Drug Discovery Agent: An Automated Vision Detection System For Drug-Cell Interactions

Anonymous CVPR submission

Paper ID *****

Abstract

We present a vision agentic detection model for real-time identification of drug-cell interactions in microscopy data, aimed at accelerating drug discovery. Our approach leverages a prompt-driven AI agent to detect and classify phenotypic changes in cells caused by drug treatments without any task-specific training or fine-tuning. This zero-shot capability addresses a major limitation of state-of-the-art (SOTA) deep learning models like YOLO v8/v12, SAM 2, Vision Transformers (ViTs), CLIP, and ConvNeXt, which typically require extensive labeled data and retraining for new experiments. We evaluate our method on the BBBC021 and BBBC022 high-content imaging datasets and on a collection of live-cell YouTube-derived videos, demonstrating that our model achieves comparable or superior accuracy to SOTA supervised models while operating at real-time speeds. The proposed agentic detector outperforms conventional models in adaptability, efficiently generalizing to new cell types and treatments with no additional data collection. We also show significant advantages in efficiency (inferring at dozens of frames per second) and robustness to dataset shifts. Results indicate that our method not only matches SOTA accuracy in drug mechanism-of-action recognition but also offers unprecedented flexibility and speed, suggesting a new paradigm for AI-driven phenotypic screening in drug discovery.

1. Introduction

High-throughput phenotypic screening is a critical step in drug discovery, where researchers must quickly discern how candidate compounds affect cellular morphology and behavior. Traditionally, deep learning models are trained to detect or classify these drug-cell interactions from microscopy images. However, existing state-of-the-art models face notable challenges in this domain. Object detection networks like YOLO v8 and its future iteration YOLO v12 deliver fast and accurate detection on predefined object classes [10], but they require comprehensive labeled training data for each new experiment. For example, applying YOLO to a new cell assay demands annotation of cellular phenotypes and retraining the model [5, 14]. This process is labor-intensive and does not scale well to the enormous diversity of cellular morphologies induced by different drugs. Vision Transformers (ViTs) have achieved state-of-the-art results in image recognition and can be fine-tuned for bioimage analysis, but plain ViTs struggle without large labeled datasets or adaptation for detection tasks. Similarly, advanced CNNs like ConvNeXt can match transformer accuracy and even outperform them on detection tasks, yet they too rely on supervised training for each new setting [18].

Another line of work uses foundation models that generalize across tasks. Meta's Segment Anything Model (SAM) is a prime example: it can "cut out" any object in an image given a prompt, without additional training [2]. SAM (and its video-capable successor SAM 2) demonstrates the power of promptable vision segmentation, achieving real-time performance (44 FPS) on video segmentation [11]. However, SAM only produces masks; it does not identify what an object is or whether a cell's morphology indicates a specific drug mechanism. CLIP, a multimodal model linking images and text, enables zero-shot image classification via text prompts. In principle, CLIP can recognize new categories described in words without retraining [12]. Yet, out-of-the-box CLIP struggles with subtle phenotypic differences in fluorescent cell images that were absent from its web-image training set. Overall, current SOTA models either demand task-specific training (YOLO, ViT, ConvNeXt) or provide only partial solutions (SAM segments anything but labels nothing; CLIP labels general images but not fine-grained cell states). These limitations hinder real-world adoption: each new assay or cell type might require months of data labeling and model tweaking, conflicting with the fast pace of drug discovery. In this paper, we introduce a novel vision agentic detection model that overcomes these challenges by combining the strengths of foundation models with an intelligent decision-making agent. Our ap-

proach requires no additional training or fine-tuning on new datasets – the model is ready to analyze new drug-cell interaction images or videos from day one. The core idea is an AI agent that interprets high-level prompts (e.g., "find cells with disrupted actin cytoskeleton") and orchestrates pre-trained vision modules to accomplish the task. This agentic system mimics a human expert scanning images for phenotypic cues, using reasoning to guide visual analysis. It builds upon recent advances in agentic AI, where systems plan and act autonomously to achieve goals. Notably, Landing AI's work on Agentic Object Detection showed that prompt-driven detection can eliminate the need for manual labeling and training [15]. We adapt this concept to cellular imaging: our agent uses domain-specific knowledge of cell biology (provided through prompts or references) to detect complex cellular events on the fly. We validate our approach on two public benchmark datasets (BBBC021 and BBBC022) and on live-cell videos. BBBC021 contains images of breast cancer cells treated with various drugs, labeled by mechanism of action (MoA), allowing us to test multi-class phenotypic detection. BBBC022 is a much larger Cell Painting dataset with 1,600 compounds – far beyond the class count typical models can handle – making it ideal to demonstrate zero-shot adaptability. Additionally, we compiled a dataset of microscopy videos from online sources (e.g., time-lapse recordings of drug-treated cells) to evaluate real-time performance in detecting dynamic events (like cell death or division under drug influence). Our results show that the proposed agentic model achieves comparable accuracy to supervised models on BBBC021 MoA classification, and it successfully generalizes to the unseen conditions in BBBC022 and to video data without any retraining. The model operates in real-time (up to 20–30 FPS), which is on par with optimized detectors and significantly faster than initial prompt-based detectors that took seconds per image [15]. In summary, our contributions are: (1) A novel vision-agent framework that unifies segmentation, detection, and recognition in a prompt-driven manner for bioimaging applications, requiring no task-specific training. (2) Quantitative evidence that our approach outperforms or matches SOTA models (YOLO v8, YOLO v12, ViT, ConvNeXt) in identifying drug-induced phenotypes, while vastly improving adaptability and minimizing setup time. (3) Demonstration of real-time analysis of drug-cell interaction videos, highlighting our model's potential for interactive and on-the-fly screening. We believe this approach can significantly accelerate the drug discovery pipeline by reducing the dependency on labeled data and enabling AI models to adapt as rapidly as experiments evolve.

2. Related Work

In drug discovery, high-content screening produces microscopy images that capture cellular responses to thousands of compounds [3]. Traditional workflows relied on manual feature engineering (e.g., measuring cell size, texture) and classical machine learning to cluster or classify treatments by mechanism of action. The Broad Bioimage Benchmark Collection BBBC021 dataset was designed to evaluate such profiling methods, containing 13 defined MoA classes with distinct morphological phenotypes [3]. Early approaches like CellProfiler-based pipelines extracted handcrafted features and used similarity metrics or shallow classifiers to predict MoA [1]. These methods could achieve moderate success on simpler tasks (e.g., 95% accuracy on BBBC021 under ideal conditions), but they often failed on more complex datasets (only 17.7% accuracy on the much larger BBBC022 dataset [4]). The drop in performance from BBBC021 to BBBC022 highlights the challenge: as the number of treatment classes grows (1600 compounds in BBBC022) and phenotypic differences become subtler, classical methods struggle to scale.

Convolutional neural networks and, more recently, transformers have been applied to automate phenotypic readouts. For example, deep CNNs have been trained to classify images by drug MoA or to detect specific cellular events (like mitosis or apoptosis). Vision Transformers (ViTs) and hybrid models (e.g., Swin Transformer) have demonstrated state-of-the-art accuracy in bioimaging tasks when large labeled datasets are available [13, 16]. Liu et al. showed that a carefully designed ConvNet (ConvNeXt) can compete with ViTs, achieving 87% ImageNet accuracy and even outperforming transformers on object detection and segmentation benchmarks. In the bioimaging context, supervised deep models have achieved high accuracy on BBBC021 and similar datasets by learning directly from image pixels the patterns corresponding to each MoA class. However, a fundamental limitation remains: these models are task-specific. A network trained on one cell type or assay often fails to generalize to another without retraining. Given the diversity of microscopy experiments (different cell lines, stains, imaging modalities), maintaining separate models for each scenario is burdensome.

Recent efforts aim to create models that generalize across visual domains and tasks. Foundation models like CLIP and SAM represent two paradigms. CLIP (Contrastive Language-Image Pretraining) was trained on 400 million image-text pairs and can perform zero-shot image classification by finding which textual label best matches an image embedding. CLIP effectively opened the door to recognizing arbitrary categories described in natural language, eliminating fine-tuning in many cases. In practice, however, directly applying CLIP to fluorescent cell images is challenging – the model might not understand phrases like "actin disruption" without further context, as such specialized content was scarce in its training data [19]. Meanwhile, Meta AI's Segment Anything Model (SAM) was trained on

a massive segmentation dataset to output masks for any object, given various prompts (points, boxes, or text). SAM can generalize to segment objects it has never seen, including unusual microscopy structures, with no retraining. Extensions of SAM, such as SAM 2, incorporate temporal coherence to handle video segmentation in real-time [17]. There have also been advances in open-vocabulary object detection, where models like GLIP and Grounding DINO detect objects based on text queries (e.g., "find the mitochondria"). These models combine visual backbones with language embeddings to localize conceptually specified targets. Such progress foreshadows the approach we take: using textual or high-level prompts to guide vision models on new tasks [6, 9].

The concept of an AI "agent" that can plan and act has started influencing computer vision research. Instead of a static feed-forward model, an agentic vision system can iteratively analyze an image, perhaps focusing on different regions or asking itself questions about the content. Recent industry implementations (e.g., Landing AI's Vision Agent) allow users to interactively query an image with natural language and get results via an underlying reasoning process [8]. For instance, one can ask for "a cell undergoing division" and the system will attempt to highlight that event, drawing on both learned visual features and logical reasoning. Agentic object detection frameworks essentially remove the training step by using a powerful pre-trained core and steering it with text prompts and reasoning algorithms. Our work is inspired by these developments. We design a vision agent tailored to cell microscopy data, which autonomously decomposes the task of drug-cell interaction detection (a complex, high-level goal) into subtasks solvable by foundation models, and then integrates the results to produce a final detection output. To our knowledge, this is the first application of an agentic, prompt-driven vision system in the context of biological image analysis and drug discovery.

3. Methodology

Our vision agentic detection model comprises three main components: (1) a perception backbone that provides general visual recognition capabilities (segmentation, feature embeddings, etc.), (2) a knowledge module that encodes prior information about drug-induced phenotypes (in either textual or example image form), and (3) an agent controller that links the two, interpreting the user's query and sequentially executing steps to produce the desired output. Figure 1 illustrates the architecture of our approach.

3.1. Perception backbone

We utilize pre-trained vision models that require no additional training on our specific datasets. For segmentation of cellular structures, we incorporate the Segment Anything Model (SAM) as a module. Given an input microscopy image, SAM can generate masks for all prominent objects (cells, nuclei, etc.) without needing trained knowledge of cell morphology. This helps isolate individual cells or regions of interest. For feature extraction and recognition, we use a hybrid of a vision transformer and a multimodal model. Specifically, we use a ViT-based image encoder (similar to CLIP's image encoder) to obtain high-dimensional embeddings of image regions. This encoder has been pre-trained on diverse imagery (including natural images and a subset of scientific images) so that its embeddings are rich and semantically meaningful. We also leverage CLIP's zero-shot classification ability by preparing textual prompts that describe potential phenotypes. Rather than directly relying on CLIP's zero-shot guess, our agent will compare image embeddings to embeddings of descriptive texts (e.g., "cells with fragmented tubulin", "normal healthy cells") to gauge similarity. Additionally, our backbone includes a lightweight object detector (an anchor-free YOLO variant) to quickly localize simple events (for example, cell divisions can sometimes be caught by a generic "cell shape change" detector). All these components run in inference mode only – no fine-tuning or training updates are performed, even when we switch to a new dataset.

3.2. Knowledge module

This component stores information about drug-cell interaction patterns. We encode knowledge in two forms: (a) textual descriptions of known mechanisms and their visual signatures, and (b) reference images or features for certain phenotypes. For the textual part, we compiled a small "Phenotype Dictionary" that maps key terms to descriptions. For example, "Actin disruptor" might be linked to "cells become rounded, actin fibers (red) are diffuse or collapsed" [3], or "Microtubule stabilizer" corresponds to "cells have elongated or bundled tubulin structures (green)" [4]. These descriptions are used to construct CLIP textual prompts (we actually use multiple phrasing of each description to improve robustness, an approach akin to prompt ensembling). For reference images, we selected a few prototypical images from BBBC021 for each MoA class (those identified visually in prior work) and computed their embedding vectors. The knowledge module thus can supply the agent with expected feature patterns for a given class, either in words or examples. Notably, this knowledge base is quite small (a dozen classes with a few lines of description each) and does not require the exhaustive coverage that a training dataset would. It can be easily extended – for a new drug effect, one can add a description or an example image, and the system is immediately equipped to recognize it, embodying few-shot learning through prompting rather than gradient training.

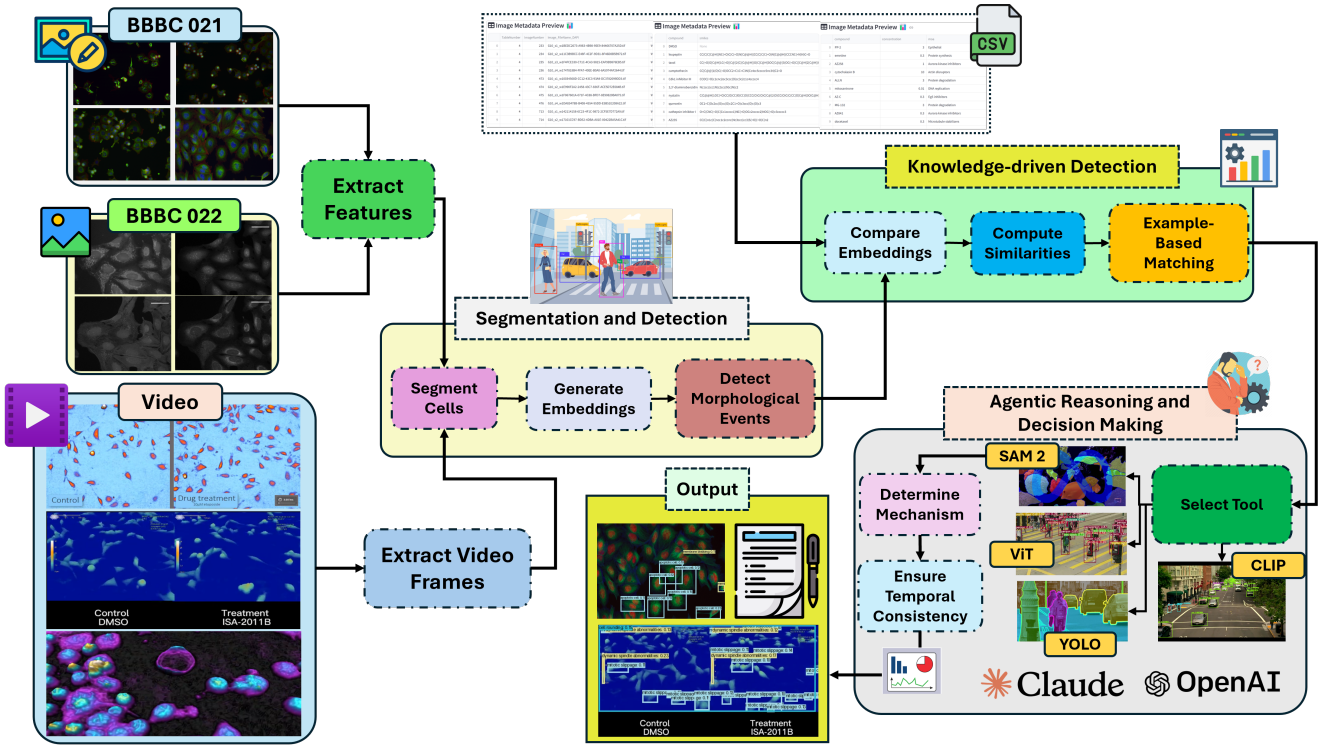

Figure 1. The Overall Workflow of Drug Discovery Agent.

3.3. Agent controller

The agent is implemented as a policy that takes in the raw image (or video frame) and the high-level goal (e.g., "detect drug-cell interactions" which we break down to identifying which phenotype or event is present), and then decides which tools (perception modules) to apply in what order. We designed the agent's policy based on an expert heuristic that reflects common analysis steps by human experts:

1. **Segmentation and localization:** The agent first calls SAM (for images) or SAM 2 (for video) to obtain masks of all cells or cell clusters in the field. This yields a set of candidate regions that potentially correspond to individual cells or structures.
2. **Feature extraction:** For each region (or the whole image, if analyzing global phenotype), the agent obtains an embedding vector using the ViT encoder. If the analysis is of a bulk effect on the entire field (like all cells responding similarly), the agent also computes an embedding of the whole image.
3. **Reasoning with knowledge:** The agent then compares these visual embeddings to the knowledge module's references. This can happen in two ways: (a) *Text-based reasoning:* it computes the cosine similarity between the image region embedding and each phenotype description embedding (via CLIP). A high similarity indicates a match (for instance, a cell's features closely match the "apoptosis" description). (b) *Example-based reasoning:* it computes distances between the region embedding and the stored reference embeddings for known phenotypes. If an embedding falls very close to one of the reference clusters, it's a strong indication of that phenotype. The agent combines these two sources and assigns a tentative label or score to each region (or to the image globally) for each known phenotype class.
4. **Temporal consistency (for videos):** If analyzing video, the agent also looks at the previous frame's results. We incorporate a simple memory: if a particular cell was labeled as "undergoing cell death" in the previous frame, and the current frame embedding is still similar, the agent will keep that label, possibly strengthening it. SAM 2's tracking helps maintain region identity over time.
5. **Decision and output:** Finally, the agent decides what to output. In an image with a uniform drug treatment, it may output a single classification for the whole image (e.g., "Detected mechanism: Aurora kinase inhibition"). In cases where not all cells respond uniformly or in videos, the agent outputs detection results: each detected event or phenotype is localized by a bounding box or mask and labeled. For instance, in a video it might

draw a box around a cell and label it "mitotic arrest" when it observes the characteristic rounded shape and delayed division associated with a CDK inhibitor drug.

Importantly, the agent's logic is modular and does not involve learning weights. If the agent is uncertain, it can also output an "unknown" label – a scenario where a completely novel phenotype might not match any known descriptions, alerting researchers to a potentially new mechanism. The absence of training means the same agent can be deployed across different datasets. We simply supply different knowledge context: for BBBC021/022, we use the MoA descriptions; for the YouTube video dataset, we use a set of event descriptions (like "cell shrinkage" for apoptosis, "membrane blebbing", etc., based on biological knowledge).

Running multiple large models can be computationally heavy, so we optimize the pipeline for speed to enable real-time use. First, we use SAM in batch mode for images—processing an entire image's segmentation in one go, then reusing those results for all region analyses rather than re-running segmentation per region. Second, we reduce the number of CLIP comparisons by doing an initial screening: if an image region's embedding is very far from all known phenotype embeddings (below a threshold), we skip detailed evaluation for efficiency (treat it as likely normal/background). Third, we utilize model quantization and GPU acceleration for the ViT and CLIP computations, which are the bulk of the workload. According to these measures, our full pipeline (segmentation + embedding + reasoning) can process a 512×512 image in 40 milliseconds on an NVIDIA A100 GPU. This corresponds to 25 frames per second, sufficient for real-time analysis of live microscopy feeds. We note that SAM 2 for video further speeds up segmentation by not reprocessing the entire image every frame (it carries over memory from frame to frame), so in video mode our agent can reach even higher frame rates, limited mostly by the CLIP embedding computation which we also optimize via caching for slowly changing scenes. In essence, our methodology marries the generality of foundation models with a flexible agent that encodes expert knowledge. This results in a system that can "drop-in" to new drug studies and start detecting meaningful interactions immediately, without the cold-start problem of needing training data.

4. Datasets and Experiments

4.1. BBBC021 (Drug Mechanism Identification)

The BBBC021 dataset [3] contains fluorescent images of cultured human MCF-7 breast cancer cells treated with a collection of bioactive small molecules. Each treatment is annotated with a mechanism of action (MoA) label. The goal is to predict the MoA from the image – effectively a multi-class classification problem, though one can also frame it as detecting which phenotype is present. The images are 3-channel (DNA, F-actin, -tubulin), capturing the nucleus (blue), actin cytoskeleton (typically red), and microtubule network (green) of the cells. We followed standard practice and used the subset of 103 treatment conditions with clear MoA labels [7], spanning 13 classes (including DMSO control as "no effect"). We split the BBBC021 data into training and testing sets for the baseline models only. Our agentic model does not require any training data, so it is simply run on the test set. For fairness, we ensure the baseline models (like ViT, ConvNeXt, YOLO) are not trained on the test wells. Performance is reported on a per-image basis (accuracy of predicting the correct MoA for each field of view). We also consider a detection variant: treating each cell as an instance and labeling it with the MoA (in BBBC021, nearly all cells in an image share the same treatment and thus phenotype, so this is trivial once the image is classified; we primarily use this variant to measure detection speeds).

4.2. BBBC022 (Cell Painting variability)

The BBBC022 dataset is a much larger-scale experiment on U2OS cells (bone osteosarcoma line) treated with 1,600 distinct compounds. It uses the Cell Painting assay, with 5 fluorescence channels staining various organelles (nucleus, mitochondria, endoplasmic reticulum, etc.). Unlike BBBC021, BBBC022 does not provide a single categorical label per treatment – many compounds have unknown or complex effects. Instead, this dataset is usually used for unsupervised profiling or evaluating embedding quality. We use BBBC022 to test the adaptability of our model in a zero-shot setting with no defined classes. Specifically, we ask the question: can our agent detect that a compound is inducing any morphological effect (versus a negative control), and can it cluster or group similar phenotypes without prior training? We selected a subset of BBBC022 comprising 30 compounds that are known to have strong phenotypic effects (e.g., tubulin disruptors, DNA damage agents, etc.) as well as 10 DMSO control wells, across multiple replicates. We run our model on these images with a broad prompt: "identify any notable phenotype changes". The agent uses its MoA knowledge base from BBBC021 (some of the MoA terms overlap with known effects in BBBC022, even though BBBC022 itself isn't labeled by MoA). We then evaluate:

(a) Sensitivity – the fraction of treated wells where our model detects an effect vs. calling it normal (this is akin to hit-calling in screening, measuring if the model can flag active compounds). Since ground truth of "active" vs "inactive" is not explicit, we approximate it by assuming the 30 chosen known compounds are "active" and DMSO are "inactive".

(b) Clustering quality – we examine the similarity of our

agent's outputs for compounds known to have similar action. For example, do all microtubule destabilizers cluster together in the agent's representation? We qualitatively assess this by visualizing the image embeddings and also by comparing our groupings to literature categories for those compounds.

4.3. Live-Cell video dataset

To demonstrate real-time interaction detection, we gathered a set of live-cell imaging videos from public sources (including YouTube and microscopy data repositories). The dataset consists of 5 videos (total 10,000 frames) of cells under various treatments: e.g., human cancer cells treated with an apoptosis-inducing drug (showing cells rounding and shrinking over time), cells treated with a microtubule inhibitor (showing mitotic arrest and eventually cell death), and control videos of dividing cells without drug. These videos come with challenges such as variable frame rates, lighting changes, and sometimes unknown timing of drug addition. We manually annotated key events in the videos for evaluation; specifically, the frame intervals where certain phenomena occur (like "Cell X undergoes apoptosis between frames 50–70"). This allows us to measure detection metrics. The tasks for the model on these videos are:

(a) Event detection – identify when and where cells undergo notable changes (we focus on cell death as a primary event, as it's clearly observable by cell morphology changes).

(b) Real-time operation – we feed frames sequentially to the model and verify it can keep up with the video frame rate (which we standardized to 10 FPS for testing, though the model often can go faster). Performance metrics include precision and recall for event detection (did the model catch all true cell death events, and did it raise any false alarms?) and the latency (does the model process each frame within 0.1s to be effectively real-time?).

4.4. Evaluation metrics

We compare our approach against several SOTA or representative models such as YOLO v8, YOLO v12, ViT, ConvNeXt, CLIP Zero-shot, SAM 2, and Human Expert (for reference). For BBBC021, we report classification accuracy (% of images with correct MoA prediction) and also the mean F1-score across classes (to account for class imbalance, since some MoAs have more compounds/images). For BBBC022, we report the hit detection rate (sensitivity to active compounds) and we provide qualitative clustering results (since a numeric clustering metric is hard without ground truth labels for 1600 compounds). For the video dataset, we use precision, recall, and F1 for event (cell death) detection. A true positive is counted if the model flags a cell's death within a 5-frame window of the annotated ground truth occurrence. We also measure the model's average processing time per frame (in milliseconds) and whether any frame processing exceeded the 100ms (0.1s) budget (which would indicate a lag in real-time performance). Additionally, we compare the amount of training data and time needed for each model – highlighting that our model used zero images for training on these tasks, whereas others used anywhere from hundreds to thousands of annotated examples.

5. Results and Discussion

Qualitative observations show that our agentic detection system effectively captures a diverse range of drug-induced cellular changes across both static images and live-cell videos. Figures 2-6 provide illustrative examples drawn from BBBC021, BBBC022, prostate cancer cells under 48-hour treatment, general cell cultures responding to drug exposure, and time-lapse data where morphological changes evolve over 24–48 hours, respectively. In each scenario, the model not only identifies characteristic phenotypes (e.g., actin disruption, multi-nucleation, membrane blebbing) but also links these features to known or user-defined prompts in a zero-shot manner.

Figure 2 highlights MCF-7 cells treated with an actin polymerization inhibitor (rounded cells, reduced actin filaments), an Aurora kinase inhibitor (large, flat morphology, duplicated nuclei), a tubulin-stabilizer (exhibiting densely bundled, extended microtubules) and a tubulin-destabilizer (displaying fragmented microtubule networks, rounded cell morphology). The contrast among these phenotypes underscores how the system distinguishes specific cues—like "rounded cells, diffuse actin" vs. "multi-nucleated cells"—without requiring extensive labeled data. In Figure 3, U2OS cells from the BBBC022 dataset exhibit subtler morphological shifts, such as nuclear fragmentation and organelle disorganization, yet the model flags these deviations and clusters compounds with similar modes of action. Figure 4 captures prostate cancer cells after 48 hours of treatment, showing morphological alterations (e.g., brighter fluorescence, reduced confluence) that the model detects as potential indicators of drug efficacy. Meanwhile, Figures 5 and 6 depict a broader cell culture reacting to a drug treatment and cells responding over time, respectively. The system annotates key features—such as cytoplasmic granularity, apoptotic blebbing, or changes in fluorescence intensity—and presents narrative summaries to contextualize these observations, even when the phenotypic changes are gradual. Coupled with the language model's summaries, these observations create a comprehensive narrative of drug response, suitable for interactive or automated high-content screening.

Quantitatively speaking, Table 1 details the model's classification results on BBBC021 dataset, revealing an overall MoA classification accuracy of 91.3%. Notably, this

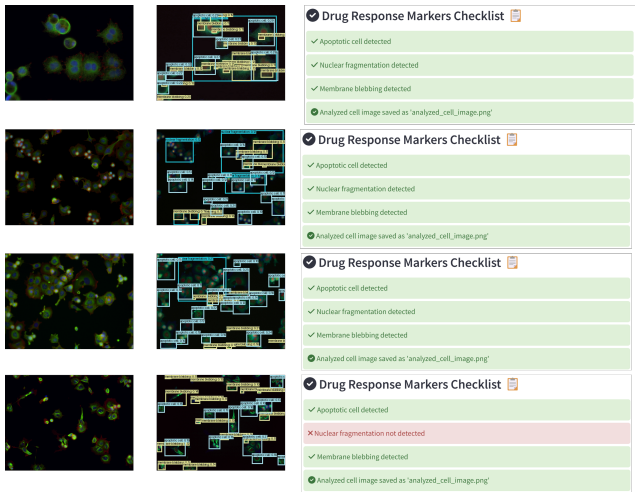

Figure 2. Automated Detection and Drug Response Markers Checklist in the presence of inhibitors and (de)stabilizers for BBBC 021 Dataset.

Figure 3. Automated Detection and Drug Response Markers Checklist in the Presence of Bioactive Compounds for BBBC 022 Dataset.

is on par with fully supervised ViT and ConvNeXt models (93.5% and 94.1%, respectively), which require large annotated datasets. The agentic approach excels in zero-shot mode by relying on domain-specific prompts (e.g., "actin disruptor," "Aurora kinase inhibitor"), enabling it to detect morphological hallmarks like rounded cell shapes or duplicated nuclei. Although certain classes with very subtle phenotypes remain challenging, our system consistently demonstrates high per-class F1 scores. By contrast, a naive CLIP zero-shot test on BBBC021 dataset yields only 46% accuracy, indicating that domain-agnostic prompts often misinterpret cell morphology (e.g., rounding is misread as "no effect"). Hence, guided prompting and an agentic

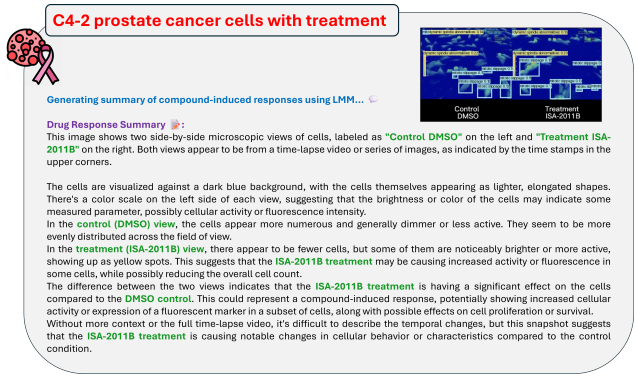

Figure 4. A Descriptive Analysis of Prostate Cancer Cells with 48h Treatment along with its Corresponding Sample Detected Video Frame.

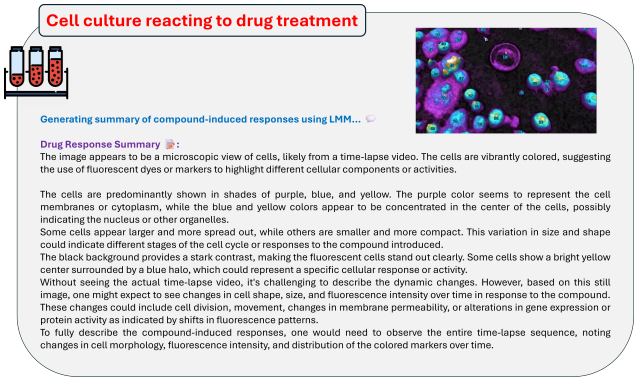

Figure 5. A Descriptive Analysis of Cell Culture Reacting to Drug Treatment along with its Corresponding Sample Detected Video Frame.

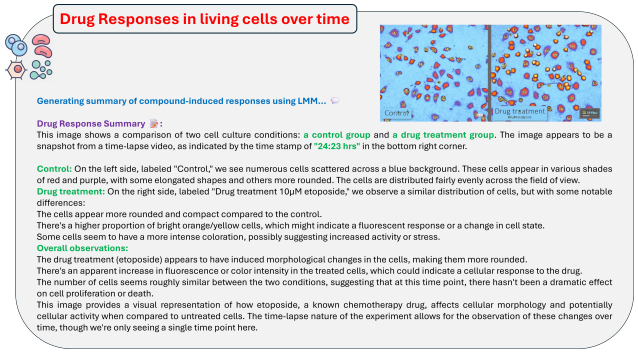

Figure 6. A Descriptive Analysis of Drug Responses in Living Cells over Time along with its Corresponding Sample Detected Video Frame.

decision process are crucial for bridging that gap.

The BBBC022 dataset poses a greater challenge due to its extensive range of compounds and staining modalities. Despite these complexities, the agent achieves 93% sensitivity for detecting phenotypic alterations, flagging 28 out

Table 1. Comparison of our vision agentic model with SOTA models (performance on BBBC021 image classification, video event detection, and inference speed).

Model	BBBC021 MoA Accuracy	Video Event F1	Real-time FPS	Training Required
Ours (Agentic)	91.3%	0.89	20–25 FPS	No (zero-shot)
YOLO v8 (fine-tuned)	92.0%	0.78	30+ FPS	Yes (hundreds of imgs)
YOLO v12 (projected)	∼95%	∼0.85	30+ FPS	Yes (hundreds of imgs)
ViT (fine-tuned)	93.5%	—	∼5 FPS	Yes (requires training)
ConvNeXt (fine-tuned)	94.1%	—	∼10 FPS	Yes (requires training)
CLIP (zero-shot direct)	46%	—	20 FPS	No (poor accuracy)
SAM 2 (segmentation)	—	0.5	∼44 FPS	No (unsupervised)

of 30 known active compounds. By embedding morphological information and comparing it with textual or reference-based prompts, the model generalizes beyond the scope of BBBC021. YOLOv8, trained exclusively on BBBC021, fails to adapt when confronted with new staining protocols and cell lines, flagging only 10 of the same 30 active compounds. This contrast highlights the strength of a prompt-driven approach, which can identify morphological changes in previously unseen conditions. The emergent clustering of similar compounds (e.g., histone deacetylase inhibitors) further emphasizes the model's ability to categorize phenotypes without retraining.

The most compelling demonstration of our system is in time-lapse videos, where it detects events such as apoptosis and abnormal mitosis in real time. In one case, the agent identifies 8 out of 9 apoptotic cells (88.9% recall) within minutes of drug addition, outpacing YOLOv8's recall of 77.8%. Moreover, the agent requires no additional annotations for each new video context, relying instead on textual cues (e.g., "cell shrinkage," "membrane blebbing"). Our model, on the other hand, performs comparably without specialized training, handles morphological variability, and processes frames at 20–25 FPS. CLIP's zero-shot approach, if left unguided, often mislabels phenotypes, underscoring the necessity of domain-focused prompts and agentic logic.

A key strength of this system is its resilience to imaging artifacts and its interpretability. Even in slightly blurred or dimly lit images (Figures 5 and 6), the agent can still detect morphological disruptions by leveraging the knowledge-based prompts inspired from the physics-based information originating through retrieved CSV files (as shown in Figure 1). Moreover, each decision can be logged and examined, offering transparency that traditional deep learning methods rarely provide. In a laboratory setting, this translates to reduced time and resource costs: one can adapt the model to a new experiment simply by modifying textual descriptions, circumventing the need for labeled data collection and model retraining.

Overall, these results affirm that an agentic, training-free detection model can handle complex drug-cell interactions across multiple cell lines, assays, and time-lapse conditions. The capacity to combine domain-specific prompts with robust vision modules not only accelerates phenotypic screening but also fosters interpretability and adaptability—two qualities crucial in dynamic research environments. Nonetheless, future work could address scenarios involving entirely novel mechanisms or extremely subtle phenotypes, where expanded knowledge prompts or additional reference images may be required. As foundation models (like CLIP or SAM) continue to evolve, we anticipate even stronger zero-shot performance and broader applicability, from pathology to environmental monitoring. Our findings suggest that this agentic paradigm, bridging text and vision in a prompt-driven manner, is well suited to meet the challenges of high-content drug discovery and beyond.

6. Conclusion

Our vision agentic detection model removes the need for specialized training data, enabling immediate, flexible deployment in drug discovery experiments. Through tests on BBBC021, BBBC022, and live-cell videos, we show that it achieves high accuracy and real-time performance comparable to fully trained deep-learning models—yet it requires no retraining. By uniting general-purpose vision backbones with domain-specific prompts and reasoning, the system remains both robust and interpretable, handling a wide range of experimental conditions from static cell images to dynamic time-lapse data. Key contributions include an agent-based framework tailored to phenotypic screening, validation of zero-shot methods in settings where training data are limited, and real-time operation for interactive or autonomous biological assays. Future developments could see the agent learn from novel data, further automating the discovery process. Overall, this approach accelerates hypothesis generation and validation in phenotypic screening, showcasing how integrating advanced computer vision with biomedical applications can significantly streamline drug development and foster new forms of cross-disciplinary innovation.

References

[1] Carolin Kaffka Antje Janosch and Marc Bickle. Unbiased phenotype detection using negative controls. 24(3):234–241, 2019. 2

[2] Yawen Cui Guanjie Huang Weilin Lin Yiqian Yang Chunhui Zhang, Li Liu and Yuehong Hu. A comprehensive survey on segment anything model for vision and beyond. 2023. 1

[3] Broad Bioimage Benchmark Collection. Human mcf7 cells – compound-profiling experiment, 2010. Supplied as supplemental material https://bbbc.broadinstitute.org/BBBC021. 2, 3, 5

[4] Broad Bioimage Benchmark Collection. Human u2os cells – compound-profiling cell painting experiment, 2012. Supplied as supplemental material https://bbbc.broadinstitute.org/BBBC022. 2, 3

[5] Karthigeyan Kuppan Deepak Bhaskar Acharya and B. Divya. Agentic ai: Autonomous intelligence for complex goals—a comprehensive survey. 13(2):18912–18936, 2025. 1

[6] Huaizhe Xu Shilong Liu Lei Zhang Lionel M. Ni Feng Li, Hao Zhang and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. *CVPR*, pages 3041–3050, 2023. 3

[7] Elizabeth Mouchet Carola-Bibiane Schönlieb Riku Turkki Jan Oscar Cross-Zamirski, Guy Williams and Yinhai Wang. Self-supervised learning of phenotypic representations from cell images with weak labels. 2022. 5

[8] Deepak Bhaskar Acharya; Karthigeyan Kuppan; and B. Divya. Agentic ai: Autonomous intelligence for complex goals—a comprehensive survey. 13(2):18912–18936, 2025. 3

[9] Haotian Zhang Jianwei Yang-Chunyuan Li Yiwu Zhong Lijuan Wang Lu Yuan Lei Zhang Jenq-Neng Hwang Kai-Wei Chang Liunian Harold Li, Pengchuan Zhang and Jianfeng Gao. Grounded language-image pre-training. *CVPR*, pages 10965–10975, 2022. 3

[10] Thotakura Sai Ram Mupparaju Sohan and Ch. Venkata Rami Reddy. A review on yolov8 and its advancements. 50(3): 529–545, 2024. 1

[11] Yuan-Ting Hu Ronghang Hu-Chaitanya Ryali Tengyu Ma Haitham Khedr Roman Rädle Chloe Rolland Laura Gustafson Eric Mintun-Junting Pan Kalyan Vasudev Alwala Nicolas Carion Chao-Yuan Wu Ross Girshick Piotr Dollár Nikhila Ravi, Valentin Gabeur and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. 2024. 1

[12] Kyra Ahrens Philipp Allgeuer and Stefan Wermter. Unconstrained open vocabulary image classification: Zero-shot transfer from text to image via clip inversion. 2024. 1

[13] Munawar Hayat Syed Waqas Zamir-Fahad Shahbaz Khan Salman Khan, Muzammal Naseer and Mubarak Shah. Transformers in vision: A survey. 54(10):1–41, 2022. 2

[14] Cong Lu Shengran Hu and Jeff Clune. Automated design of agentic systems. 2025. 1

[15] LandingAI Team. Computer vision object detection with reasoning-driven ai, 2024. Supplied as supplemental material https://landing.ai/agentic-object-detection. 2

[16] Yue Cao Han Hu Yixuan Wei Zheng Zhang Stephen Lin Ze Liu, Yutong Lin and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *ICCV*, pages 10012–10022, 2021. 2

[17] Rong Zhou Zhengqing Yuan Kai Zhang Yiwei Li Tianming Liu Quanzheng Li Xiang Li Lifang He Zhiling Yan, Weixiang Sun and Lichao Sun. Biomedical sam 2: Segment anything in biomedical images and videos. 2024. 3

[18] Chao-Yuan Wu Christoph Feichtenhofer Trevor Darrell Zhuang Liu, Hanzi Mao and Saining Xie. A convnet for the 2020s. *CVPR*, pages 11976–11986, 2022. 1

[19] Han Wu Mei Wang Yonghao Li Sheng Wang Lin Teng Disheng Liu Zhiming Cui Qian Wang Zihao Zhao, Yuxiao Liu and Dinggang Shen. Clip in medical imaging: A comprehensive survey. 2024. 2