Language-Conditioned Object Highlighting for Simulated Prosthetic Vision

Julia Tomas-Barba[1], Celia Gines-Alcober[1] Alejandro Perez-Yus[1], Jesus Bermudez-Cameo[1], and Michael Beyeler[2]

[1] Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, Spain
{j.tomas,840910,alperez,bermudez}@unizar.es
[2] Department of Computer Science, Department of Psychological and Brain Sciences. University of California, Santa Barbara, USA
mbeyeler@ucsb.edu

Abstract. Retinal prostheses offer a promising avenue for partial vision restoration, yet their utility remains severely hindered by the perceptual limitations of current implants. Existing pipelines often attempt a global reconstruction of the visual scene; however, given the extremely low information bandwidth of these devices, this approach inevitably sacrifices critical detail, causing small or distant objects to vanish. While recent paradigms have explored scene simplification, they often rely on rigid saliency maps, lack realistic egocentric dynamics, or neglect non-linear distortions like axonal deformations. To address these challenges, we propose a modular framework that bridges open-vocabulary semantic understanding with biophysical phosphene optimization. Our pipeline leverages vision-language and segmentation foundation models to dynamically isolate target objects from natural language prompts. This object-conditioned visual target is then processed by a bio-inspired phosphene autoencoder, trained with a novel region-weighting strategy that optimizes electrical stimuli to maximize the perceptual fidelity of user-prioritized targets. In simulation, our framework significantly improves the preservation of small, task-relevant objects across varying shapes and scales, outperforming global reconstruction and heuristic baselines in both perceptual fidelity and geometric alignment.

Keywords: Simulated prosthetic vision · Retinal prostheses · Egocentric vision · Language-grounded segmentation · Phosphene encoding

1 Introduction

According to the World Health Organization (WHO), over 2.2 billion people worldwide live with visual impairment, including more than four million who are profoundly blind [37]. This motivates the development of visual neuroprostheses that can convey useful scene information through sparse, device-mediated percepts [2]. Across retinal and cortical implants, a central computational challenge is not high-fidelity image reconstruction, but deciding which visual information

should be preserved under severe bandwidth and distortion constraints. Devices like the Argus II [18] and the PRIMA system [23] have demonstrated the feasibility of eliciting artificial visual perceptions (phosphenes) by translating video feeds from a wearable camera into electrical stimuli via external image processing units. However, severe constraints of the implants such as low spatial resolution, perceptual instability, and complex spatial distortions persist, restricting the resulting visual acuity to legal blindness criteria [1, 24] and severely hindering patient autonomy and reliable interaction with complex, dynamic environments.

To mitigate these constraints, recent paradigms have shifted toward deep learning architectures such as end-to-end autoencoders. Given an input target image, the encoder generates the optimized electrical stimuli, which are then passed to a Simulated Prosthetic Vision (SPV) decoder. Current biophysically inspired decoders realistically model the complex spatial distortions involved in phosphene perception [4, 32]. Traditionally, the underlying objective of these autoencoders is to optimize the stimuli so that the elicited visual perception matches the target image pixel-wise as accurately as possible, given the known implant configuration and patient-specific variations. However, defining visual fidelity through uniform, pixel-wise reconstruction is poorly matched to low-bandwidth, task-driven prosthetic vision. Attempting to reproduce complex natural scenes forces the limited stimulation capacity to be distributed across the entire image, as these models lack semantic awareness and treat all pixels equally. Consequently, small or distant informative cues vanish from the perceptual experience, as activating phosphenes for minor targets fails to overcome the global reconstruction loss of the complete scene. Ultimately, this weakens the representation of task-relevant objects and may reduce the utility of the percept for object localization, interaction, or navigation.

In response to real-world visual complexity, alternative paradigms have explored leveraging computer vision and AI for scene simplification and object-highlighting strategies, shifting the focus from uniform scene reconstruction toward downstream task utility. These models aim to emphasize semantically relevant information (such as salient objects or navigable paths) while removing irrelevant background noise. However, three critical limitations in object-aware encoding still persist: First, they often rely on highly simplified, static simulation models, such as standard Gaussian profiles or binary activations, that completely neglect non-linear clinical distortions like axon map deformations and patient-specific variations [5,12]. Second, experimental evaluation is frequently restricted to curated, static datasets featuring low visual complexity or pre-segmented scenes [12,13]. These approaches fail to account for the chaotic, egocentric video dynamics of real-world environments faced by prosthesis wearers, where unpredictable lighting, continuous camera movement, and severe occlusions drastically degrade performance. Third, existing systems offer limited autonomy for user-driven object selection. Current frameworks rely on automatic saliency maps, or they restrict active interaction to rigid, closed-set modalities (e.g., specific object classes or localized gaze-cropping). Consequently, the patient cannot easily prioritize specific visual elements based on their needs or context unless adapting

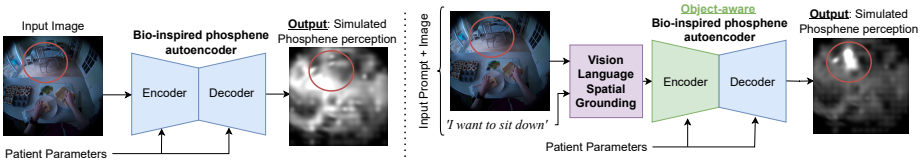


Fig. 1: Comparison of purely pixelwise reconstruction-based autoencoders (left) with our proposed pipeline that highlights specific objects in the scene given natural language expressions (right).

to the constraints of the system. Ideally, an assistive device should do the exact opposite: adapt to the patient's intent through flexible and natural queries.

To address these challenges, we propose a modular, object-aware framework that bridges open-world semantic understanding with biophysical phosphene optimization. Our two-stage system contains a Vision-Language Spatial Grounding module to isolate target objects via flexible natural language prompts, and a bio-inspired phosphene autoencoder that optimizes electrical stimuli to highlight these objects over the background (see Fig. 1). We argue that this two-stage pipeline is the most efficient and robust way to address this problem: Rather than attempting to learn complex semantics from scratch, we leverage the state-of-the-art capabilities of interchangeable foundation models and LLMs, seamlessly integrating them with an enhanced phosphene autoencoder. We instantiate and evaluate this framework in retinal prosthetic vision by utilizing a biophysically validated retinal model that incorporates realistic axonal distortions [12]. By fine-tuning this autoencoder on a real-world egocentric dataset [27] with object masks, our framework ensures that highlighted objects remain functional and perceivable regardless of scale or distance. However, the broader encoding problem is not retina-specific: the same language-grounded selection-and-encoding strategy could be adapted to other visual neuroprostheses (e.g., cortical implants) simply by replacing the perceptual decoder.

In summary, the main contributions of this work are as follows:

- We propose an object-aware training methodology for a bio-inspired phosphene autoencoder. By introducing a novel region-weighting strategy, we successfully shift the network's behavior from generic scene reconstruction to target-specific highlighting, preserving small and hard-to-perceive objects.
- We integrate a Vision-Language Spatial Grounding mechanism into phosphene-based vision, enabling users to isolate targets dynamically through flexible natural language instructions without closed-set limitations.
- We evaluate the effectiveness of our approach using real-world egocentric vision benchmarks and simulated prosthetic vision. Experimental validation includes rigorous quantitative evaluation of the encoding strategy and qualitative demonstrations of the complete language-guided pipeline, providing a computational testbed that paves the way for future wearable deployment and human-subject validation.

2 Related work

2.1 Deep Learning for visual prostheses encoding

Early efforts in bionic vision relied on hand-crafted biophysical models, such as Gaussian profiles [7], to simulate phosphene perception. However, subsequent clinical studies reported phosphenes as elongated, irregular shapes due to the activation of passing axon pathways. Mathematical models have been proposed to characterize these effects, including non-linear axon deformations [3, 5, 11] and temporal persistence [15]. These collective advancements culminated in standardized simulation frameworks like *pulse2percept* [4]. To avoid manual engineering of phosphene elicitation, recent works shifted toward end-to-end autoencoders [32]. These frameworks use a deep learning decoder as a proxy for patient perception, enabling the encoder to optimize stimuli directly. While some of these frameworks use fixed patient models [13, 38], recent paradigms introduced dynamic clinical variability with patient-specific prosthesis models, also integrating a human-in-the-loop optimization protocol for parameter calibration [12]. Contemporary architectures have built upon this decoder mode, adopting Vision Transformers to process more complex, unconstrained imagery keeping the desired personalized parametrization [33]. We adopt this state-of-the-art end-to-end paradigm as our base autoencoder model to apply our proposed region-weighting training strategy.

2.2 Object-Oriented Strategies in Prosthetic Vision

Several approaches have explored object-oriented representations to improve the functional utility of prosthetic vision. Saliency estimators and semantic segmentation models have been used to identify and isolate task-relevant regions in both static [19] and dynamic scenes [14], combined with scene layout structure [31], or attention-based architectures during mobility and navigation tasks [36]. In [26] they propose to segment and highlight objects while dimming the background instead of pure background removal, to keep scene context. In addition, interaction-driven strategies leverage user signals to guide visual selection more directly, including speech-based filtering [35], or gaze [22, 25]. However, these strategies are fundamentally restricted to automatic saliency maps, rigid closed-set vocabularies or require specialized hardware (e.g., an eye tracker). They fail to provide the flexible, open-world spatial selection necessary for true user autonomy, which remains a core objective of our work.

2.3 Vision-Language and Segmentation Models for Assistive Perception

Recent advances in open-world object detection [30, 34] have enabled semantic understanding in complex visual scenes without closed-set restrictions. Additionally, promptable segmentation models like SAM [29] have been upgraded to

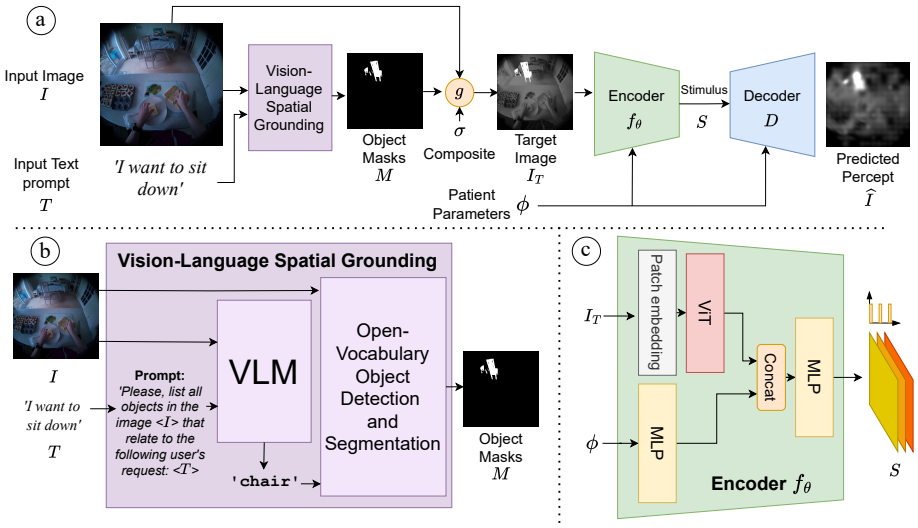


Fig. 2: Architecture of the proposed framework. (a) Complete pipeline, from the input image from the wearable camera (I) and user's request text prompt (T) to the final predicted percept $\hat{I}$. (b) Detail of the Vision-Language Spatial Grounding module. (c) Inner architecture of the Encoder (f_θ).

support concept segmentation in the latest version [6], achieving highly accurate masks for arbitrary concepts. Concurrently, large Vision-Language Models (VLMs) [8,39] have demonstrated strong capabilities in high-level reasoning and structured scene description, spawning a family of methods that combine VLMs with segmentation [40,41]. These models are increasingly integrated into assistive systems to provide context-aware feedback for visually impaired users [16,17,20]. However, these approaches primarily focus on delivering secondary outputs like auditory descriptions or ranked object summaries [21]. Therefore, these outputs remain decoupled from the non-linear, low-bandwidth constraints of electrical neurostimulation, as they lack the ability to actively shape stimulation patterns. Our work bridges this gap by integrating text-conditioned foundation models directly with differentiable biophysical decoders, transforming natural language intent into optimized electrical stimulation patterns.

3 Methods

3.1 System Overview

The proposed framework translates a real-world visual scene into an optimized prosthetic stimulation pattern conditioned on user-specified task instructions. As illustrated in Fig. 2a, the pipeline operates over three primary inputs: (i) a visual frame $I \in \mathbb{R}^{H \times W}$ captured by a wearable egocentric camera, (ii) a natural language request T provided by the user, and (iii) a patient-specific

parameter vector ϕ calibrating the physiological constraints of the individual's retina according to established clinical frameworks [12].

The execution pipeline is organized into two main sequential stages. First, the *Vision-Language Spatial Grounding* stage (Sec. 3.2) parses the natural language instruction alongside the input image I to generate a binary semantic mask M isolating the target objects, which is then blended with the scene to construct an object-conditioned visual target I_T. Subsequently, the *Bio-inspired Phosphene Autoencoder* stage (Sec. 3.3) processes the spatial target I_T and the simulated patient-specific model parameters ϕ to render the predicted phosphene percept $\hat{I}$. Specifically, this stage employs a neural encoder f_θ to generate an electrical stimulation matrix S, which is then routed through a biophysical retinal model D [12] to approximate the expected percept $\hat{I}$ under the assumptions of the simulated patient-specific model.

Unlike previous autoencoder-based approaches that uniformly reconstruct the input image, our objective is to improve the simulated perceptual fidelity and geometric preservation of prioritized target regions, for which we propose a novel region-weighting training strategy as another core contribution (Sec. 3.4). Our modular architecture leverages off-the-shelf foundation models for spatial grounding, isolating the training process to the autoencoder. This design provides long-term flexibility, allowing each module to be independently upgraded.

3.2 Vision-Language Spatial Grounding

To support open-world user intent, the first stage of the framework leverages a Large Language Model paired with an open-vocabulary foundation model to perform spatial grounding from natural language. As shown in Fig. 2b, given a frame I as visual context and a natural language prompt T (e.g., a textual command, necessity, or request given by the user), the Vision Language Model (VLM) reasons over the prompt and provides a list of objects $L = \{\texttt{'chair'}, \texttt{'door'}, \dots\}$ present in I that are related to T. This step exploits zero-shot capabilities of modern VLMs, providing enough flexibility for a wide range of tasks without requiring specific fine-tuning. Then, an open-vocabulary object detector and segmentation model finds all instances of L that are present in I, obtaining pixel-wise grounding. To accomplish this, we leverage models such as SAM 3 [6] that are able to segment instances directly from textual prompts, with high recall and contour accuracy.

Therefore, this stage maps the textual context onto the spatial coordinates of the image, generating a binary semantic mask $M \in \{0,1\}^{H \times W}$. This mask is subsequently utilized by the composition function $g(I, M)$ to construct an object-conditioned visual target I_T:

$$I_T = g(I, M) = M + \sigma \cdot (1 - M) \odot I, \tag{1}$$

where $\odot$ denotes the element-wise product, and $\sigma \in [0, 1)$ is a contrast hyperparameter. Here, the foreground is mapped to a high-contrast target rather than preserving within-object texture, and the background is attenuated by σ to emphasize the target region.

3.3 Bio-inspired Phosphene Autoencoder

The stimulation optimization framework utilizes an end-to-end differentiable autoencoder paradigm inspired by [12,33]. As illustrated in Fig. 2, the encoder f_θ translates the target image I_T and the patient parameter vector ϕ into the electrical stimulation matrix $S = f_\theta(I_T, \phi)$, where $S \in \mathbb{R}^{N_e \times 3}$ defines the pulse amplitude, frequency, and duration across the array of N_e electrodes. Although our encoder builds upon [33], in our implementation, we decouple the image from patient-specific parameters processing. Specifically, the ViT branch focuses exclusively on extracting global spatial features from I_T, while a parallel multilayer perceptron maps ϕ into a low-dimensional embedding. The outputs of both branches are then concatenated along the feature dimension and routed through a final synthesis MLP to output the localized electrical parameters S (see Fig. 2c).

To simulate the patient's visual perception, S and ϕ are processed by the non-trainable biophysical decoder D [12] to render the simulated predicted phosphene percept $\hat{I} = D(S, \phi)$. During training, the differentiability of D allows perceptual gradients to flow backward to update all trainable encoder parameters θ. This end-to-end optimization guides the ViT's attention mechanisms to align the electrical stimulation with the structural region of the target object defined in I_T.

3.4 Region-weighted Optimization

Unlike conventional phosphene autoencoders optimized for uniform scene reconstruction, our framework explicitly prioritizes target structures while retaining essential contextual cues. We optimize the network using a joint loss function composed of global pixel fidelity and localized target mask alignment:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{global}} + (1 - \alpha)\mathcal{L}_{\text{mask}}, \tag{2}$$

where α is a scalar hyperparameter that balances the optimization objectives.

Specifically, Global Pixel Fidelity ($\mathcal{L}_{\text{global}}$) computes the standard Mean Squared Error (MSE) between the target image I_T and the reconstructed phosphene simulation $\hat{I}$ to preserve the overall layout. To focus on target objects when they occupy small regions, Target Mask Alignment ($\mathcal{L}_{\text{mask}}$) isolates them using the semantic mask M. This term calculates the MSE exclusively within the object area, normalizing by its total pixel count:

$$\mathcal{L}_{\text{mask}} = \frac{1}{\sum M} \| M \odot (I_T - \hat{I}) \|^2, \tag{3}$$

This formulation encourages the encoder to allocate representational capacity to task-relevant regions while maintaining sufficient contextual information for spatial orientation and scene interpretation.

4 Experiments

4.1 Experimental Setup

Dataset. We develop and validate our framework using static frames extracted from the HD-EPIC dataset [27], which captures unconstrained egocentric activities in complex kitchen environments. The dataset provides a challenging benchmark characterized by large scale variations, severe occlusions, and highly cluttered backgrounds. From this repository, we curate a collection of 50,000 high-resolution RGB frames.

To obtain semantic supervision, pseudo-label instance masks are generated using the open-vocabulary Segment Anything Model 3 (SAM 3) [6], prompted with text queries describing relevant kitchen elements such as tools, containers, and food items. These masks are used as training and evaluation targets, but should not be interpreted as manually verified ground-truth annotations for open-world language grounding. To model varying clutter levels and multi-object scenarios, we construct combined target masks M for each frame. Given an image containing N_f valid masks, we sample a subset of k instances, where $k \sim \mathcal{U}(1, N_f)$, and merge the selected masks into a single target segmentation mask. This stochastic mask aggregation strategy encourages the model to learn stimulation patterns that generalize across varying numbers of objects, semantic categories, and spatial densities.

During training, each sample consists of a pair $\{I_T, M\}$. The target image I_T, synthesized via the composition function detailed in Sec. 3.2 with a $\sigma = 0.5$, serves as the primary visual input to the network, while the binary semantic mask M is utilized to compute the region-specific objective inside the loss function.

Biophysical Phosphene Simulator. To simulate patient perception, we employ the non-trainable differentiable biophysical decoder proposed in [12]. We simulate an epiretinal prosthesis consisting of a 15×15 square electrode array ($N_e = 225$ active channels) with a fixed electrode pitch of $400\,\mu$m. The decoder maps the generated stimulation pattern together with a 13-dimensional patient parameter vector ϕ to the final predicted phosphene image $\hat{I}$. The simulator model remains fixed throughout all training and evaluation procedures. During training, the patient parameters ϕ are randomly sampled within physiologically plausible ranges to improve robustness, except for the implant position parameters, which are held constant to preserve alignment between predictions and target images. During testing, the parameter vector ϕ is fixed for each $\{I_T, \phi\}$ pair to ensure reproducible evaluation.

Evaluation Metrics. To balance pixel-level correctness with human-centric perception, performance is evaluated across three dimensions:

- **Pixel-Level Reconstruction.** We report the Mean Squared Error (MSE) computed over the entire frame (MSE_{gl}). To evaluate region-specific reconstruction, we use the ground-truth semantic mask M to compute the target mask region error (MSE_{mask}), allowing us to analyze localized accuracy alongside global scene preservation.

- **Geometric Alignment.** To evaluate shape preservation under phosphene distortions, the simulated perception $\hat{I}$ is binarized across a range of thresholds τ_{IoU}. We select $\tau_{\text{IoU}} = 0.75$ as our primary threshold, because it corresponds to the mathematical midpoint between the attenuated background intensity ($\sigma = 0.5$) and the maximum foreground intensity (1.0). The Intersection over Union (IoU) is computed between the binarized perception and the pseudo-label mask M, and we report the mean IoU, over all test samples, denoted as $\text{mIoU}_{(0.75)}$, as one operating-point measure of shape preservation.
- **Semantic Consistency.** To quantify high-level semantic preservation under severe visual degradation, we extract visual embeddings using a pre-trained CLIP model [28] and compute the cosine similarity between the normalized feature representations of the simulated perception $\hat{I}$ and the target image I_T. We treat this metric as an exploratory proxy for semantic consistency, since CLIP embeddings are not validated measures of recognizability under phosphene vision.

Implementation Details. The proposed framework is implemented in PyTorch and optimized end-to-end utilizing an NVIDIA RTX 4090 GPU. The visual encoder is based on a Vision Transformer (ViT) backbone [10]. To prevent environmental data leakage, the curated dataset is partitioned into strict splits comprising 40,000 training frames, 9,000 validation frames, and 1,000 test frames. The final optimized model is trained for 15 epochs using the Adam optimizer with a batch size of 16. The learning rate is initialized at $\eta = 5 \times 10^{-6}$ and dynamically adjusted following a cosine annealing schedule. For the joint objective function formulated in Eq. (2), the loss weights α is set according to the optimal configuration identified in the ablation study (Sec. 4.2).

Baselines To assess the impact of our object-aware conditioning, we compare our method against two alternative pipelines based on the same global reconstruction network:

- **Global Reconstruction Baseline:** This baseline corresponds to the visual encoding network optimized using a combination of global MSE and SSIM on ImageNet [9], without any mask-based conditioning, following [33]. All image regions are treated uniformly, making it the reference for standard global scene reconstruction.
- **Heuristic Morphological Baseline:** This variant applies a rule-based preprocessing stage to the target masks before passing them through the same global reconstruction network. We enlarge small object masks with adaptive dilation depending on size and aspect ratio before composing the object-conditioned target I_T. This baseline tests whether hand-designed mask operations can recover targets lost under aggressive spatial downsampling.

For a fair interpretation, all comparisons should be read in light of the training data, objective functions, and preprocessing available to each baseline. The primary controlled comparison is between models evaluated on the same object-conditioned input I_T.

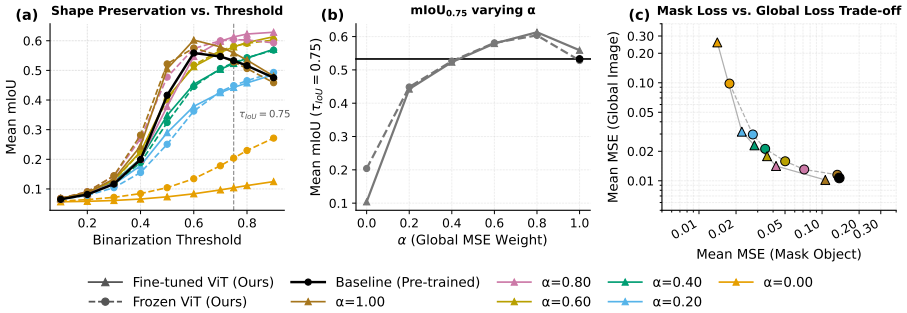


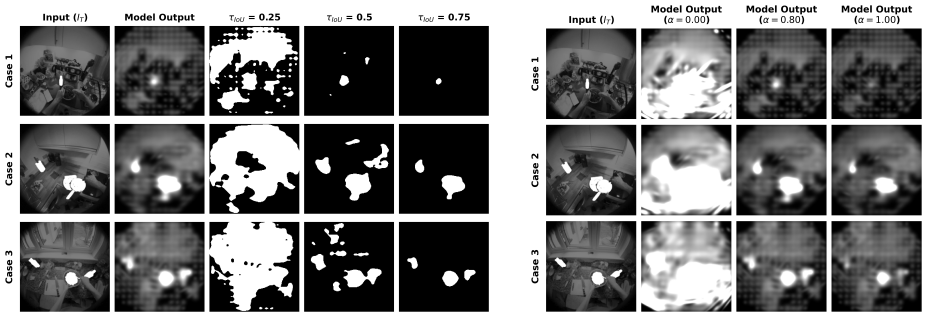
Fig. 3: Quantitative ablation study of hyperparameter sweeps and architectural constraints. The unified composite panel illustrates: (a) system empirical sensitivity across continuous binarization thresholds τ_{IoU}, (b) sensitivity analysis evaluating the mIoU performance under a fixed binarization threshold of $\tau_{\text{IoU}} = 0.75$ for the distinct α parameters of the joint loss, contrasting the performance of frozen versus fine-tuned ViT backbone, and (c) tracking of the pixel-wise Mean Squared Error (MSE) trade-offs between localized target accuracy and global scene reconstruction. The legend segregates network training strategies (left) from experimental parameter configurations (right) across all subplots.

4.2 Ablation Study

To identify the optimal configuration for our joint optimization objective, defined in Eq. (2), we conduct a systematic ablation study evaluating three key components: the binarization threshold $\tau_{\text{IoU}} \in [0.1, 0.9]$, the weighting parameter $\alpha \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, and the encoder training strategy (*Frozen* vs. *Fine-tuned* ViT). To reduce computational cost during hyperparameter exploration, all configurations are trained for 5 epochs using a representative subset of 10,000 training images and 1,000 validation samples. Results are shown quantitatively in the graphs in Fig. 3, and qualitatively in Fig. 4.

Binarization threshold τ_{IoU}. We first analyze the sensitivity of the mIoU metric across different thresholds to establish a robust evaluation criterion. As shown in Fig. 3a, all configurations experience a severe performance collapse below $\tau_{\text{IoU}} = 0.5$. This occurs because thresholds lower than our background attenuation baseline ($\sigma = 0.5$) mistakenly incorporate background structures into the foreground mask, leading to widespread activations (see Fig. 4a). Once the threshold exceeds this floor, the metric effectively isolates the true geometric shape of the target objects. Following the evaluation protocol introduced in Sec. 4.1, we adopt $\tau_{\text{IoU}} = 0.75$ for all subsequent analyses, as this threshold yields a reasonable boundary between object masks values (1.0) and attenuation factor ($\sigma = 0.5$).

Weighting parameter α. Using this evaluation criterion, we investigate the balance between target alignment and global scene preservation via α. Optimizing solely for the target mask ($\alpha = 0.0$) leads to localized over-optimization, causing saturated phosphene patterns and distorted boundaries, whereas prioritizing only global reconstruction ($\alpha = 1.0$) preserves background context but



(a) Binarization threshold (τ_{IoU}). (b) Joint loss weight sensitivity (α).

Fig. 4: Qualitative evaluation of the spatial optimization framework. (a) Analysis of background flooding below the attenuation baseline ($\sigma = 0.5$). (b) Structural sensitivity analysis showing the trade-off between mask optimization ($\alpha = 0.00$) and global reconstruction ($\alpha = 1.00$) across representative test cases.

causes object contours to fade. This can be observed in Fig. 4b, where $\alpha = 0.0$ yields highly distorted regions, and $\alpha = 1.0$ prioritizes global reconstruction so much that some objects vanish. Therefore, we look for intermediate values of α to establish a well-behaved trade-off between both objectives. Specifically, we found that $\alpha = 0.8$ strikes the optimal balance by avoiding both saturation artifacts and contour degradation. This selection is quantitatively validated in Fig. 3b, where $\alpha = 0.8$ explicitly achieves the absolute peak mIoU under the selected $\tau_{\text{IoU}} = 0.75$ threshold. This performance correlates directly with the qualitative results in Fig. 4b, which illustrate that $\alpha = 0.8$ yields well-defined object boundaries and well-preserved structures.

Freeze/Fine-tune ViT. Lastly, we evaluate the encoder optimization strategy by comparing a partially frozen backbone (*Frozen ViT*) against full end-to-end training (*Fine-tuned ViT*). As illustrated in Fig. 3b, while both strategies perform similarly across lower weighting values, the *Fine-tuned ViT* configuration achieves the absolute peak mIoU of the study at our selected setup ($\alpha = 0.8$). This demonstrates that adapting the encoder enables the latent representations to better align with the downstream phosphene rendering process when using the optimal balance parameter. This advantage is further supported by the underlying pixel-wise reconstruction trade-off tracked in Fig. 3c, where the *Fine-tuned ViT* baseline shows a more efficient trade-off curve, offering a better balance between mask and scene reconstruction errors. These analyses consistently identify the combination of $\alpha = 0.8$ with the *Fine-tuned ViT* as the optimal configuration of the proposed framework, which we use for subsequent comparison experiments as *Our* approach.

4.3 Main Results and Baseline Comparison

The empirical evaluation reveals a clear trade-off between global reconstruction fidelity and target structure preservation. As shown in Tab. 1, the *Global Base-*

line achieves the lowest global error (MSE_{gl}), indicating a strong bias toward background-dominated, pixel-wise reconstruction. However, this comes at the expense of degraded target reconstruction, as evidenced by the highest mask error (MSE_{mask}) and lowest geometric alignment ($\text{mIoU}_{(0.75)}$), confirming that specialized highlighting methods positively impact target-specific results. Both the *Heuristic Baseline* and our target-conditioned fine-tuning improve upon these metrics. Specifically, our approach improves MSE_{mask} by over 64% and increases the $\text{mIoU}_{(0.75)}$ from 0.5328 to 0.6423 relative to the *Global Baseline*, representing a substantial margin of +0.1095 (+10.95%). Comparatively, the *Heuristic* baseline achieves a more moderate improvement over the *Global Baseline* (57.97% reduction in MSE_{mask} and +5.35% in $\text{mIoU}_{(0.75)}$), supporting the conclusion that the region-weighted objective better preserves target-mask structure under the simulated phosphene model. Interestingly, our approach not only yields a lower target reconstruction error but also maintains a global reconstruction error (MSE_{gl}) lower than the *Heuristic* approach, demonstrating that our learned method is also superior at preserving the whole scene context.

Furthermore, this improvement extends to semantic-level alignment. Our method achieves the highest CLIP cosine similarity (0.6428), outperforming both the global (0.6262) and heuristic (0.6398) configurations. This suggests that improved local structural fidelity may also improve embedding-level semantic consistency, although this metric should not be interpreted as direct evidence of human recognizability.

To complement these quantitative results, Fig. 5 presents qualitative comparisons on representative, challenging kitchen scenarios, including small or distant objects, complex shapes, and multiple concurrent targets. We first compare the *Global Baseline* using the original input image (I) versus the target-conditioned input (I_T), where the latter is blended with masks following Eq. (1). The I_T preprocessing improves target visibility by attenuating irrelevant background regions, even without region-weighted fine-tuning. However, this input modification alone does not prevent small objects from vanishing in the phosphene representation, as shown in Rows 1 and 3 of Fig. 5.

When using the same I_T input, we compare the two baselines with our region-weighted training method. While the *Heuristic* approach recovers some vanished small or distant objects, it occasionally over-dilates them, resulting in a diffuse response. In contrast, our method produces a compact and well-localized representation. Notably, our framework is the only one capable of recovering complex

Table 1: Quantitative comparison of phosphene generation performance on the test set. Bold indicates the best performance.

Method	$\text{MSE}_{gl} \downarrow$	$\text{MSE}_{mask} \downarrow$	$\text{mIoU}_{(0.75)} \uparrow$	CLIP $\uparrow$
Global Baseline	**0.0107**	0.1380	0.5328	0.6262
Heuristic Baseline	0.0133	0.0580	0.5863	0.6398
Ours	0.0126	**0.0488**	**0.6423**	**0.6428**

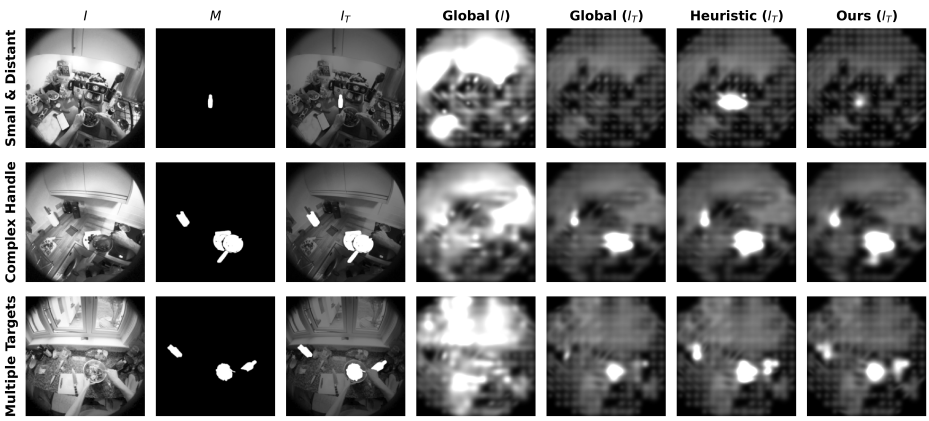


Fig. 5: Qualitative comparison of phosphene generation strategies under different input and optimization configurations. From left to right: Original Image (I), Target Mask (M), Preprocessed Input (I_T), Global Baseline model evaluated on I, Global Baseline model evaluated on I_T, Heuristic Baseline on I_T, and Our region-weighted optimization on I_T. Comparing Columns 4 and 5 demonstrates how the preprocessing module successfully redirects network focus toward user-specified regions. Across the rows, our method consistently achieves superior structural fidelity, preserving fine geometric details (such as the pan handle in Row 2) and correctly isolating distinct targets (Row 3) without the over-dilation artifacts present in the heuristic approach.

geometries (e.g., the pan's handle in Row 2). Overall, our approach preserves both object count and geometric structure, producing a more balanced, spatially consistent representation. This demonstrates that a data-driven approach is more versatile for the diverse objects encountered in daily life compared to heuristic methods, which are difficult to parametrize for robust performance across all plausible object shapes under severe compression.

4.4 Complete Vision-Language Spatial Grounding pipeline

In this section we qualitatively evaluate the behavior of the complete pipeline that involves the complete Vision-Language Spatial Grounding module, including the VLM component, as illustrated in Fig. 6. In particular, we use Qwen3.5 20B model [39] as VLM. First, intermediate results consist of the list of objects identified by the VLM and their corresponding confidence (Fig. 6b). Then, a set of masks related with these particular objects are extracted with SAM 3 and composited via alpha-blending with the original image to get the target image as described in Sec. 3.2 (Fig. 6c). Finally, the trained autoencoder processes the target image with the patient parameters to get the output percept (Fig. 6d). These qualitative examples show that the complete pipeline can identify objects related to the user instruction, generate corresponding masks, and produce simulated phosphene outputs adapted to a particular patient-parameter setting.

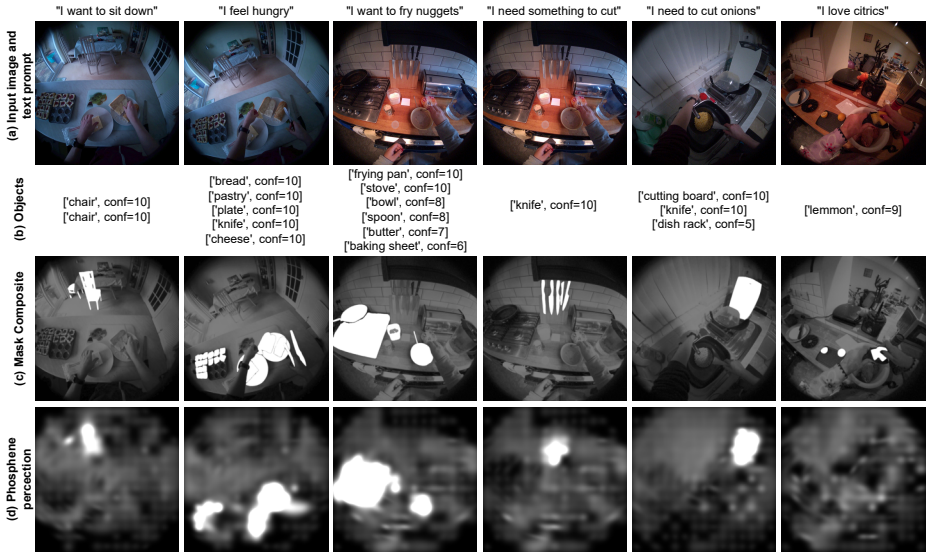


Fig. 6: Qualitative results of the complete pipeline. (a) Egocentric input image captured by the camera and a text prompt encoding the user instructions. (b) Related objects reasoned by the VLM. (c) Intermediate composite including the masked objects. (d) Output simulated percepts.

5 Conclusions

In this work, we presented a framework for phosphene stimuli encoding considering language-driven segmentation and biophysical patient adapted models of prostheses. Our proposal allows us to bridge the gap between an open-vocabulary user instruction and an augmented representation of a set of objects to interact with. For that we combine a Vision-Language Model, a Segmentation Model and a custom stimulus encoder refined to emphasize the mask of the detected objects of interest in the target phosphene representation. Our results show that region-weighted fine-tuning improves target-mask alignment over a general stimulus encoder (+10.95% mIoU at the selected operating threshold), while retaining attenuated background context. We also present a set of qualitative examples of the complete pipeline. Because the present evaluation is simulation-based, future work should test whether these improvements translate into better localization, recognition, or navigation performance in human-subject SPV experiments and, ultimately, implanted users.

Acknowledgements

This work was supported by projects PID2024-158322OB-I00, and grant PRE2022-104910 (MCIN/AEI/10.13039/501100011033/ FEDER, UE), and project JIUZ2024-IyA-07.

References

1. Alqahtani, H.B., Votruba, M., Regini, J.: Do retinal implants provide long-term efficacy, safety and improve quality of life? a systematic review. Therapeutic Advances in Ophthalmology **17**, 25158414251385884 (2025) 2
2. Ayton, L.N., Barnes, N., Dagnelie, G., Fujikado, T., Goetz, G., Hornig, R., Jones, B.W., Muqit, M.M., Rathbun, D.L., Stingl, K., Weiland, J.D., Petoe, M.A.: An update on retinal prostheses. Clinical Neurophysiology **131**(6), 1383–1398 (2020) 1
3. Beyeler, M.: Biophysical model of axonal stimulation in epiretinal visual prostheses. In: 9th International IEEE/EMBS Conference on Neural Engineering (NER). pp. 348–351 (2019) 4
4. Beyeler, M., Boynton, G.M., Fine, I., Rokem, A.: pulse2percept: A Python-based simulation framework for bionic vision. In: 16th Python in Science Conference (SciPy). pp. 81–88 (2017) 2, 4
5. Beyeler, M., Nanduri, D., Weiland, J.D., Rokem, A., Boynton, G.M., Fine, I.: A model of ganglion axon pathways accounts for percepts elicited by retinal implants. Scientific reports **9**(1), 9199 (2019) 2, 4
6. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris Coll-Vinent, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: SAM 3: Segment anything with concepts. In: International Conference on Learning Representations (ICLR). pp. 138846–138923 (2026) 5, 6, 8
7. Chen, S.C., Suaning, G.J., Morley, J.W., Lovell, N.H.: Simulating prosthetic vision: I. Visual models of phosphenes. Vision research **49**(12), 1493–1506 (2009) 4
8. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025) 5
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: IEEE Computer Vision and Pattern Recognition Conference (CVPR). pp. 248–255 (2009) 9
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021) 9
11. Granley, J., Beyeler, M.: A computational model of phosphene appearance for epiretinal prostheses. In: 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 4477–4481 (2021) 4
12. Granley, J., Fauvel, T., Chalk, M., Beyeler, M.: Human-in-the-loop optimization for deep stimulus encoding in visual prostheses. Advances in neural information processing systems **36**, 79376–79398 (2023) 2, 3, 4, 6, 7, 8
13. Granley, J., Relic, L., Beyeler, M.: Hybrid neural autoencoders for stimulus encoding in visual and other sensory neuroprostheses. Advances in Neural Information Processing Systems **35**, 22671–22685 (2022) 2, 4
14. Guo, F., An, J., Duan, J.: Moving object detection in high dynamic scene for visual prosthesis. In: 8th Information Technology and Mechatronics Engineering Conference (ITOEC). vol. 8, pp. 1540–1544. IEEE (2025) 4
15. Hou, Y., Pullela, L., Su, J., Aluru, S., Sista, S., Lu, X., Beyeler, M.: Predicting the temporal dynamics of prosthetic vision. In: 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2024) 4

16. Huang, Y., Xu, J., Pei, B., He, Y., Chen, G., Zhang, M., Yang, L., Nie, Z., Liu, J., Fan, G., et al.: An egocentric vision-language model based portable real-time smart assistant. arXiv preprint arXiv:2503.04250 (2025) 5
17. Huh, M., Xue, Z., Das, U., Ashutosh, K., Grauman, K., Pavel, A.: Vid2Coach: Transforming how-to videos into task assistants. In: 38th Annual ACM Symposium on User Interface Software and Technology (2025) 5
18. Humayun, M.S., Dorn, J.D., da Cruz, L., Dagnelie, G., Sahel, J.A., Stanga, P.E., Cideciyan, A.V., Duncan, J.L., Eliott, D., Filley, E., Ho, A.C., Santos, A., Safran, A.B., Arditi, A., Del Priore, L.V., Greenberg, R.J.: Interim results from the international trial of Second Sight's visual prosthesis. Ophthalmology **119**(4), 779–788 (2012) 2
19. Khalifa, N., Selim, S., Al-Atabany, W.: Evaluating the efficacy of smart saliency detection system for visual prosthesis users: An experimental comparison across various visual prosthesis implants. Artificial Organs (2026) 4
20. Li, F.M., Ng, K., Zhu, B., Carrington, P.: OSCAR: object status and contextual awareness for recipes to support non-visual cooking. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (2025) 5
21. Liang, J., Li, H., Chai, X., Gao, Q., Zhou, M., Guo, T., Chen, Y., Di, L.: An audiovisual cognitive optimization strategy guided by salient object ranking for intelligent visual prothesis systems. Journal of Neural Engineering **21**(6), 066021 (2024) 5
22. Nejad, A., Küçükoğlu, B., de Ruyter van Steveninck, J., Bedrossian, S., Heutink, J., de Haan, G.A., Cornelissen, F.W., van Gerven, M.: Point-SPV: End-to-end enhancement of object recognition in simulated prosthetic vision using synthetic viewing points. Frontiers in human neuroscience **19**, 1549698 (2025) 4
23. Palanker, D., Le Mer, Y., Mohand-Said, S., Muqit, M., Sahel, J.A.: Photovoltaic restoration of central vision in atrophic age-related macular degeneration. Ophthalmology **127**(8), 1097–1104 (2020) 2
24. Palanker, D., Weiland, J.D., Rosin, B., Sahel, J.A.: Restoration of sight with electronic retinal prostheses. Annual review of vision science **11**(1), 125–147 (2025) 2
25. Papadopoulos, E., Güçlütürk, Y.: Interactive semantic segmentation for phosphene vision neuroprosthetics. arXiv preprint arXiv:2509.19957 (2025) 4
26. Perez-Yus, A., Santos-Villafranca, M., Tomas-Barba, J., Bermudez-Cameo, J., Montano-Olivan, L., Lopez-Nicolas, G., Guerrero, J.J.: RASPV: A robotics framework for augmented simulated prosthetic vision. IEEE Access **12**, 15251–15267 (2024) 4
27. Perrett, T., Darkhalil, A., Sinha, S., Emara, O., Pollard, S., Parida, K.K., Liu, K., Gatti, P., Bansal, S., Flanagan, K., et al.: HD-EPIC: A highly-detailed egocentric video dataset. In: IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR). pp. 23901–23913 (2025) 3, 8
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763. PMLR (2021) 9
29. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: SAM 2: Segment anything in images and videos. In: International Conference on Learning Representations (ICLR). pp. 28085–28128 (2025) 4

30. Ren, T., Jiang, Q., Liu, S., Zeng, Z., Liu, W., Gao, H., Huang, H., Ma, Z., Jiang, X., Chen, Y., et al.: Grounding DINO 1.5: Advance the "edge" of open-set object detection. arXiv preprint arXiv:2405.10300 (2024) 4
31. Sanchez-Garcia, M., Martinez-Cantin, R., Guerrero, J.J.: Semantic and structural image segmentation for prosthetic vision. Plos one **15**(1), e0227677 (2020) 4
32. van Steveninck, J.d.R., Güçlü, U., van Wezel, R., van Gerven, M.: End-to-end optimization of prosthetic vision. Journal of Vision **22**(2), 20–20 (2022) 2, 4
33. Tomas-Barba, J., Perez-Yus, A., Martinez-Cantin, R., Guerrero, J.J., Bermudez-Cameo, J.: Adaptive vision transformer for enhanced perception in visual prostheses. In: 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2025) 4, 7, 9
34. Wang, A., Liu, L., Chen, H., Lin, Z., Han, J., Ding, G.: YOLOE: Real-time seeing anything. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 24591–24602 (2025) 4
35. Wang, Y., Wang, Z., Hahn, N., Cheng, L., Liu, S.C.: Investigation of audio controlled edge semantic segmentation system for visually impaired people. In: 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2025) 4
36. White, J., Ruiz-Serra, J., Petrie, S., Kameneva, T., McCarthy, C.: Self-attention based vision processing for prosthetic vision. In: 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (2023) 4
37. World Health Organization: Blindness and vision impairment. World Health Organization Fact Sheets (2026), https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment 1
38. Wu, Y., Karetić, I., Stegmaier, J., Walter, P., Merhof, D.: A deep learning-based in silico framework for optimization on retinal prosthetic stimulation. In: 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2023) 4
39. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L.C., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S.Q., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y.C., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 5, 13
40. Yuan, H., Li, X., Zhang, T., Sun, Y., Huang, Z., Xu, S., Ji, S., Tong, Y., Qi, L., Feng, J., et al.: Sa2VA: Marrying SAM2 with LLaVA for dense grounded understanding of images and videos. arXiv preprint arXiv:2501.04001 (2025) 5
41. Zhang, Z., Ma, Y., Zhang, E., Bai, X.: PSALM: Pixelwise segmentation with large multi-modal model. In: European Conference on Computer Vision (ECCV). pp. 74–91. Springer (2024) 5