Not Your Stereo-Typical Estimator: Combining Vision and Language for Volume Perception

Supplementary Material

7. GPT-5 Experiment Prompts

For our experiments with GPT-5, the prompts are designed to ensure clarity in how the model is instructed and what data it receives. We used two different prompting structures depending on whether monocular (single image) or stereo (two images) were used for volume estimation.

Single Image Prompt The single-image prompt asks the model to simply estimate the volume and ensure a numerical output. Listing 1 showcases a "context-free" version, whereas Listing 2 is the "context-aware" alternative.

Listing 1. Single Image Volume Estimation Prompt (no context)

```
"Answer with ONLY a single floating-point
    number (milliliters). No units, no extra
     text.
Estimate the object's volume in milliliters
     from the image.
Return ONLY a single floating-point number
    (milliliters), no units, no words, no
    punctuation, no JSON, no code fences."
```

Listing 2. Single Image Volume Estimation Prompt (with context)

```
"Answer with ONLY a single floating-point
    number (milliliters). No units, no extra
     text.
Given this is an image of {context_text},
    estimate its volume in milliliters.
Return ONLY a single floating-point number
    (milliliters), no units, no words, no
    punctuation, no JSON, no code fences."
```

Stereo Image Prompt For the two-view task, a more detailed and structured prompt is passed to the model to rely on stereo cues and return a single numeric estimate. The prompt, in detail, is found in Listing 3

Listing 3. Stereo Image Volume Estimation Prompt

```
"You are given TWO images of the SAME
    object, captured from different
    viewpoints.
Use both images jointly (stereo cues,
    parallax, shape consistency) to estimate
     the object's volume in milliliters.
Assume similar scale and camera distance;
    modest viewpoint change is present.
RESPONSE FORMAT (STRICT JSON, one object,
    no code fences, no extra text):
{\n  \"volume_ml\": <float>,\n  \"
    explanation\": \"<2-4 concise sentences
    on the visual cues you used>\"\n}
Rules:
- Return ONLY the JSON object above (no
    markdown, no reasoning sections, no
    additional keys).
- \"volume_ml\" MUST be a single floating-
    point number (no units, no commas).
Return ONLY the final JSON object; do not
    include chain-of-thought or extra text."
```

8. Generalizability

To evaluate our model's ability to generalize to unseen object categories, we performed experiments on the MetaFood3D dataset [3]. We used a random train-test split, resulting in 415 training and 104 testing samples, which ensures that the test set contains categories absent from training. Given the dataset's limited size, strong generalization performance is not expected. As shown in Table 8, we compare our method against the trainable "RGB Only" baseline under these challenging conditions.

Method	MAE (mL)	MAPE (%)
Baseline	150.02	553.96
RGB Only	221.07	299.52
Ours	**40.50**	**22.38**

Table 8. **Random Split Generalization.** The random split into training and testing splits ensures that there are categories and items that the model has never "seen" and yet the performance is significantly better than the RGB estimation method which was also trained with the same training and testing split.

To further assess the model's zero-shot generalization capabilities, we trained our method exclusively on the OmniObject3D dataset [43] and evaluated it directly on the MetaFood3D dataset [3]. This cross-dataset evaluation introduces a significant domain shift, as the model must contend with numerous out-of-distribution (OOD) images and categories. As presented in Table 9, while performance is understandably limited by this domain gap, our method still surpasses the "RGB Only" baseline. This result indicates a greater understanding of object sizes and context cues,

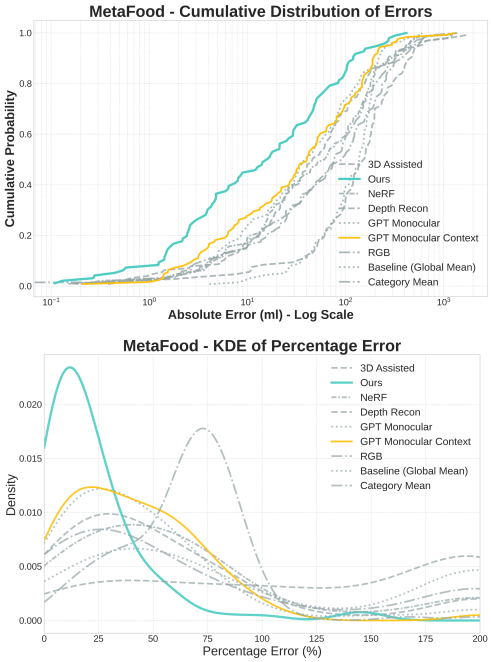

Figure 5. **Error Distribution of Volume Estimation Methods.** We highlight our method (blue) against the second-best performing method, GPT-5 with context (orange), with regards to MAPE. **(Top)** The Cumulative Distribution Function (CDF) of absolute errors on a log scale where a shift or curve to the upper left indicates better performance. **(Bottom)** The Kernel Density Estimation (KDE) of percentage errors where a tall peak near zero with significant delay indicates better performance.

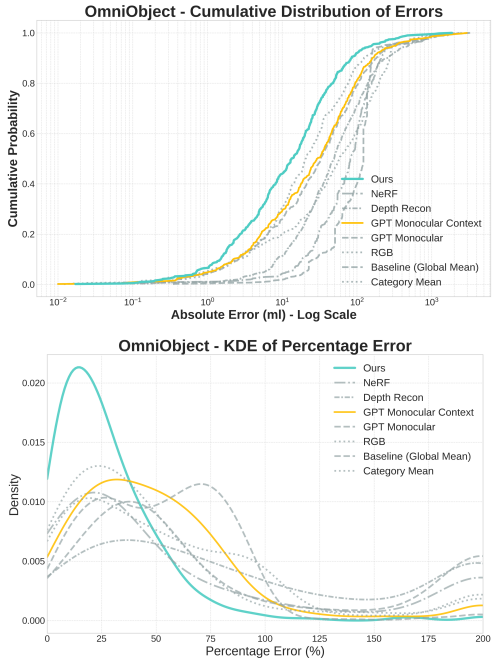

Figure 6. **Error Distribution of Volume Estimation Methods.** We highlight our method (orange) against the second-best performing method, GPT-5 with context (blue), with regards to MAPE. **(Top)** The Cumulative Distribution Function (CDF) of absolute errors on a log scale where a shift or curve to the upper left indicates better performance. **(Bottom)** The Kernel Density Estimation (KDE) of percentage errors where a tall peak near zero with significant delay indicates better performance.

though achieving true zero-shot performance will necessitate training on larger and more varied datasets.

Method	MAE (mL)	MAPE (%)
Baseline	165.75	836.50
RGB Only	174.57	212.96
GPT-5	92.38	198.83
Ours	**96.08**	**178.86**

Table 9. **Cross Dataset Validation.** The model tested on "out-of-distribution" images still outperforms the baseline and the RGB Only method which was trained and tested on the same data as our method.

9. Additional Error Visualizations

In Figure 6, we perform the same error analysis for OmniObject as we did in Figure 5 (From main paper). We similarly observe that our method outperforms all other monocular methods, having more effective absolute error and absolute percentage error distributions.

Figures 7 and 8 plot the estimated volumes against the ground truth volumes for MetaFood and OmniObject, respectively. Similar to our error analysis figures, we also compare our method against the second-best performing method in GPT-5 with context. A "line of best fit" is overlaid these scatter plots to help visualize the correlation between the estimations and a "perfect estimation" line. We observe how, for both datasets, our method's line of best fit is more closely aligned to the perfect estimation line.

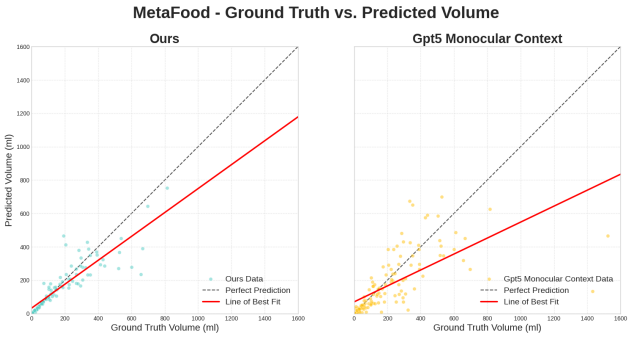

Figure 7. Predicted volumes plotted against ground truth volumes for MetaFood. We plot our method against the second-best performing monocular method, GPT-5 with context, and overlay a line of best fit to help visualize correlation with a "perfect estimation" line.

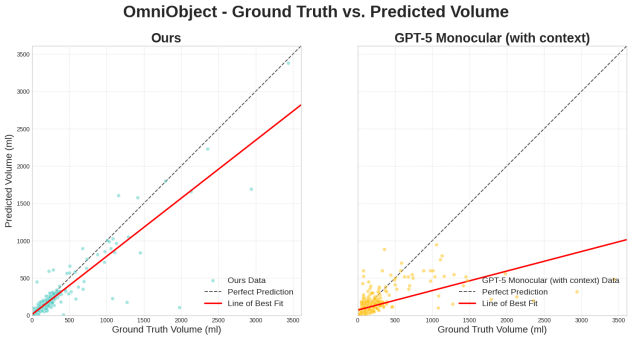

Figure 8. Predicted volumes plotted against ground truth volumes for OmniObject. We plot our method against the second-best performing monocular method, GPT-5 with context, and overlay a line of best fit to help visualize correlation with a "perfect estimation" line.