# WEAK-TO-STRONG ON-POLICY DISTILLATION VIA PERCEPTION GAP SUPERVISION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

On-policy distillation (OPD) supervises a student on its own rollouts using token-level targets from a teacher. Existing OPD methods typically rely on a teacher model that is stronger than the student. Training such a teacher can be expensive or even impractical. We study weak-to-strong OPD for vision-language models, where the teacher is weaker than the student, typically being smaller or less extensively trained. Directly distilling the weak teacher's predictions can degrade student performance. We introduce Weak-to-Strong Distillation via Perception Gap (W2S-PG), which constructs supervision from the discrepancy between the same weak teacher's predictions under two answer-preserving visual perturbations. While teacher-derived supervision is effective, the student becomes stronger during training, and its own predictions may eventually provide more reliable supervision than the teacher-derived signal. To select the more reliable source of supervision, we further propose Adaptive Supervision Signal Selection (AS3), a token-level mechanism that adaptively chooses between teacher-derived perception-gap supervision and self-training according to model confidence. Across six perception benchmarks and four teacher–student pairs, W2S-PG consistently improves the student and outperforms other OPD variants such as vanilla OPD and previous weak-to-strong distillation approaches. When combined with W2S-PG, AS3 provides further gains of 0.3–2.4 points across all benchmarks. Ablations of augmentation and discrepancy choices demonstrate the robustness of W2S-PG, and evaluations on seven additional benchmarks show that its benefits extend beyond perception to visual reasoning and other tasks.

## 1 INTRODUCTION

On-policy distillation (OPD) (Agarwal et al., 2024) has become an important approach for the post-training of foundation models (Yang et al., 2025; Xu et al., 2026). OPD follows the standard distillation setting and involves two models: a student and a teacher. In the OPD setting, the teacher provides the student with token-level, fine-grained supervision on trajectories sampled from the student's own policy. OPD provides token-level supervision signals for the student, and on-policy training helps mitigate exposure bias.

In vanilla OPD, a stronger teacher model is usually required. Typically, teacher models are larger than student models (Yang et al., 2025) (e.g., a 32B teacher for a 4B student) or are specialized for a particular domain through reinforcement learning (Xu et al., 2026). However, training larger models or expert models across multiple domains requires substantial computation. To address these limitations and eliminate the need for more powerful teacher models, weak-to-strong distillation (Burns et al., 2024; Feng et al., 2026; Zhu et al., 2025) has emerged as a promising paradigm, enabling student models to learn from weaker teachers when stronger teachers are unavailable.

The main insight behind recent weak-to-strong distillation approaches is to avoid directly distilling model outputs and instead leverage the discrepancy between different models, such as models before and after RL training, as the supervision signal. In this work, we explore a different direction for weak-to-strong distillation by leveraging the *perception gap* (W2S-PG) in vision-language models (VLMs) (Liu et al., 2023). Our approach is motivated by the observation that VLMs can produce different output distributions for different visual views of the same input. We exploit the discrepancy between the predictions of a *single* weak teacher under different visual inputs as a proxy supervision



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

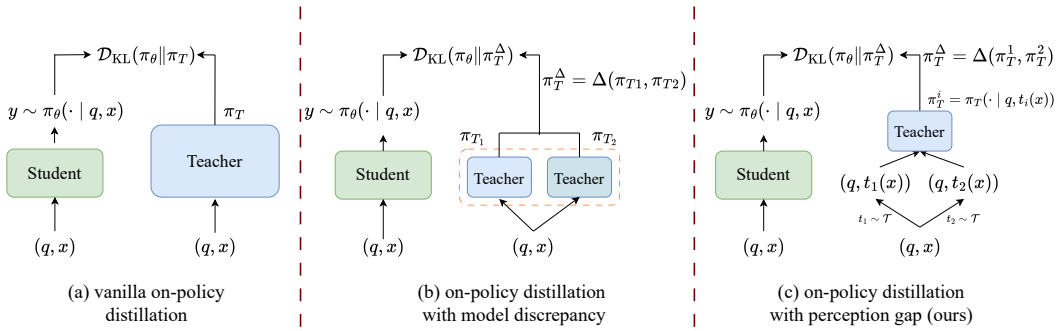


Figure 1: **Conceptual comparison of three on-policy distillation mechanisms.** The three distillation mechanisms differ in *where their supervision signals come from*. Given an on-policy sample $y \sim \pi_\theta(\cdot \mid q, x)$ for a text query $q$ and an image $x$: **(a)** vanilla on-policy distillation learns from a more powerful teacher model $\pi_T$. **(b)** Previous weak-to-strong distillation approaches (Yu et al., 2026; Zhu et al., 2025) use output differences between two different models, $\pi_{T_1}$ and $\pi_{T_2}$, as supervision signals. **(c)** Our W2S-PG uses the output differences of a single weak model $\pi_T$ under different visual conditions, $t_1(x)$ and $t_2(x)$, as the supervision signal.

signal, thereby eliminating the need for a stronger teacher model or multiple weak teacher models. A conceptual comparison of different OPD variants is shown in Figure 1.

While W2S-PG provides effective supervision for on-policy distillation, another concern arises as the student becomes stronger during training: is supervision derived from the perception gap of a weak teacher always more reliable than the student's own predictions? To obtain more reliable supervision signals, we further propose adaptive supervision signal selection (AS3), a token-level mechanism that adaptively selects supervision for each token from either the teacher or the student based on their confidence.

To validate the effectiveness of W2S-PG, we first compare it with other distillation approaches on multiple perception benchmarks. The results show that W2S-PG outperforms various OPD variants with the weak teacher. Furthermore, W2S-PG consistently improves performance across different student–teacher pairs and model families. We conduct ablations on various design choices of W2S-PG and AS3 to assess the contribution of each component. Finally, we demonstrate the generalizability of W2S-PG beyond perception benchmarks by evaluating it on other domains, including visual reasoning, general visual question answering and chart understanding tasks.

To summarize, our main contributions are as follows:

- We propose W2S-PG, a weak-to-strong distillation approach that leverages the output discrepancy of a single weak teacher model under different visual inputs as the supervision signal. W2S-PG eliminates the need for stronger teacher models in on-policy distillation for vision-language models.
- We propose AS3, which adaptively selects reliable supervision signals from the teacher or the student at the token level based on their confidence.
- Extensive experiments across benchmarks spanning multiple domains, particularly perception, and across four student–teacher pairs demonstrate the effectiveness and generalizability of W2S-PG and AS3.

## 2 Method

In this section, we introduce the formulation of the proposed W2S-PG and AS3. An overview of W2S-PG with AS3 is presented in Algorithm 1.

### 2.1 Preliminaries: On-Policy Distillation

Given a text query $q$ and an image $x$ from the training set $\mathcal{S}$, the likelihood of a generated sequence $y$ under a vision-language model $\pi_\theta$ is defined as $\pi_\theta(y \mid q, x) = \prod_{j=1}^{|y|} \pi_\theta(y_j \mid q, x, y_{<j})$ where

**Algorithm 1** Weak-to-Strong Distillation with Perception Gap (W2S-PG) and Adaptive Supervision Signal Selection (AS3)

**Require:** Training sample $(q, x)$, weak teacher $\pi_T$, student $\pi_\theta$, augmentation set $\mathcal{T}$, number of training steps $K$

**for** $k = 1, \ldots, K$ **do** ▷ Training steps
  Sample transformations $t_1 \sim \mathcal{T}$ and $t_2 \sim \mathcal{T}$, and obtain two augmented views $x_1 = t_1(x), x_2 = t_2(x)$
  Generate a student rollout $y \sim \pi_\theta(\cdot \mid q, x)$
  Forward the two augmented views through $\pi_T$ to obtain $y_j^T$ using Eq. (2)
  **for** each token position $j$ in $y$ **do**
    Compute confidence scores $c_j = \max y_j$ and $c_j^T = \max y_j^T$
    **if** $c_j \geq c_j^T$ **then**
      Set $\alpha_j = 1$
    **else**
      Sample $\gamma \sim U(0, 1)$ and compute $\alpha_j$ with Eq. (4)
    **end if**
  **end for**
  Optimize $\pi_\theta$ using Eq. (7)
**end for**
**return** Trained student model $\pi_\theta$

$y_{<j}$ denotes the sequence of tokens preceding position $j$, and $y_j \in \Delta^{V-1}$ denotes the predictive distribution over the vocabulary at position $j$, conditioned on $y_{<j}$. Here, $V$ is the vocabulary size, and $\Delta^{V-1}$ denotes the corresponding probability simplex. On-policy distillation (OPD) (Agarwal et al., 2024; Gu et al., 2024) supervises the student on its own generated trajectories as

$$\mathcal{L}_{\text{OPD}}(\theta) = \mathbb{E}_{(q,x)\sim\mathcal{S},\, y\sim\pi_\theta(\cdot|q,x)} \left[\mathcal{D}_{\text{KL}}\left(\pi_\theta(y \mid q, x) \,\|\pi_T(y \mid q, x)\right)\right], \tag{1}$$

where $\pi_T$ denotes the frozen teacher model, usually larger (Yang et al., 2025) or specialized for a particular domain (Xu et al., 2026). $\mathcal{D}_{\text{KL}}$ denotes Kullback–Leibler (KL) divergence.

### 2.2 Weak-to-Strong Distillation via Perception Gap

Training models with vanilla OPD, as formulated in Eq. (1), requires a stronger model as the teacher $\pi_T$. However, training larger or specialized models is usually expensive. We propose **weak-to-strong distillation via perception gap** (W2S-PG) to eliminate the need for stronger teacher models. Inspired by the observation that VLMs generate different outputs under different visual inputs, we use the discrepancy between the predictions of a *single* weak teacher under different visual inputs as the supervision signal.

Given two different transformations from the augmentation family $\mathcal{T}$, $t_1 \sim \mathcal{T}$ and $t_2 \sim \mathcal{T}$, we generate two augmented views of the image $x$, $x_1 = t_1(x)$ and $x_2 = t_2(x)$. Given a rollout $y \sim \pi_\theta(\cdot \mid q, x)$ of the student $\pi_\theta$, rather than directly learning from the weak teacher model $\pi_T$, we use the discrepancy between outputs under different visual inputs as the supervision signal. We denote the output difference of the weak teacher under different visual inputs as $y^T = \pi_T^{w2s}(y \mid q, x_1, x_2)$. For the $j$-th token, the supervision signal is obtained as

$$y_j^T = \pi_T^{w2s}(y \mid q, x_1, x_2) = \text{softmax}\left((\pi_T(\cdot \mid q, x_1, y_{<j}) - \pi_T(\cdot \mid q, x_2, y_{<j}))^2\right). \tag{2}$$

With the supervision signal $\pi_T^{w2s}$, we reformulate OPD in Eq. (1) as

$$\mathcal{L}_{\text{W2S-PG}}(\theta) = \mathbb{E}_{(q,x)\sim\mathcal{S},\, y\sim\pi_\theta(\cdot|q,x)} \left[\mathcal{D}_{\text{KL}}\left(\pi_\theta(y \mid q, x) \,\|\pi_T^{w2s}(y \mid q, x_1, x_2)\right)\right]. \tag{3}$$

The key difference between Eq. (1) and Eq. (3) is the supervision signal. For vanilla OPD (Eq. (1)), we directly use outputs $\pi_T(y \mid q, x)$ from the teacher model as supervision because the teacher is usually more powerful. However, in the weak-to-strong distillation scenario, we use the discrepancy $\pi_T^{w2s}(y \mid q, x_1, x_2)$ of the *weaker* teacher model across different visual inputs in Eq. (3). This signal serves as a "proxy" supervision signal when a stronger teacher model is unavailable.

### 2.3 Selecting Reliable Supervision Adaptively

A key challenge in weak-to-strong distillation is how a weak foundation model can effectively supervise a stronger counterpart. Although discrepancies across different visual inputs provide effective supervision, the student becomes stronger during training, raising another question:

*should the student trust its own predictions or the supervision from the weaker teacher model?*

A short answer is **both**. We empirically find that relying exclusively on either the student's own predictions or the weak-to-strong supervision in Eq. (3) is suboptimal (see Figure 3). Therefore, we propose **adaptive supervision signal selection** (AS3) to adaptively select more reliable supervision signals from the student and the teacher. We detail AS3 as follows.

For the $j$-th token in the generated sequence $y$, given $y_j \in \Delta^{V-1}$ and $y_j^T \in \Delta^{V-1}$, we use their confidence scores, $c_j = \max y_j$ and $c_j^T = \max y_j^T$ as proxies for their reliability. If the student's confidence is higher, we use its own prediction rather than the supervision from the teacher model. Otherwise, we probabilistically select between student and teacher supervision according to their relative confidence. We introduce a token-level balance factor $\alpha_j$ to control the selection of supervision signals. Specifically, if $c_j \geq c_j^T$, we set $\alpha_j = 1$, i.e., we use the signal from the student itself. Otherwise, we sample $\gamma$ from the uniform distribution $U(0,1)$. If $\gamma > c_j/c_j^T$, we set $\alpha_j = 0$ and use $y_j^T$ as the supervision signal; otherwise, we use $y_j$. Formally, $\alpha_j$ is defined as

$$\alpha_j = \begin{cases} 1, & c_j \geq c_j^T, \\ \mathbb{I}\left(\gamma \leq c_j/c_j^T\right), & c_j < c_j^T, \quad \gamma \sim U(0,1), \end{cases} \tag{4}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function.

With the token-level balance factor $\alpha_j$, if the supervision signal at token position $j$ is obtained from the student, we directly use the sequence generated by the student using the following loss:

$$\mathcal{L}_s(j) = -\log \pi_\theta(y_j \mid q, x, y_{<j}). \tag{5}$$

This is equivalent to self-training on the model's own generated sequence. If the supervision signal is obtained from the teacher model $\pi_T$, the loss function is formulated as

$$\mathcal{L}_t(j) = \mathcal{D}_{\text{KL}}\left(\pi_\theta\left(y_j \mid q, x, y_{<j}\right) \| \pi_T^{w2s}\left(y_j \mid q, x_1, x_2, y_{<j}\right)\right). \tag{6}$$

This corresponds to the standard W2S-PG objective in Eq. (3) at token position $j$. Combining the two supervision sources, we define the adaptive hybrid loss as

$$\mathcal{L}_{\text{W2S-PG}}^{\text{AS3}} = \frac{1}{|y|}\sum_{j=1}^{|y|} \alpha_j \mathcal{L}_s(j) + (1-\alpha_j)\mathcal{L}_t(j). \tag{7}$$

Note that W2S-PG (Eq. (3)) is a special case of Eq. (7) when we set $\alpha_j = 0$ for all token positions $j$. We discuss different choices of $\alpha_j$ in Sec. 3.3.

### 2.4 Design of Augmentation

We consider the following augmentation family $\mathcal{T}$ in our experiments.

- **Downscaling.** We randomly downscale the input image by a factor of 2 or 4.
- **Vision token dropout.** We randomly drop vision tokens produced by the vision encoder in the VLM at a rate of 10% or 20%.
- **Identity transformation.** No transformation is applied to the input image.
- **Gaussian noise.** We add Gaussian noise to the image with variance selected from $\{0.08, 0.12, 0.18, 0.26, 0.38\}$.
- **Gaussian blur.** We blur the image using a Gaussian filter that attenuates high-frequency signals.

Table 1: **Performance of different distillation approaches on six perception benchmarks**. We highlight results related to W2S-PG and the **best** results.

| | VStar | HR-4K | HR-8K | MME-RW | ZoomBench | CV-Bench |
|---|---|---|---|---|---|---|
| *Teacher models for reference* | | | | | | |
| Qwen3-VL-2B | 72.8 | 71.8 | 66.0 | 58.1 | 41.3 | 79.0 |
| Qwen3-VL-32B | 88.5 | 82.0 | 79.8 | 68.4 | 49.3 | 86.8 |
| *Qwen3-VL-4B distillation* | | | | | | |
| Qwen3-VL-4B | 80.1 | 78.3 | 72.9 | 63.5 | 40.2 | 85.0 |
| OPD w/ Qwen3-VL-32B | 84.0 | 81.1 | 74.4 | 64.7 | 44.5 | 86.3 |
| OPD-$\Delta$ | 84.5 | 82.0 | 74.6 | 64.4 | 46.1 | 85.8 |
| OPD w/ Qwen3-VL-2B | 78.4 | 76.5 | 72.0 | 63.1 | 40.4 | 81.2 |
| W2S-PG | 92.1 | 82.2 | 80.2 | 67.1 | 58.2 | 88.0 |
| W2S-PG w/ AS3 | **93.6** | **83.4** | **82.6** | **68.5** | **59.3** | **88.4** |
| *Qwen3-VL-8B distillation* | | | | | | |
| Qwen3-VL-8B | 86.4 | 78.9 | 74.6 | 66.0 | 37.9 | 85.4 |
| OPD w/ Qwen3-VL-32B | 87.0 | 81.1 | 77.5 | 68.2 | 43.8 | 86.4 |
| OPD-$\Delta$ | 88.4 | 82.0 | 77.9 | 68.0 | 44.1 | 86.8 |
| OPD w/ Qwen3-VL-2B | 82.4 | 75.1 | 72.9 | 64.6 | 37.4 | 83.6 |
| W2S-PG | 93.2 | 85.0 | 82.8 | 70.3 | 58.5 | 88.7 |
| W2S-PG w/ AS3 | **94.1** | **86.8** | **84.5** | **71.6** | **59.5** | **89.0** |

The main principle behind the augmentation family $\mathcal{T}$ is to construct different views of the same image without changing the answer to a given query $q$. Before training, we perform an offline verification of the augmented training data and filter out samples for which the applied augmentations change the answer. During training, we force the sampled augmentation operations $t_1 \neq t_2$ to induce a sufficient perception gap between the two views.

### 2.5 What Does W2S-PG Learn from the Weak Teacher?

Instead of distilling *what* the weak teacher predicts, W2S-PG distills *how* its predictions change in response to variations in the visual input. This provides a complementary learning signal that reflects the visual evidence underlying the teacher's predictions.

To see this, consider two augmented views of an image $t_1(x)$ and $t_2(x)$. If the perturbation is relatively small and additive, we can approximate $t_i(x) \approx x + \delta_i$, where $\delta_i$ denotes the perturbation induced by the $i$-th augmentation. For a sufficiently small perturbation $\delta_i$, the teacher's prediction can be locally approximated using a first-order Taylor expansion:

$$\pi_T(y \mid q, t_i(x)) \approx \pi_T(y \mid q, x) + \nabla_x \pi_T(y \mid q, x)^\top \delta_i. \tag{8}$$

Consequently, the difference between the teacher's predictions on two augmented views satisfies

$$\pi_T(y \mid q, t_1(x)) - \pi_T(y \mid q, t_2(x)) \approx \nabla_x \pi_T(y \mid q, x)^\top (\delta_1 - \delta_2). \tag{9}$$

This relation provides an intuitive interpretation of the signal distilled by W2S-PG. While the absolute prediction $\pi_T(y \mid q, x)$ primarily conveys the teacher's belief about the answer, the prediction difference captures its *local sensitivity* along a particular direction of visual variation $\delta_1 - \delta_2$. It reveals which visual changes affect the teacher's decision and which leave it largely unchanged.

Across diverse augmentations, these responses expose different projections of the teacher's input sensitivity. Visual variations that strongly affect task-relevant evidence tend to induce larger changes in the teacher's predictions, whereas variations to which the teacher is invariant induce smaller changes. By matching these prediction responses rather than only the predictions themselves, W2S-PG encourages the student to inherit the teacher's dependence on visual evidence—i.e., how visual information is used to support its predictions—without requiring the weak teacher's absolute predictions to be treated as reliable targets.

Table 2: **Performance of W2S-PG on various teacher-student pairs**. **+x** indicates improvement compared with the corresponding student model.

| Teacher | Student | VStar | HR-4K | HR-8K | MME-RW | ZoomBench | CV-Bench |
|---|---|---|---|---|---|---|---|
| *Baselines* | | | | | | | |
| – | Qwen3-VL-2B | 72.8 | 71.8 | 66.0 | 58.1 | 41.3 | 79.0 |
| – | Qwen3-VL-4B | 80.1 | 78.3 | 72.9 | 63.5 | 40.2 | 85.0 |
| – | Qwen3-VL-8B | 86.4 | 78.9 | 74.6 | 66.0 | 37.9 | 85.4 |
| – | Qwen3.5-2B | 83.3 | 73.8 | 72.3 | 61.4 | 58.6 | 78.9 |
| – | Qwen3.5-4B | 84.3 | 81.6 | 74.9 | 66.0 | 57.2 | 85.0 |
| – | Qwen3.5-9B | 83.3 | 82.4 | 77.4 | 71.4 | 52.1 | 85.8 |
| *Trained with W2S-PG (w/ AS3)* | | | | | | | |
| Qwen3-VL-2B | Qwen3-VL-4B | 93.6 (+13.5) | 83.4 (+5.1) | 82.6 (+9.7) | 68.5 (+5.0) | 59.3 (+19.1) | 88.4 (+3.4) |
| | Qwen3-VL-8B | 94.1 (+7.7) | 86.8 (+7.9) | 84.5 (+9.9) | 71.6 (+5.6) | 59.5 (+21.6) | 89.0 (+3.6) |
| Qwen3.5-2B | Qwen3.5-4B | 93.8 (+9.5) | 84.2 (+2.6) | 81.0 (+6.1) | 69.2 (+3.2) | 64.8 (+7.6) | 88.6 (+3.6) |
| | Qwen3.5-9B | 94.4 (+11.1) | 88.4 (+6.0) | 85.2 (+7.8) | 72.3 (+0.9) | 65.1 (+13.0) | 89.5 (+3.7) |

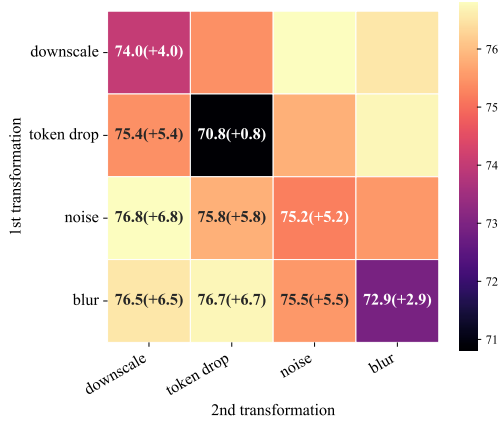


Figure 2: **Ablation study on different augmentation types** using individual augmentations or pairs of augmentations. Diagonal entries correspond to a single transformation with one branch as the identity transformation, and off-diagonals correspond to pairs of two different transformations (one applied to each view).

# 3 Experiments

## 3.1 Experiment Settings

**Evaluation settings.** We evaluate W2S-PG on six perception evaluation settings across five benchmarks: VStar (Wu & Xie, 2024), MME-RealWorld (Zhang et al., 2025), HR-Bench at 4K and 8K resolutions (Wang et al., 2025b), CV-Bench (Tong et al., 2024), and ZoomBench (Wei et al., 2026). We report the average performance over three evaluation runs for each setting. Following the official evaluation protocols, we use top-$p = 0.8$, top-$k = 20$, a maximum generation length of 16,384 tokens, and a temperature of 0.7.

**Model training.** We use 37K synthetic samples from ZwZ (Wei et al., 2026) as our training set. We use Qwen3-VL-4B/8B (Bai et al., 2025) as the student models and Qwen3-VL-2B as the weak teacher model. Training hyperparameters are reported in Table 5.

**Baselines.** We consider different OPD variants as baselines. For vanilla OPD, we use Qwen3-VL-2B and Qwen3-VL-32B as the weak and strong teachers, respectively. Furthermore, we include OPD-$\Delta$ (Feng et al., 2026), another weak-to-strong distillation approach that uses the output difference between the pre- and post-RL models. We first train Qwen3-VL-2B with DAPO (Yu et al., 2025) and then use the output difference between the RL-trained and original Qwen3-VL-2B models as the teacher signal. We keep all other settings fixed for a fair comparison with other OPD variants. We also verify the effectiveness of AS3 in this section.

## 3.2 Results

**Comparison with other training approaches.** Table 1 shows the results of different training approaches using the same data on six perception benchmarks. We find that vanilla OPD with a stronger teacher model (i.e., Qwen3-VL-32B) consistently improves the student model. In contrast, using a weaker teacher (Qwen3-VL-2B) decreases the performance of student models. OPD-$\Delta$

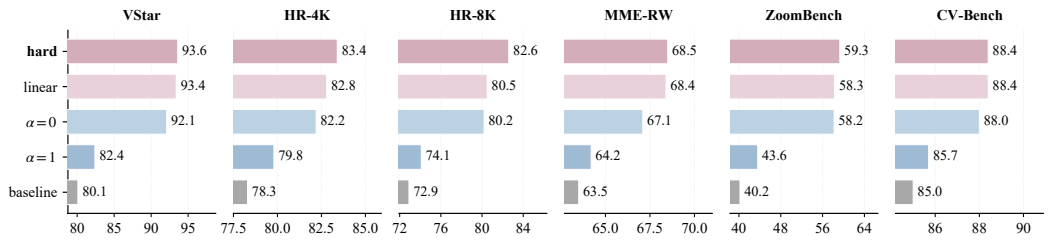


Figure 3: **Ablation study on different design choices for $\alpha_j$.** The numbers on the x-axis represent performance on the corresponding benchmark. $\alpha_j = 1$ and 0 denote that we fully trust supervision signals from the student or the teacher, respectively. Linear (Eq. (10)) and hard (Eq. (4)) (the default setting in AS3) are two choices that adaptively select supervision signals based on the confidence.

improves the student model on most benchmarks by distilling the output differences of the weak model before and after RL training. The proposed method, W2S-PG, consistently improves the student model across all benchmarks by a substantial margin. Notably, W2S-PG improves Qwen3-VL-4B by 12 points on VStar and 18 points on ZoomBench. AS3 yields further gains, suggesting that combining teacher-derived and self-training signals is beneficial in the weak-to-strong setting.

**Different student–teacher pairs.** We consider two additional Qwen3.5 student–teacher pairs: Qwen3.5-4B with Qwen3.5-2B and Qwen3.5-9B with Qwen3.5-2B[1] in Table 2. We observe that W2S-PG generalizes well across different model families. Consistent improvements are observed across all benchmarks and four student–teacher pairs. Notably, Qwen3.5-9B trained with W2S-PG and AS3 achieves gains of 13.0 points on ZoomBench and 7.8 points on HR-8K.

### 3.3 Ablation Study

In this section, we conduct ablation studies to verify the impact of each component of W2S-PG and AS3. We also investigate the effects of different augmentation types and the design choice of $\alpha_j$. All experiments are conducted with Qwen3-VL-4B as the student model and Qwen3-VL-2B as the teacher model.

**Augmentation ablation.** To understand the effects of individual data augmentations and the importance of augmentation composition, we investigate the performance of W2S-PG when applying augmentations from $\mathcal{T}$ in a controlled study using one or two augmentations. We study whether each augmentation is effective and whether composing two augmentations improves performance.

Figure 2 shows results across six perception benchmarks. Diagonal entries correspond to applying one transformation to one view while leaving the other view unchanged (i.e., using the identity transformation). Gaussian noise yields the best performance, while the other augmentations also provide improvements. For off-diagonal entries, a different augmentation is applied to each of the two views. These results show that combining two different transformations generally improves performance compared with a single transformation. In the main experiments of Table 1, we combine all augmentation types to provide more diverse visual conditions. Using all augmentation types outperforms using only two.

**Choice of $\alpha_j$ and divergence.** We set $\alpha_j$ to the binary values 0 and 1 to entirely trust the supervision signal from either the teacher or the student, respectively. We also consider another variant that uses a soft value of $\alpha_j$, formulated as

$$\alpha_j = c_j/(c_j + c_j^T). \tag{10}$$

This variant uses a linear weighting scheme based on the confidence scores.

Figure 3 shows the results with different choices of $\alpha_j$. $\alpha_j = 0$ corresponds to using Eq. (3) as the training objective, i.e., relying solely on the weak teacher. $\alpha_j = 1$ corresponds to training the student on its own generated outputs. We find that $\alpha_j = 0$ significantly improves performance

[1] https://huggingface.co/collections/Qwen/qwen35

Table 3: **Ablation study on different divergence measures of** $\pi_T^{w2s}(\cdot|q, x_1, x_2)$. $\mathcal{D}_{\text{JSD}}$ denotes Jensen–Shannon divergence. We highlight the default setting.

| | HR-8K | ZoomBench |
|---|---|---|
| Qwen3-VL-4B | 72.9 | 40.2 |
| $\log(\pi_T(\cdot\|q, x_1)/\pi_T(\cdot\|q, x_2))$ | 78.1 | 56.8 |
| $\mathcal{D}_{\text{KL}}(\pi_T(\cdot\|q, x_1)\|\pi_T(\cdot\|q, x_2))$ | 78.7 | 55.1 |
| $\mathcal{D}_{\text{JSD}}(\pi_T(\cdot\|q, x_1)\|\pi_T(\cdot\|q, x_2))$ | 79.8 | 57.7 |
| $(\pi_T(\cdot\|q, x_1) - \pi_T(\cdot\|q, x_2))^2$ | 80.2 | 58.2 |

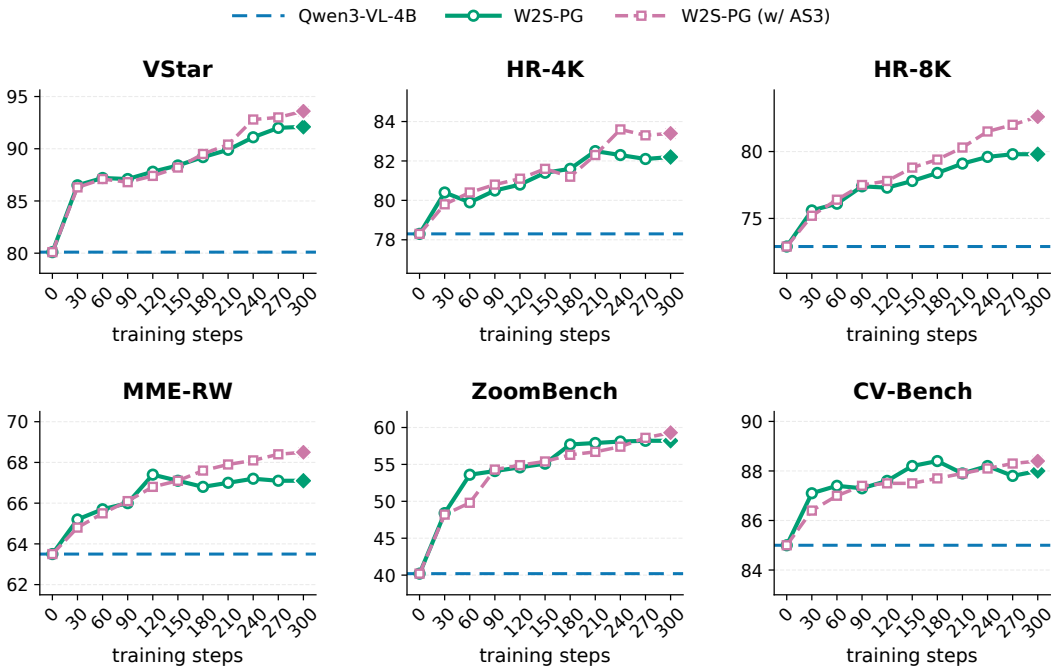


Figure 4: Performance of different benchmarks during training of W2S-PG and AS3.

in weak-to-strong distillation. $\alpha_j = 1$ also improves performance, but is suboptimal. The two hybrid-supervision variants consistently outperform both, with the simple hard-weighting design performing slightly better. These results demonstrate the importance of adaptive hybrid supervision and robustness across different designs of $\alpha_j$.

We ablate different discrepancy functions for constructing $\pi_T^{w2s}(\cdot \mid q, x_1, x_2)$ in Table 3. Different divergence functions consistently improve over the baseline by substantial margins. Furthermore, symmetric divergences generally outperform asymmetric ones. This is intuitive because the discrepancy between the two outputs should be invariant to the ordering of the two augmented views. These results show that perception-gap supervision from the weak teacher is effective and robust to the choice of discrepancy function.

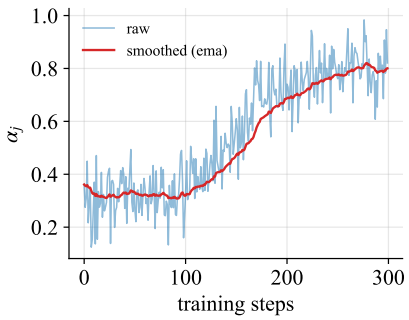


Figure 5: Average of $\alpha_j$ (cf. Eq. (4)) in one batch during training. We plot raw values and the smoothed curve.

**Training analysis.** We plot performance on each benchmark throughout training in Figure 4. We find that the performance with AS3 is less affected in the early stage of training. However, with the help of AS3, the performance continues to improve in the later stages of training, while W2S-PG alone tends to saturate. To examine how AS3 shifts between the two supervision sources during training, we visualize the batch-averaged value of $\alpha_j$ during training in Figure 5. We find that $\alpha_j$ increases during training, which shows that the proportion of student-derived

Table 4: **Evaluation results on 7 benchmarks** including reasoning and visual question answering. **+x** indicates improvement compared with the corresponding student model.

| | MathVista$_{mini}$ | MathVision | MathVerse$_{mini}$ | MMStar | CharXiv$_{reason}$ | MMMU$_{val}$ | MMMU-Pro |
|---|---|---|---|---|---|---|---|
| Qwen3-VL-2B | 61.3 | 31.6 | 52.1 | 58.3 | 26.8 | 53.4 | 36.5 |
| Qwen3-VL-4B | 73.7 | 51.6 | 46.8 | 69.8 | 39.7 | 67.4 | 53.2 |
| w/ W2S-PG | 77.4 (+3.7) | 56.5 (+4.9) | 48.1 (+1.3) | 71.4 (+1.6) | 51.2 (+11.5) | 71.8 (+4.4) | 57.3 (+4.1) |
| Qwen3-VL-8B | 77.2 | 53.9 | 62.1 | 70.9 | 46.4 | 69.6 | 55.9 |
| w/ W2S-PG | 82.9 (+5.7) | 60.4 (+6.5) | 64.8 (+2.7) | 72.1 (+1.2) | 57.5 (+11.1) | 76.3 (+6.7) | 60.4 (+4.5) |

supervision increases from approximately 0.2 to 0.8 as training progresses. This trend is consistent with the benchmark performance, which shows that student-derived supervision gradually becomes dominant as the student improves during training.

### 3.4 Generalization across Different Domains

Beyond perception benchmarks, we further evaluate W2S-PG on visual reasoning, visual question answering, and chart understanding tasks to verify whether W2S-PG generalizes to other tasks.

**Setup.** We take ViRL39K (Wang et al., 2025a), a diverse and high-quality multimodal dataset, as the training set. We use Qwen3-VL-4B/8B as the student models and Qwen3-VL-2B as the teacher. We evaluate on seven benchmarks, including MathVista (Lu et al., 2024), MathVision (Wang et al., 2024a), MathVerse (Zhang et al., 2024), MMStar (Chen et al., 2024), CharXiv (Wang et al., 2024b), MMMU (Yue et al., 2024), and MMMU-Pro (Yue et al., 2025). We use the same training hyperparameters as in Sec. 3.1.

**Results.** Table 4 shows the results for models trained with ViRL39K. We observe that W2S-PG remains effective across visual reasoning, visual question answering, and chart understanding tasks. For example, W2S-PG improves Qwen3-VL-8B by 11.1 points on CharXiv and Qwen3-VL-4B by 11.5 points. Significant gains are also observed on other benchmarks across different model scales. The results in Table 4 indicate the effectiveness of W2S-PG on tasks beyond perception.

## 4 Related Work

**On-Policy Distillation.** OPD (Gu et al., 2024; Agarwal et al., 2024) optimizes policy models with self-generated sequences under supervision from a stronger teacher. It is widely used in training small models (Yang et al., 2025) and merging multiple expert models (Xu et al., 2026). Standard OPD usually requires a more powerful teacher, typically a stronger or specialized model. In contrast, weak-to-strong distillation aims to learn from weaker teacher models, which eliminates the need for stronger teacher models. Concurrent works (Feng et al., 2026; Yu et al., 2026) achieve weak-to-strong distillation by contrasting model outputs before vs. after RL training, across model sizes, or under different hints. In contrast to these methods, we use the output difference of one model under different visual conditions as the supervision signal. Furthermore, we propose AS3 to adaptively select more reliable supervision rather than fully trusting the teacher signal.

**Weak-to-Strong Generalization.** Weak-to-strong generalization (Burns et al., 2024) aims to use a weak model to supervise a stronger model. A central motivation for weak-to-strong generalization is the need for scalable oversight when reliable human supervision or stronger models are unavailable. Recent work has extended this paradigm to the post-training of language models (Zhu et al., 2025; Feng et al., 2026). In this work, we study weak-to-strong generalization in the vision-language setting and explore output differences under different perception conditions as the supervision signal.

## 5 Conclusion

In this paper, we study weak-to-strong on-policy distillation, in which a policy model is trained using supervision from a weaker teacher. Rather than directly using the teacher model's outputs, we propose W2S-PG, which uses the output discrepancy of the same model under different visual conditions. To obtain more reliable supervision from both the student and the teacher, we propose AS3 that adaptively selects supervision signals at the token level. We verify the effectiveness of W2S-PG and AS3 on multiple benchmarks and show that W2S-PG outperforms vanilla OPD with a stronger or weaker teacher model. Furthermore, we conduct an ablation study to analyze the contributions of the proposed components and their key design choices.

### AI USE STATEMENT

In this work, we used generative AI tools solely for language polish and coding assistance. Specifically, GPT6 was used during the writing process to improve the clarity of the manuscript and correct grammatical errors. Codex was used as a coding agent during the experimental process to assist with code generation. Generative AI tools were not used to generate research ideas, formulate scientific claims or conclusions, or interpret experimental results.

All AI-assisted content was carefully reviewed and verified by the authors. In particular, all AI-generated code was manually inspected by the author team, and its correctness was verified through execution and examination of the resulting outputs. The entire manuscript, including AI-assisted language edits, was manually reviewed by the authors to ensure technical accuracy and consistency with the intended meaning. All experimental results and conclusions were independently verified by the author team.

We take full responsibility for the final content of this work, including all text, claims, code, experimental results, and other artifacts produced with the aid of generative AI.

## REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *ICLR*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *ICML*, 2024.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024.

Shiyuan Feng, Huan-ang Gao, Haohan Chi, Hanlin Wu, Zhilong Zhang, Zheng Jiang, Bingxiang He, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Weak-to-strong generalization via direct on-policy distillation. *arXiv preprint arXiv:2607.05394*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *ICLR*, 2024.

Jack Hong, Chenxiao Zhao, ChengLIn Zhu, Weiheng Lu, Guohai Xu, and XingYu. Deepeyesv2: Toward agentic multimodal model. In *ICLR*, 2026.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. In *ICLR*, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. In *NeurIPS*, 2025.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. VL-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. In *NeurIPS*, 2025a.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-vision dataset. In *NeurIPS*, 2024a.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *AAAI*, 2025b.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal LLMs. In *NeurIPS*, 2024b.

Lai Wei, Liangbo He, jun lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, Zhuosheng Zhang, and Weiran Huang. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. In *ICML*, 2026.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *CVPR*, 2024.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Fangxu Yu, Zinan Lin, Xiaodong Liu, Weijia Xu, Michael Xu, Tianyi Zhou, and Jianfeng Gao. Weak-to-strong on-policy distillation. *arXiv preprint arXiv:2607.26246*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *NeurIPS*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *ACL*, 2025.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *ECCV*, 2024.

Yifan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. Mme-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *ICLR*, 2025.

YiFan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think beyond images. In *ICLR*, 2026.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and XingYu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In *ICLR*, 2026.

Wenhong Zhu, Zhiwei He, Xiaofeng Wang, Pengfei Liu, and Rui Wang. Weak-to-strong preference optimization: Stealing reward from weak aligned model. In *ICLR*, 2025.

Table 5: Training hyperparameters of W2S-PG.

| Hyperparameter | Value |
|---|---|
| Max Generation Length | 16384 |
| Learning Rate | 2e-6 |
| LR Scheduler | Constant |
| Rollout Temperature | 1.0 |
| Batch Size | 128 |
| Training Steps | 300 |
| Training Backend | FSDP2 |
| Rollout Backend | SGLang |

Table 6: The system prompt used in training models with ZwZ 37K (Wei et al., 2026).

```
You are a helpful assistant.  The assistant first thinks about
the reasoning process in the mind and then provides the user
with the answer.  The answer is enclosed within <answer>
</answer> tags, respectively, i.e., reasoning process here
<answer> answer here </answer>.
```

## A Experimental Setup Details

**Training.** Table 5 summarizes the training hyperparameters. Table 6 shows the prompt used in training.

**Evaluation.** In Sec. 3.1, for perception benchmarks, we follow the evaluation settings used in ZwZ (Wei et al., 2026). For evaluations on other tasks in Sec. 3.4, we use `lmms-eval`[2] to ensure a fair comparison across different benchmarks. We use decoding parameters following the official evaluation guidelines and summarize them in Table 7.

## B Comparison with Agentic Baselines

To address perception tasks involving high-resolution images, another popular line of research explores agentic approaches, i.e., "thinking with images". They typically use external tools to zoom in on or crop regions of interest and then re-encode these regions during inference. We compare against several representative agentic baselines: Pixel-Reasoner (Su et al., 2025), DeepEyes (Zheng et al., 2026; Hong et al., 2026), Thyme (Zhang et al., 2026), Mini-o3 (Lai et al., 2026). Furthermore, we add ZwZ (Wei et al., 2026), a region-to-image distillation approach for reference.

Table 8 shows a model-level comparison. W2S-PG is competitive with, and in some cases even surpasses, publicly reported agentic models. Compared with ZwZ, W2S-PG combined with AS3 outperforms ZwZ by a clear margin, demonstrating the effectiveness of perception-gap supervision.

[2] https://github.com/evolvinglmms-lab/lmms-eval

Table 7: Summary of decoding parameters during evaluation.

| Model | Temperature | Top-p | Top-k | Presence penalty | Repetition penalty | Max tokens |
|---|---|---|---|---|---|---|
| Qwen3.5 | 1.0 | 0.95 | 20 | 1.5 | 1.0 | 32768 |
| Qwen3-VL | 0.7 | 0.8 | 20 | 0.0 | 1.0 | 16384 |

Table 8: Comparison of W2S-PG models (single forward pass) with agentic models on several perception benchmarks.

| Models | VStar | HR-4K | HR-8K | MME-RW-en |
|---|---|---|---|---|
| *Base models (single forward pass)* | | | | |
| Qwen3-VL-4B | 80.1 | 78.3 | 72.9 | 63.5 |
| Qwen2.5-VL-7B | 78.5 | 71.6 | 67.9 | 58.3 |
| Qwen3-VL-8B | 86.4 | 78.9 | 74.6 | 66.0 |
| *"Thinking-with-Images" agentic models*† | | | | |
| Pixel-Reasoner | 84.3 | 72.6 | 66.1 | 64.4 |
| DeepEyes | 85.6 | 75.1 | 72.6 | 64.1 |
| Thyme | 82.2 | 77.0 | 72.0 | 64.8 |
| DeepEyesV2 | 81.8 | 77.9 | 73.8 | 64.9 |
| Mini-o3 | 88.2 | 77.5 | 73.3 | 65.5 |
| *Qwen3-VL with crop tool* | | | | |
| Qwen3-VL-4B+tool | 88.0 | 81.3 | 74.4 | 65.4 |
| Qwen3-VL-8B+tool | 90.1 | 82.3 | 78.0 | 66.0 |
| *Trained models with ZwZ-37K (single forward pass)* | | | | |
| ZwZ-7B | 88.5 | 75.4 | 73.3 | 66.2 |
| ZwZ-4B | 92.7 | 81.8 | 79.5 | 68.5 |
| W2S-PG | 92.1 | 82.2 | 80.2 | 67.1 |
| W2S-PG w/ AS3 | 93.6 | 83.4 | 82.6 | 68.5 |
| ZwZ-8B | 91.1 | 84.4 | 82.0 | 69.9 |
| W2S-PG | 93.2 | 85.0 | 82.8 | 70.3 |
| W2S-PG w/ AS3 | 94.1 | 86.8 | 84.5 | 71.6 |

†: Evaluation results are from (Wei et al., 2026). Results are included for reference only because their base models and training sets differ from ours.