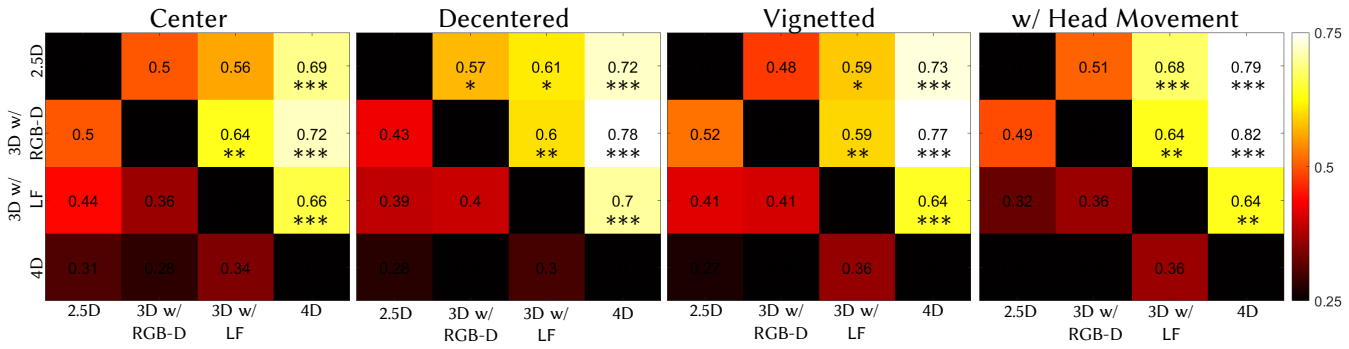

Fig. S19. Normalized matrices of comparisons based on varying viewing conditions are provided, reflecting preferences estimated by the vote counts favoring the column option over the row option. Additionally, the figure presents statistical significance determined via the one-tailed Wilcoxon signed rank test using the preferences of twenty-six subjects (*: p<0.05, **: p<0.01, and ***: p<0.001). The colorbar spans from 0.25 to 0.75, representing the range between -1 JOD and +1 JOD in the converted scale.

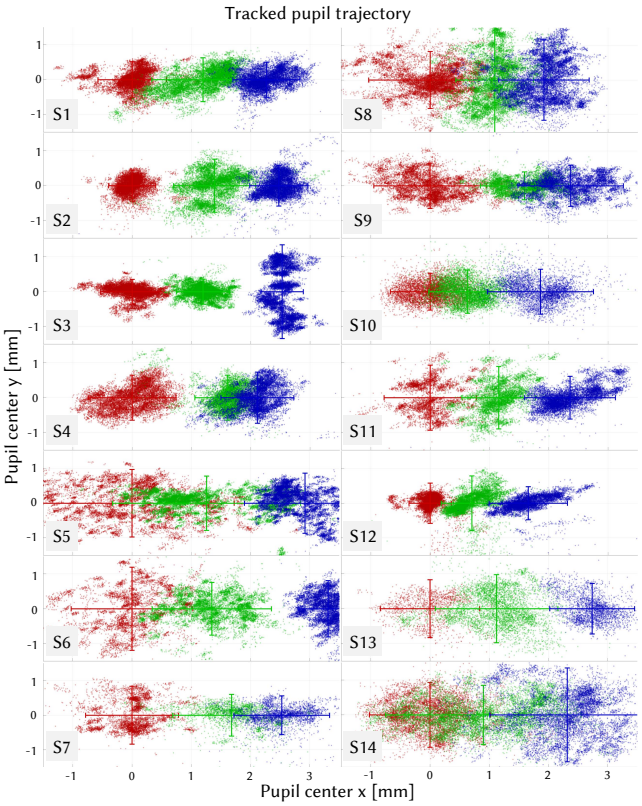

Fig. S20. Recorded eye trajectory of fourteen subjects during a single session when experiencing the holographic image of landscape_day scene. Each color dot indicates the recorded pupil position with different viewing conditions (red: *Center*, green: *Decentered*, blue: *Vignetted*). The horizontal displacement of the *Center* condition is regarded as zero, whereas the vertical displacements have been adjusted to have an average of zero for improved visualization. The error bars represent the 95% confidence interval of the pupil displacement recorded in each viewing condition.

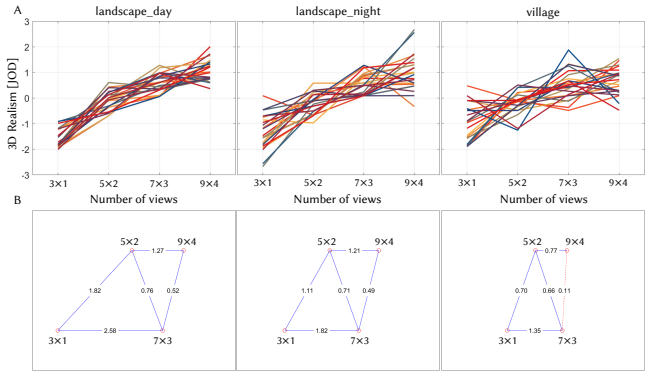

Fig. S21. Additional experimental results of user experiment #2 (1st column: landscape_day, 2nd column: landscape_night, 3rd column: village): (1st row) JOD scores based on the responses from individual subjects. Each line indicates the JOD scores of each view number case with an average of zero. (2nd row) The graphical demonstrations of the JOD scaling. The red circle indicates the number of view conditions, and the lines are interconnected with neighboring options. The blue line represents the statistical significance of the JOD score difference (p<0.05) between the paired conditions evaluated with a two-tailed z-test, as opposed to the red dashed line. The value on the individual line indicates the mean JOD difference of the paired conditions.

We provide additional experimental results as Fig. S21 acquired with the pairwise comparisons with conditions that differ in the number of views used in 4D CGH supervision. Fig. S21(A) demonstrates the scaled JOD scores in individual subjects depending on the stimuli. Notably, the experiment conducted with the landscape_day scene exhibited consistent and prominently positive slopes in individual preference results. In contrast, the experiment featuring the landscape_night scene displayed relatively smaller slopes, accompanied by greater variability in responses among individuals. Furthermore, the disparity in 3D realism based on the number of views proved to be notably minimal, and a significant number of subjects exhibited an inverted JOD in the paired option (7×3 vs 9×4).

The effect of perceived 3D realism depending on the number of views used for 4D CGH supervision is evaluated with a two-tailed z-test on the scaled JOD scores. Fig. S21(B) shows the graphical representation of the scaling and simultaneously demonstrates the statistical results. Most paired options elicited very strong statistical significance on the difference (p<0.001). The paired option of 7×3 vs 9×4 in the landscape_night scene showed strong evidence (p<0.01), while the paired option of 7×3 vs 9×4 in the village scene showed no evidence of the significance (p=0.45).

S7 DISCUSSION

In this section, we provide additional discussions not held in the main paper.

S7.1 Display types for 3D perceptual testbeds

As summarized in Table 1, data formats and CGH supervision techniques can be evaluated based on standard visual cues and scene representation capacities like defocus blur and view-dependent effects. Light field and its supervision for holographic displays uniquely support accurate view-dependent effects, including occlusion, parallax, and specular highlights. Table S1 compares various accommodation-supporting displays for 3D scene perceptual testbeds [Hoffman et al., 2008, Mercier et al., 2017, Shibata et al., 2011, Zhong et al., 2021]. Holographic displays, might now offer the best testbed for perceptual studies, with their flexibility to represent arbitrary data formats, and their improved image quality with recent advances [Peng et al., 2020, Shi et al., 2021]. Despite the limitations of a small eyebox, our study finds that parallax content (4D light field) greatly enhances perceptual realism, even with a narrow eyebox. We expect the perceptual impact to grow as the étendue of holographic displays increases.

S7.2 Trade-off between computation efficiency and quality

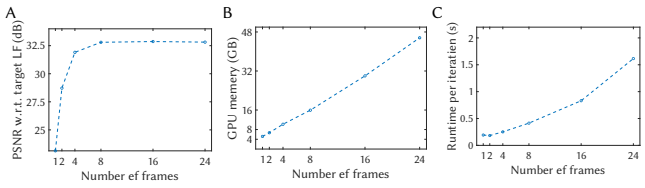

Fig. S22. Trade-off in computation efficiency and performance in time-multiplexed holographic displays. Here, we show (A) light field fidelity measured by PSNR, (B) reserved GPU memory, and (C) runtime per iteration using various numbers of frames.

In this work, we have demonstrated that light field optimization using gradient-descent-based methods, which directly optimize for the raster light fields, is the most effective approach for achieving perceptual realism in 3D holographic displays. Notably, this achievement is enabled by time-multiplexed holographic display engines utilizing fast SLMs [Choi et al., 2022, Kim et al., 2022]. Consequently, the computational resources required for calculating multiple frames of phase/amplitude patterns increase. Here, we present data on the time, performance, and memory usage in relation to the number of frames. In Fig. S22, we report our results from simulating our holographic display system with varying numbers of frames across four different scenes. We calculate the average performance (A), reserved GPU memory (B), and runtime per iteration (C). We run 2000 iterations for 7×3 views on an NVIDIA RTX A6000 GPU for this comparison, and the optimization almost converged after 1000 iterations (See Fig. S3). We note that the fidelity plateaus after 8 frames for our binary case, which implies that we could effectively allocate frames for different figures of merit, such as 'étendue'. This would be an interesting avenue for future work. It is important to note that when using 24 frames, the simulation may take hours to complete thousands of iterations. Faster generation of light field holograms is probably one of the most interesting problems to tackle in future work, which we expect to solve using the deep learning-based method [Peng et al., 2020, Shi et al., 2021] that already has shown exciting progress.

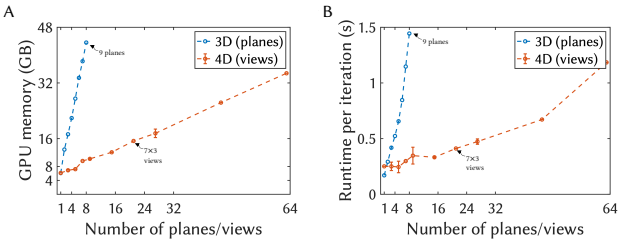

Fig. S23. (*left*) GPU memory and (*right*) computation time as a function of the number of planes for 3D (blue, focal stack) and views for 4D (red, light field) CGH optimizations. The memory and runtime for each optimization type are represented on the same graph, with the x-axis indicating the number of planes or views, respectively. The errorbar in 4D CGH supervision represents the standard deviation of each measured value with combinatorial candidates (e.g. 9: 9×1, 3×3).

In Fig. S23, we also compare the computation speed (per iteration) and memory usage for focal stack and light field optimizations on the same GPU. Using 8 frames, we were only able to optimize up to 9 planes using a 48 GB GPU, whereas for light field optimizations, we could optimize up to 63 views ($= 9 \times 7$). We could optimize a larger number of views than planes because the STFT operations used in light field optimizations share correlations between the graphs.

Similar to CGH supervision using light field data, 3D CGH supervision encounters limitations in memory and runtime per iteration. These approaches necessitate substantial memory allocation to retain the focal stack targets for the sampled depths. Additionally, the wave propagation relies on the fast Fourier transform-based angular spectrum method [Goodman, 2005], causing resource demands to increase proportionally with the number of planes.

S7.3 Trade-off between degrees of freedom, the number of constraints, and étendue

In Fig. S24, we present the loss values for various numbers of degrees of freedom, constraints, and CGH methods. Specifically, we run the gradient-descent-based optimization for CGH methods as described in the Appendix of the manuscript with different degrees

	resolution	eye-tracking required	retinal blur class	monocular occlusion/parallax	image quality	eyebox
fixed focus [Cakmakci and Rolland, 2006]	high	no	incorrect	not supported	high	wide
varifocal [Mercier et al., 2017]	high	yes	rendered	not supported	high	wide
fixed multifocal [Zhong et al., 2021]	moderate	yes	near-correct	optimized	moderate	narrow
attenuated layers [Huang et al., 2015]	low	optional	near-correct	correct	moderate	moderate
integral imaging [Lanman and Luebke, 2013]	moderate	no	near-correct	correct	high	moderate
holographic [Kim et al., 2022]	high	no	correct	correct	mid-high	narrow

Table S1. Assessment of various displays that support accommodation for a perceptual testbed, based on optical and perceptual criteria. A large portion of the criteria and evaluations are adapted from Matsuda et al. [2017]. It is noteworthy that recent advances in holographic displays have significantly improved image quality, which we classify as mid-high [Peng et al., 2020, Shi et al., 2021]. Eyebox defined as 'moderate' ranges from 5–10 mm. We exclude computational complexity and form factor from the comparison, as perceptual studies can be conducted on prototypes with pre-rendered content.

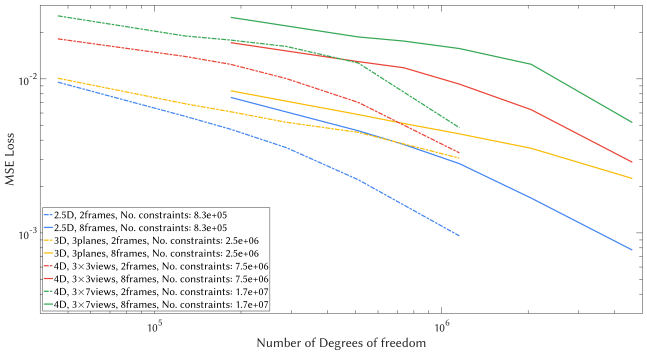

Fig. S24. Loss values vs the number of degrees of freedom. We perform additional simulations on these factors to verify the trend. As expected, increased degrees of freedom or a smaller number of constraints lead to low loss values.

of freedom in SLMs and number of frames. The number of degrees of freedom is calculated as (Number of optimizable pixels) × (Number of frames). To tune the number of optimizable pixels, we assume the SLMs with the same size but with larger pixel pitches. In practice, we set superpixels, so it has optimizable pixels number of $1920 \times 1200, 960 \times 600, \cdots, 192 \times 120$, and upsample them so that they match the original SLM pixel resolution in the optimization pipeline. This implementation allows us to match the feature size in the simulation pipeline while varying the number of optimizable pixels. Thus, having the same number of degrees of freedom with different frame counts implies larger superpixels, leading to lower performance within the same color category. The number of constraints is calculated by (Number of pixels in the target (ROI)) × (number of views) × (number of planes) . We ran 1,000 iterations, set the ROI as central 1190×700 pixels, and the learning rate to 0.1 for 2 frames and 0.4 for 8 frames. As expected, increased degrees of freedom or a smaller number of constraints lead to low loss values. However, the commonly used mean square error metric should not directly represent the perceptual performance and exploring metric functions for constraining the spatial-angular information of light or perceptual realism would be an interesting direction for future study [Kiran Adhikarla et al., 2017].

S7.4 Eye rotation requirements

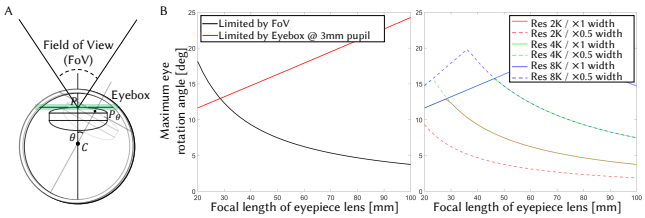

Fig. S25. Required eye rotation angle in the given holographic near-eye displays that presents a trade-off relationship between FoV and size of eyebox. (A) Schematic of eye when the near-eye displays present an image with a specific field of view (FoV) and eyebox. The eye rotates based on the center of rotation (C). Here, the FoV and the eyebox are defined by the focal length of the eyepiece lens and the SLM's resolution and pixel pitch. (B) (left) The FoV and the eyebox limit the overall maximum rotation angle and (right) the maximum rotation angle can be plotted in different SLM specifications.

Throughout the user study, our primary assumption revolved around the near-eye display's eyebox size being smaller than the pupil size, which we refer to underfilled pupil. However, we highlight the effectiveness of 4D CGH supervision when the eyebox surpasses the eye pupil, allowing partial sampling and clearer image disparity. It is important to note that not all near-eye display systems facilitate proper eye rotation movement. To elucidate the necessity of eye rotation in near-eye display configurations, we provide Figure S25. In Figure S25(A), a schematic eye interacting with a near-eye display, featuring a specific field of view and eyebox, is illustrated. Here, our assumption involves the eye rotating around the center of rotation (C) without additional translational movement, and aligns the visual axis and optical axis. The center of rotation is approximated as 10 mm behind the center of the iris (P).

In this context, the eye's rotation range is constrained by two key factors: the FoV and the eyebox. The maximum eye rotation within the FoV refers to the highest angle the visual axis can pivot to reach the edge of the FoV. Simultaneously, the eye's rotation within the eyebox is computed by multiplying the distance between the eye pupil and the rotation center ($\overline{PC}$) by the rotation angle (θ), ensuring it remains within the limits of the eyebox. We also

account for an additional rotational angle allowance of half of the pupil diameter. The determination of the near-eye display system's maximum required rotation angle involves selecting the smaller value between these two calculations.

Analyzing Fig. S25(B), as the focal length of the eyepiece lens increases, the rotation angle limited by the FoV decreases, while the restriction imposed by the eyebox expands. However, our primary focus remains on the minimum value between these limitations. For this simulation, a pupil diameter of 3 mm was assumed. Consequently, a shorter focal length for the eyepiece, which has a large chance of being an underfilled pupil, induces maximum eye rotation in the near-eye display setup. Conversely, opting for a longer focal length eyepiece, resulting in an overfilled pupil scenario, does not significantly prompt eye rotation due to its highly restricted FoV. This trend relaxes in an étendue-expanded scenario with high-resolution SLMs, as depicted in Fig. S25. However, a decrease in the pixel pitch of the SLM to suit the near-eye display setup necessitates an eyepiece with a shorter focal length. Note that recent near-eye displays develop in shortening the eye relief along with the focal length to minimize the overall size.

S7.5 Debate in multi-focal vs. multi-view

In conventional autostereoscopic 3D displays, specific data formats are mandated by systematic constraints. For instance, a multi-layer scheme supports focal-stack-based imagery exclusively, while the multi-view scheme displays multiple view images but with reduced spatial resolution. However, holographic displays can reconstruct both 3D data formats (multi-focal and multi-view) and the spatial-angular resolution trade-off of the multi-view scheme is relatively relaxed compared to the displays with integral imaging. Please refer to the Table S1 for the comparisons. This intriguing capability of holographic display sparks a debate that warrants thorough discussion.

The target of i-th focal stack (fs_i) can be acquired with the given light field map as follows:

$$fs_i = \sum_v w_v \cdot l_v \left(x - x_{i,v}, y - y_{i,v}\right), \tag{S6}$$

where, l_v is a v-th 2D slice of the light field, w_v is a constant stating the view-dependent weight, and $(x_{i,v}, y_{i,v})$ is the coordinate translation depending on the view and depth. This indicates that the focal stack is a linear combination of the translated 2D slice of the light field and the view-dependent weight is usually unitary.

The perceived retinal image depending on the pupil state (p) and the focal state (j) of the eye can be simplified as follows:

$$\begin{aligned} I_{p,j} &= \sum_i psf_{p,i,j} * fs_i \\ &= \sum_i psf_{p,i,j} * \sum_v w_v \cdot l_v \\ &\neq \sum_v psf_{p,j,v} * l_v \end{aligned} \tag{S7}$$

Here, $psf_{p,i,j}$ is the displacement-dependent point spread function depending on the depth of the focal stack and the focal state of the eye. As stated in Sec. S5.1.4, the point spread function is also a function of pupil displacement, and it should not be pre-defined in the rendering stage for precise visualization. Thus, the approximation of a volumetric scene into a set of focal stacks can suppress the reconstruction of view-dependent information and ultimately deteriorate the 3D visual experience.

S8 APPENDIX

In our study, we opted to scale the perceptual difference based on vote counts obtained through pairwise comparisons, as opposed to employing a direct rating system (mean-opinion score, as referenced by Hossfeld et al. [2016]). Direct rating requires pre-trained subjects to assign a unified score to results with multi-dimensional differences, and individual rating scores tend to vary. In contrast, pairwise comparison is a simpler method that is well-suited for non-experts. Furthermore, this approach offers results with low measurement errors [Shah et al., 2016] and can utilize sparse sampling with adaptive experimental procedures [Mantiuk et al., 2012].

S8.1 Statistics in pairwise comparison

For the analysis of the results obtained from the pairwise comparison, we referred to the work of Perez-Ortiz and Mantiuk [2017] and utilized their released code from the GitHub repository (https://github.com/mantiuk/pwcmp). Here, we briefly summarize the analysis for a better understanding.

Analyzing the statistical difference using JOD-scaled data is more complex compared to the statistical test with direct rating experiments or accumulated vote counts. This is because the JOD scores are interconnected and not independent from one another. In detail, the correlation between the scaled JOD scores among the options can be determined by examining the covariance matrix C obtained during JOD scaling. In that case, if we consider the pairwise comparison of n conditions, which are scaled in JOD scores as $q = (q_1, \ldots, q_n)$, the score difference between two conditions, say i and j, can be calculated as $q_{ij} = q_i - q_j$. The variance for this score difference is given by $v_{ij} = c_{ii} + c_{jj} - 2c_{ij}$, where c_{ii} and c_{jj} represent the diagonal elements and c_{ij} represents the off-diagonal element in the covariance matrix C. Based on this, we can assume that the z-value of the score difference follows a normal distribution, represented as $z_{ij} = q_{ij}/\sqrt{v_{ij}} \sim N(0, 1)$. A two-tailed z-test is employed to determine if we can reject the null hypothesis that there is no difference in JOD scores between the two conditions, with a specified level of confidence.

REFERENCES

O. Cakmakci and J. Rolland. Head-worn displays: a review. *Journal of display technology*, 2(3):199–216, 2006.

S. Choi, M. Gopakumar, Y. Peng, J. Kim, M. O'Toole, and G. Wetzstein. Time-multiplexed neural holography: a flexible framework for holographic near-eye displays with fast heavily-quantized spatial light modulators. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022.

A. Duane. Normal values of the accommodation at all ages. *Journal of the American Medical Association*, 59(12):1010–1013, 1912.

J. W. Goodman. *Introduction to Fourier optics.* Roberts and Company publishers, 2005.

B. Guenter, M. Finch, S. Drucker, D. Tan, and J. Snyder. Foveated 3d graphics. *ACM transactions on Graphics (tOG)*, 31(6):1–10, 2012.

D. M. Hoffman, A. R. Girshick, K. Akeley, and M. S. Banks. Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of vision*, 8(3): 33–33, 2008.

T. Hoßfeld, P. E. Heegaard, M. Varela, and S. Möller. Qoe beyond the mos: an in-depth look at qoe via better metrics and their relation to mos. *Quality and User Experience*, 1:1–23, 2016.

F.-C. Huang, D. P. Luebke, and G. Wetzstein. The light field stereoscope. In *SIGGRAPH emerging technologies*, pages 24–1, 2015.

D. Kim, S.-W. Nam, B. Lee, J.-M. Seo, and B. Lee. Accommodative holography: improving accommodation response for perceptually realistic holographic displays. *ACM Transactions on Graphics (TOG)*, 41(4):1–15, 2022.

V. Kiran Adhikarla, M. Vinkler, D. Sumin, R. K. Mantiuk, K. Myszkowski, H.-P. Seidel, and P. Didyk. Towards a quality metric for dense light fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 58–67, 2017.

R. Konrad, A. Angelopoulos, and G. Wetzstein. Gaze-contingent ocular parallax rendering for virtual reality. *ACM Transactions on Graphics (TOG)*, 39(2):1–12, 2020.

D. Lanman and D. Luebke. Near-eye light field displays. *ACM transactions on graphics (TOG)*, 32(6):1–10, 2013.

B. Lee, D. Kim, S. Lee, C. Chen, and B. Lee. High-contrast, speckle-free, true 3d holography via binary cgh optimization. *Scientific reports*, 12(1):2811, 2022.

C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5): 978–994, 2010.

R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk. Comparison of four subjective methods for image quality assessment. In *Computer graphics forum*, volume 31, pages 2478–2491. Wiley Online Library, 2012.

N. Matsuda, A. Fix, and D. Lanman. Focal surface displays. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017.

O. Mercier, Y. Sulai, K. Mackenzie, M. Zannoli, J. Hillis, D. Nowrouzezahrai, and D. Lanman. Fast gaze-contingent optimal decompositions for multifocal displays. *ACM Transactions on Graphics (TOG)*, 36(6):1–15, 2017.

I. C. on Non-Ionizing Radiation Protection et al. Guidelines on limits of exposure to laser radiation of wavelengths between 180 nm and 1,000 μm. *Health Physics*, 71(5): 804–819, 1996.

J.-H. Park. Recent progress in computer-generated holography for three-dimensional scenes. *Journal of Information Display*, 18(1):1–12, 2017.

Y. Peng, S. Choi, N. Padmanaban, and G. Wetzstein. Neural holography with camera-in-the-loop training. *ACM Trans. Graph.*, 39(6):1–14, 2020.

M. Perez-Ortiz and R. K. Mantiuk. A practical guide and software for analysing pairwise comparison experiments. *arXiv preprint arXiv:1712.03686*, 2017.

N. B. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramch, M. J. Wainwright, et al. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*, 17(58):1–47, 2016.

L. Shi, B. Li, C. Kim, P. Kellnhofer, and W. Matusik. Towards real-time photorealistic 3d holography with deep neural networks. *Nature*, 591(7849):234–239, 2021.

T. Shibata, J. Kim, D. M. Hoffman, and M. S. Banks. The zone of comfort: Predicting visual discomfort with stereo displays. *Journal of vision*, 11(8):11–11, 2011.

B. A. Wandell. *Foundations of vision*. sinauer Associates, 1995.

Z. Zhang and M. Levoy. Wigner distributions and how they relate to the light field. In *2009 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10. IEEE, 2009.

F. Zhong, A. Jindal, Ö. Yöntem, P. Hanji, S. Watt, and R. Mantiuk. Reproducing reality with a high-dynamic-range multi-focal stereo display. *ACM Transactions on Graphics*, 40(6):241, 2021.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009