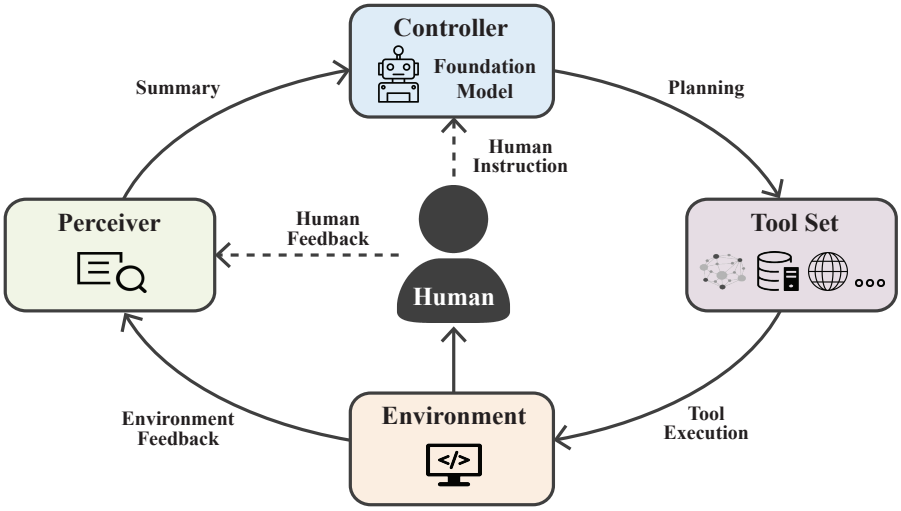

Figure 3: Illustration of the tool learning framework, where we display the human user and four core ingredients of the framework: tool set, controller, perceiver, and environment. The user sends an instruction to the controller, which then makes decisions and executes tools in the environment. The perceiver receives feedback from both the environment and the user and summarizes them to the controller.

representation of the tool, while a real environment involves actual interaction with the physical tool. Virtual environments have the advantage of being easily accessible and replicable, allowing for more cost-effective training for models. However, virtual environments may not fully replicate the complexities of the real-world environment, leading to overfitting and poor generalization (Hansen et al., 2021). On the other hand, real environments provide a more realistic context but may be more challenging to access and involve greater costs.

Controller. The controller $\mathcal{C}$ serves as the "brain" for tool learning framework and is typically modeled using a foundation model. The purpose of the controller $\mathcal{C}$ is to provide a feasible and precise plan for using tools to fulfill the user's request. To this end, $\mathcal{C}$ should understand user intent as well as the relationship between the intent and available tools, and then develop a plan to select the appropriate tools for tackling tasks, which will be discussed in § 3.2.1. In cases where the query is complex and targets a high-level task, $\mathcal{C}$ may need to decompose the task into multiple sub-tasks, which requires foundational models to have powerful planning and reasoning capabilities (§ 3.2.2).

Perceiver. The perceiver $\mathcal{P}$ is responsible for processing the user's and the environment's feedback and generating a summary for the controller. Simple forms of feedback processing include concatenating the user and environment feedback or formatting the feedback using a pre-defined template. The summarized feedback is then passed to the controller to assist its decision-making. By observing this feedback, the controller can determine whether the generated plan is effective and whether there are anomalies during the execution that need to be addressed. Under more complex scenarios, the perceiver should be able to support multiple modalities, such as text, vision, and audio, to capture the diverse nature of feedback from the user and the environment.

3.1.2 Connecting the Components

Formally, assume we have a tool set $\mathcal{T}$, which the controller can utilize to accomplish certain tasks. At time step t, environment $\mathcal{E}$ provides feedback e_t on the tool execution. The perceiver $\mathcal{P}$ receives the user feedback f_t and the environment feedback e_t, and generates summarized feedback x_t. Typically, the perceiver can be achieved by pre-defined rules (e.g., concatenating f_t and e_t) to form x_t, or modeled with complex neural models. The controller $\mathcal{C}$ generates a plan a_t, which selects and executes an appropriate tool from $\mathcal{T}$. This process can be formulated as the following probability distribution:

$$p_{\mathcal{C}}(a_t) = p_{\theta_{\mathcal{C}}}(a_t \mid x_t, \mathcal{H}_t, q), \tag{1}$$

where $\theta_{\mathcal{C}}$ denotes the parameters of $\mathcal{C}$, q denotes the user query or instruction, and $\mathcal{H}_t = \{(x_s, a_s)\}_{s=0}^{t-1}$ denotes the history feedback and plans. In its simplest form, a generated plan a_t can simply be a specific action for tool