Pilsner: A Compositionally Verified Compiler for a Higher-Order Imperative Language

Georg Neis1 Chung-Kil Hur2 Jan-Oliver Kaiser1 Craig McLaughlin3 Derek Dreyer1 Viktor Vafeiadis1
1MPI-SWS, Germany 2Seoul National University 3University of Glasgow

Abstract

Compiler verification is essential for the construction of fully verified software, but most prior work (such as CompCert) has focused on verifying whole-program compilers. To support separate compilation and to enable linking of results from different verified compilers, it is important to develop a compositional notion of compiler correctness that is modular (preserved under linking), transitive (supports multi-pass compilation), and flexible (applicable to a wide variety of languages and program transformations).

In this work, building on prior work of Hur et al. [6, 7], we develop a novel approach to compositional compiler verification based on parametric inter-language simulations (PILS). PILS are as modular and flexible as state-of-the-art logical-relations models, but are transitive as well. We demonstrate the effectiveness of PILS by using them to verify Pilsner, a simple but non-trivial multi-pass optimizing compiler for an ML-like language programmed in Coq. Pilsner is the first compiler for a higher-order imperative language to be compositionally and mechanically verified. This has been a significant undertaking, involving several person-years of work and over 48,000 lines of Coq.

1. Introduction

Most verification tools operate on programs written in high-level languages, which must be compiled down to machine-level languages prior to execution. The compiler is simply trusted to “preserve the semantics” of its source language (and hence preserve confidence prior to execution. The compiler is simply trusted to “preserve the semantics” of its source language (and hence preserve confidence prior to execution). Unfortunately, this task is well founded. For instance, recent work of Le et al. [12] identified 147 confirmed bugs in the industrial-strength GCC and LLVM compilers, of which 95 were violations of semantics preservation.

The goal of compiler verification is to eliminate the need for trust in compilation by providing a formal, machine-checked guarantee that a compiler is semantics-preserving. Toward this end, the most successful project so far has been CompCert [13], a verified optimizing compiler for a significant subset of C that was developed by Leroy and collaborators using the Coq proof assistant. Indeed, Le et al. [12] report that, despite extensive testing, they were unable to uncover a single bug in CompCert.

Now one may rightly ask: what does it mean for a compiler to “preserve the semantics” of the source program it is compiling? The standard answer, adopted by CompCert, is as follows: given a whole source program $p_S$, the compiler’s output $tr(p_S)$ should be a whole target program $p_T$ that refines $p_S$. Refinement means that any observable behavior of $p_T$ should also be a valid observable behavior of $p_S$, for some common notion of “observable behavior” that both languages share (e.g., termination, I/O events).

Although the above definition of semantics preservation is perfectly suitable for whole-program compilers, it says nothing about separate compilation. In practice, many programs are linked together from multiple separately-compiled modules, some of which may be compiled using different compilers or even hand-optimized in assembly. Is it possible to define a compositional notion of semantics preservation that says what it means for a single module in a program to be compiled correctly, while assuming as little as possible about how the other modules in the program are compiled?

We are not the first to broach this question—it has been an active research topic in recent years. But we argue that all previously proposed solutions are lacking in some dimension of compositional-ity. In particular, we articulate the following three desiderata for a compositional notion of semantics preservation:

- **Modularity:** To enable verified separate compilation, semantics preservation should be defined at the level of modules, not just whole programs, and it should be preserved under linking. That is, if a source module $m_S$ is refined by a target module $m_T$, and a source module $m'_S$ is refined by a target module $m'_T$, then the source-level linking of $m_S$ and $m'_S$ should be refined by the target-level linking of $m_T$ and $m'_T$. (This has been referred to as “horizontal compositional-ity”)

- **Transitivity:** Proofs of semantic preservation should be transitively composable. That is, one should be able to prove a compiler correct by verifying its constituent passes independently and then linking the theorems together by transitivity. (This has been referred to as “vertical compositional-ity”)

- **Flexibility:** It should be possible to prove semantics preservation for a wide range of compilers and program transformations, so that the results of different verified compilers (together with hand-optimized and hand-verified machine code) can be safely linked. In particular, we seek a notion of semantic preservation that imposes as few restrictions as possible on the kinds of program transformations and the kinds of source, intermediate, and target languages to which it is applicable.

Some previous approaches based on logical relations [1, 6] are modular and flexible but not transitive. Another approach, based on contextual equivalence and multi-language semantics [17], is transitive but not fully flexible or modular. Recently, a compositional version of CompCert has been developed [2], but its flexibility is limited in that it bakes in the assumption that the source and target languages of the compiler share the same (C-like) memory model, which is not the case for compilation of higher-level languages.

In this paper, we present Pilsner, a verified compiler for a higher-order imperative (ML-like) language, programmed in Coq. Pilsner translates a λ-calculus with recursive types, abstract types, and general references (the Source language $S$), through a CPS-based Intermediate language (I), down to an idealized assembly-like Target language (T). After CPS conversion, Pilsner performs several optimizations at the I level prior to code generation. These optimizations include function inlining, contification, dead code elimination, and hoisting (Figure 1). Although Pilsner is relatively simple—it is not nearly as realistic as the (whole-program) verified CakeML compiler, for instance [11]—it is the first compiler for a higher-order imperative language to be compositionally verified. Furthermore, it has been fully mechanized: this has been a significant undertaking, involving several person-years of work and over 48,000 lines of Coq.
We consider the setting of a high-level ML-like source language completely different source and target languages. As a proof of concept, they demonstrated the extreme flexibility of PILS (parametric inter-language simulations) between ML and assembly [7].

Parametric bisimulations are a simulation method for higher-order imperative languages, designed to support proofs that are modular, transitive, and capable of principle generalizing to inter-language reasoning, i.e., reasoning about relations between programs in different languages. However, Hur et al. only actually used them to reason about equivalence between programs written in a single (high-level) language, leaving open the question of whether the generalization to inter-language reasoning would pan out.

Hur and Dreyer’s work, on the other hand, was precisely targeted at supporting inter-language reasoning. Their Kripke logical relations introduced a rich and flexible notion of Kripke structures (possible worlds), with which they modeled the protocols governing calling conventions and the invariants connecting the different representations of data in the source and target languages of a compiler. As a proof of concept, they demonstrated the extreme flexibility of these Kripke structures by using them to verify the correctness of a deliberately obfuscated piece of self-modifying assembly code with respect to an ML-level function. However, due to limitations of their logical-relations method—in particular the lack of transitivity—their Kripke structures were only applicable to single-pass compilers.

PILS marries the benefits of these approaches together to support modular, transitive verification of Pilsner, a multi-pass compiler with completely different source and target languages. (We have ported over Hur and Dreyer’s self-modifying code example as well, giving evidence that PILS are as flexible as their logical relations.)

In the rest of the paper, we give a high-level overview of our main results (§2), review the basic idea behind parametric bisimulations which is also the core of PILS (§3), we describe the structure and some technical details of the PILS used in Pilsner (§4–§6), we highlight interesting aspects of the Pilsner verification (§7), and we conclude with discussion and related work (§8).

2. Results

2.1 Modularity

The goal of compiler correctness is to obtain a formal guarantee that the program that comes out of the compiler behaves the same as (or refines) the program that went in, according to a mathematical model of the source and target language in question. Traditionally, research on compiler correctness has focused on whole program compilation and does not support separate compilation. In separate compilation, the source program consists of several source modules, which are independently compiled to target modules. These target modules are then linked together, creating the final program. Note that different source modules may very well be compiled by different compilers.

We now illustrate how our approach, PILS, supports such separate compilation (and in fact even more heterogeneous scenarios). We consider the setting of a high-level ML-like source language $S$ and a low-level machine target language $T$ (for details, see §4).

The main component of our development is a relation between target modules and source modules: $\Gamma \vdash M_T \preceq_{TS} M_S : \Gamma'$. Intuitively states that target module $M_T$ refines source module $M_S$ and that they import the functions listed in $\Gamma$ and export those in $\Gamma'$. The first key result, Theorem 1, applies to whole programs, i.e., well-typed modules that import nothing and export at least a main function $(F_{main} : \text{unit} \rightarrow \tau) \in \Gamma$.

Theorem 1 (Adequacy for whole programs).

\[
\vdash M_T \preceq_{TS} M_S : \Gamma \quad (F_{main} : \text{unit} \rightarrow \tau) \in \Gamma \\
\text{Behav}(M_T) \subseteq \text{Behav}(M_S)
\]

Behav($\cdot$) denotes the set of I/O and termination behaviors that a program can have. The theorem implies for instance that, if $M_S$ always successfully terminates, then so does $M_T$ and moreover they produce the same outputs.

If we have a compiler that respects our relation $\preceq_{TS}$, then Theorem 1 gives us the same result as traditional whole-program compiler verification would. However, our relation also satisfies the following crucial property.

Theorem 2 (Preservation under linking, a.k.a. modularity).

\[
\Gamma \vdash M_T \preceq_{TS} M_S : \Gamma_1 \quad \Gamma_1 \vdash M'_T \preceq_{TS} M'_S : \Gamma_2 \\
\Gamma \vdash (M_T' \bowtie M'_T) \preceq_{TS} (M_S \bowtie M'_S) : \Gamma_1 \Gamma_2
\]

(Here, $\bowtie$ is overloaded notation for the linking operation both in the source and target language.) The theorem says that if we link two target modules, each of which is related to a source module, then the resulting target module is related to the linking of those source modules. Notice that, for the linking to make sense, the types of the first module’s exported functions (in $\Gamma_1$) need to match the second module’s assumptions. Of course, if a program consists of more than two modules, this theorem can be iterated as necessary and once linking results in a whole program, we can apply Theorem 1.

Observe that these properties don’t mention any particular compiler but are stated in terms of arbitrary related modules. The missing link is a theorem saying that the desired compiler adheres to our relation. We prove this for Pilsner, our compiler from $S$ to $T$, shown in Figure 1 (and discussed in detail in §7). Its correctness theorem applies to any well-typed source module.

Theorem 3 (Correctness of Pilsner).

\[
\Gamma \vdash M_S : \Gamma' \\
\Gamma \vdash \text{Pilsner}(M_S) \preceq_{TS} M_S : \Gamma'
\]

We would like to stress two important points. PILS were designed with flexibility in mind and make only few assumptions about the translation of source programs, namely details of the calling convention and in-memory representation of values (see §5). Consequently:

1. Nothing stops us from proving a theorem analogous to the previous one for a different compiler from $S$ to $T$, which may use an entirely different intermediate language (or several ones).

2. Nothing stops us from proving the relatedness of a source and target module by hand, e.g., when the target module is not the direct result of a compiler run but was manually optimized (see §2.3 for an extreme example of this, where the target module overwrites its own code at run time).

Hence, we can preserve correctness not only by linking Pilsner-produced code with other Pilsner-produced code, but also by linking it with code produced by other compilers or even hand-translated. We only have to ensure that these translations are also correct w.r.t. $\preceq_{TS}$.

2.2 Transitivity

Proving a property like Theorem 3 can require a lot of effort: the more complex the compiler, the more complex its correctness proof.
It is thus crucial that a correctness proof can be broken up into several pieces, e.g., one sub-proof per compiler pass. PILS support such a decomposition thanks to a transitivity-like property. In our setting, where Pilsnere compiles via one intermediate language I, we can show the following:

**Theorem 4 (Transitivity).**
\[
\Gamma \vdash M_1 \preceqM I M_2 : [I'] \quad [I] \vdash M_1 \preceqM H M_2 : [I'] \\
\Gamma \vdash M_1 \preceqM IS M_2 : [I'] \\
\Gamma \vdash M_1 \preceqM TS M_2 : [I']
\]

Here, \(\preceqM I\) relates target modules to intermediate modules, \(\preceqM H\) relates intermediate modules to intermediate modules, and \(\preceqM IS\) relates intermediate modules to source modules. All are very similar to \(\preceqM TS\) and support similar reasoning principles. We will say more about them in § 5; for now suffice it to say that, since \(\preceqM I\) and \(\preceqM H\) involve only untyped languages,\(^1\) the relations themselves are "untyped" and we erase the typing annotations in their environments (e.g., written \([I]\)), leaving just a list of function labels. Notice how using the transitive closure of \(\preceqM I\) in the second premise of the rule allows us to verify each IL transformation separately.

### 2.3 Flexibility

As already mentioned above, PILS make few assumptions about details of a translation and we are confident that they are sufficiently flexible to account for other compilers and even optimizations by hand. To substantiate this claim, we have proven\(^2\) the challenging refinement from Hur and Dreyer [6] mentioned in § 1, which relies on tricky manipulations of local state, far more involved than those of any imaginable compiler.

This example is based on Pitts and Stark’s "awkward" example [18], which is easy to explain:

- \(e_0 := \lambda f. (x := 1 ; f())\)
- \(e_0 := \lambda f. (f(); 1)\)

Both expressions evaluate to higher-order functions that, when applied, call the argument “callback” function \(f\) and then return a number. In \(e_0\) this number is simply 1. In \(e_0\) it is the result of dereferencing a local (private) reference \(x\), which is initialized to 0. Notice, though, that when \(e_0\) is called for the first time, it immediately writes 1 to \(x\). Since there are no other writes, the value of \(x\) returned at the end will always be 1 as well. As a result, \(e_0\) and \(e_0\) are equivalent programs (see the next section for a proof sketch).

Hur and Dreyer [6] adapt this example by substituting for \(e_0\) a tricky self-modifying machine program that implements the same behavior, but in a rather baroque way. Figure 2 shows what this program looks like in memory. It is parameterized by the load address \(n\) and \(E(\_\_\_\_\_\)\) denotes the encoding of an instruction as a machine word. Notice that part of the code has been "encrypted" by adding 666 to its encoding. We briefly explain how the code works.

The first few lines allocate a new function closure with empty environment and code pointer \(n + 5\), and return it to the context. When this function gets called the first time, it starts out by decrypting the encrypted instructions (offsets 5–11), thus replacing the encrypted code in memory. Subsequently (offsets 12–14), it replaces its first instruction by a direct jump in order to skip over the decryption loop in future executions. The remaining code (offsets 15–23), which is also the target of that jump, simply performs the callback function call and then returns 1.

Hur and Dreyer showed that this contrived implementation refines the high-level program \(e_0\) as a demonstration that their KLRs approach is flexible enough to reason about semantically involved "transformations", even ones whose correctness relies on low-level internal state changes that clearly have no high-level counterpart. By verifying the same example (with respect to \(\preceqM TS\)), we aim to demonstrate that PILS are equally flexible. In fact, the high-level structure of our proof closely follows that of Hur and Dreyer’s proof because, as we explain in the next section, PILS and KLRs have a lot in common.

### 3. Background

PILS are essentially an inter-language generalization of the earlier work on parametric bisimulations (PBs) [7], which in turn emerged from prior work on KLRs [1, 4] in an attempt to overcome its limitations concerning transitivity. Both PBs and KLRs support the same high-level reasoning principles for higher-order imperative programs; they just do so in technically different ways. In this section, we give a bit of background on KLRs and PBs, and the similarities and differences between them.

#### What PBs and KLRs have in common: Protocols

One of the key features shared by PBs and recent work on KLRs is the ability to impose protocols on the local state of some functions, which describe how that local state is permitted to evolve over time. This idea was articulated most clearly by Dreyer et al. [1, 4], who proposed the use of various forms of state transition systems (STSs) to model these protocols and applied them to the verification of many challenging contextual equivalences.

The basic idea of protocols—and how intuitively one reasons about them—is easiest to explain by appeal to the simple “awkward” example presented in the previous section. We thus begin by walking through a proof of this example, using STS protocols, at a very informal level. At such a high level of abstraction, there is really no difference between a proof of the example using PBs vs. KLRs.

**Step 1:** Recall that we wish to prove \(e_0\) equivalent to \(e_0\). By symbolically executing \(e_0\), we see that it allocates a fresh heap location—call it \(l_x\)—which gets bound to \(x\). Since this location is fresh, and since it is kept private by \(e_0\), we can impose a protocol on it, dictating how its contents may evolve. We choose the following protocol, initially in state \(s_0\):

\[
\begin{align*}
\lambda x & \mapsto 0 & s_1 & \downarrow l_x \mapsto 1
\end{align*}
\]

This protocol expresses that the value at \(l_x\) currently is 0 (which indeed it is), but that it may eventually change to 1, after which point it will stay 1 forever.

**Remark.** This STS is very simple: besides consisting of only two states and a single transition, it also refers to only one of the two memories, namely that of \(e_0\). In general, an STS may refer to both memories, and even relate them to each other.

Now that we have installed this protocol on \(l_x\), the proof reduces to showing that the function values \(v_{\lambda x}\) and \(v_{\downarrow l_x}\) returned by \(e_0\) and \(e_0\)1
“behave equivalently” (and, in so doing, respect the protocol).

\[ v_a := \lambda f. (f \ v_b := 1; f \ v_b := 1); v_a \]

(\textit{Note that} \( v_b \) \textit{is just} \( e_2 \), \textit{since} \( e_2 \) \textit{was already a value}.)

So what does it mean to “behave equivalently”? This is really the big question, for which KLRs and PBs give different answers. Rather than try to answer it directly, we will instead describe two informal proof principles concerning behavioral equivalence that, at the level of abstraction we are working at here, are supported by both proof methods, and we will finish the proof sketch by just appealing to these proof principles. After that, we will explain how the different proof methods implement these principles.

**Principle 1 (Showing Behavioral Equivalence):** To show that \( v_a \) and \( v_b \) \textit{behave} equivalently, it suffices to show that they \textit{behave} equivalently when applied to any arguments \( f_a \) and \( f_b \) passed in from the environment, which we may \textit{assume} are equivalent.

**Principle 2 (Using Assumed Equivalence):** If \( f_a \) and \( f_b \) are \textit{assumed} equivalent, then \( f_a () \) and \( f_b () \) \textit{behave} equivalently.

Note that these proof principles make a distinction between when two functions \textit{behave} equivalently and when they \textit{are assumed} equivalent. We explain the difference between these notions below. Note also that we restricted Principle 2 here to functions with unit argument; this is merely to simplify our informal discussion.

**Step 2:** By Principle 1, suppose that we are given \( f_a \) and \( f_b \) passed in from the environment, that \textit{are assumed} equivalent; it remains to show that \( v_a, f_a \) and \( v_b, f_b \) \textit{behave} equivalently. When the execution of these functions begins, the state of \( l_i \)'s protocol could be either in \( s_0 \) or \( s_1 \), since the functions could be called at some point in the future. In either case, \( v_a, f_a \) first sets \( l_0 \) to 1, thereby either updating the state of the protocol to \( s_1 \) (if it was in \( s_0 \) to start) or leaving it as—is—either way, a legal transition.

**Step 3:** We are now trying to show that \( (f_a (); l_0) \) and \( (f_b (); l_1) \) \textit{behave} equivalently, given that the protocol is now in state \( s_1 \). Since \( f_a \) and \( f_b \) \textit{were assumed equivalent}, we know by Principle 2 that \( f_a () \) and \( f_b () \) \textit{behave} equivalently. The result therefore reduces to showing that \( l_0 \) and \( 1 \) \textit{behave} equivalently. We may assume that these expressions are executed starting in state \( s_1 \), because that is the only state accessible from the state \( s_1 \) we were in previously (\textit{i.e.}, we assume the protocol has been respected by \( f_a \) and \( f_b \)).

**Step 4:** Since we know we are still in state \( s_1 \), we know that \( l_0 \) evaluates to 1. Thus, \( l_0 \) and 1 \textit{behave} equivalently, so we are done.

**Logical relations.** Both KLRs and PBs allow one to turn the above proof sketch into a proper proof. The key difference between them is how they formalize behavioral vs. assumed equivalence.

KLRs formalize this by defining a relation, which says—one and for all—what it means for two expressions to be indistinguishable at a certain type. One then uses this \textit{same} “logical” relation as the definition of \textit{both} behavioral and assumed equivalence. For expressions, the logical relation says that they are equivalent if they either both run forever or they both evaluate to equivalent values. For function values, which both the \( e \)'s and the \( f \)'s in our example are, the logical relation says that they are equivalent if they map logically-related arguments to logically-related results. Principles 1 and 2 both fall out of this definition as immediate consequences.

The main difficulty with logical relations is that, by conflating behavioral and assumed equivalence, they introduce an inherent circularity in the construction of the logical relation. In particular, the definition of equivalence of function values refers recursively to itself in a negative position (when quantifying over equivalent arguments). Traditionally, for simpler languages (e.g., without recursive types or higher-order state), this circularity is handled by defining the logical relation by induction on the type structure. For richer languages, such as the source language of Pilsner, induction on types is no longer sufficient, but step-indexing can be used instead to stratify the construction by the number of steps of computation in the programs being related [1]. This is the approach taken by Hur and Dreyer [7] in their earlier work on compositional compiler correctness. However, it is not known how to prove transitivity of logical relations for step-indexed models.

**Parametric bisimulations.** This problem with transitivity was one of the key motivations for \textit{parametric bisimulations} (PBs). Unlike logical relations, PBs treat behavioral and assumed equivalence as distinct concepts. In particular, rather than trying to define assumed equivalence, PBs take assumed equivalence as a \textit{parameter} of the model (hence the name “parametric bisimulations”). That is, a PB proof that two expressions are behaviorally equivalent is parameterized by an arbitrary unknown \textit{relation} \( U \) representing assumed equivalence.\( ^3 \) and \( U \) could relate any functions \( f_a \) and \( f_b \).

To make use of this unknown \( U \) parameter, PBs update the definition of behavioral equivalence accordingly. For function values, one can show them behaviorally equivalent precisely as suggested by Principle 1, \textit{i.e.}, if they map \( U \)-related (assumed equivalent) arguments to behaviorally equivalent results. For expressions, behavioral equivalence extends the definition from logical relations with a new possibility, namely that the expressions may call functions \( f_a \) and \( f_b \) related by \( U \). This amounts to baking Principle 2 directly into the definition of behavioral equivalence. The reason this is necessary—\textit{i.e.}, the reason Principle 2 does not just fall out of the definition otherwise—is that \( U \) is a \textit{parameter} of behavioral equivalence. Knowing that \( f_a \) and \( f_b \) \textit{are assumed equivalent} according to \( U \) tells us absolutely nothing about them! Consequently, Principle 2 must be explicitly added to the definition of behavioral equivalence as an extra case (called the “external” call case, as it concerns calls to “external” functions passed in from the environment).

One can understand PBs as defining a \textit{“local” notion} of behavioral equivalence: two expressions are behaviorally equivalent if they behave the same in their local computations, \textit{ignoring} what happens during calls to \( (U\)-related) external functions passed in from the environment. Intuitively, this is perfectly sound: it just means each module in a program is responsible for its own local computations, not the local computations of other modules. Moreover, as we have observed already, it is largely a technical detail: PBs can support the same high-level protocol-based reasoning as KLRs do.\( ^4 \)

The major benefit of PBs over KLRs is that they avoid the problems with the circularity of KLRs which necessitated step-indexing. In particular, note that by taking Principle 1 as the definition of behavioral equivalence for function values, we avoid the negative self-reference that plagues logical relations: the arguments \( f_a \) and \( f_b \) are simply drawn from the unknown relation \( U \). This avoidance of step-indexing was essential in making it possible to establish that PBs do in fact support transitivity, but it does not mean that the proof of transitivity was easy. The interested reader can find further details in an earlier technical report [8].

4. PILS, Part 1: Languages

PILS generalize PBs to the inter-language setting of compiler verification. Recall from §2 that we are interested in compiling from an ML-like source language \( S \) to a machine language \( T \), and that our concrete compiler, Pilsner, employs an intermediate language \( I \). As

\( ^3 \) Hur \textit{et al.} [7] called this the “global knowledge”, in contrast to the “local knowledge”. We feel our terminology is more intuitive, and does not conflict with the global/local distinction as it pertains to worlds (§5.1).

\( ^4 \) There is one exception: PBs do not admit the eta law for function values. This is a known problem [7], with a known solution [9], but we leave it as future work to incorporate this solution into PILS.
argued before, besides the main similarity relation \( \leq_{TS} \) between \( T \) and \( S \) modules, we also need PILS to provide a similarity relation for each pair of languages adjacent in the compilation chain (Figure 1), i.e., \( \leq_{TT}, \leq_{TI}, \) and \( \leq_{IS}. \)

4.1 Language-Generic Approach

In order to avoid duplicate work, we define PILS in a language-generic way, i.e., we define similarity \( \leq_{AB} \) for two abstract languages \( A \) and \( B \) (for some notion of abstract language to be described), and then instantiate it with different language pairs in order to obtain the desired relations. This has two important benefits:

1. Most of the metatheory, which is quite involved, can be established once and for all. This is particularly crucial because PILS were developed from the start in Coq, and over time the definitions—and thus proofs—had to undergo countless changes. This might simply have been infeasible if it weren’t for the language-generic setup.

2. One can easily instantiate PILS with a different intermediate language (or several of them) in order to verify a different compiler. Using modularity (Theorem 2), one can then of course safely link \( T \) code produced by this compiler with Pilsner-produced code.

In (1), we say “most of the metatheory” because transitivity and the parts of modularity and adequacy that deal with details of module loading are actually not proven generically. Ultimately, it would be nice to do so but it would require some effort to properly axiomatize various properties of the abstract language that these proofs rely on. Moreover, it might require a distinction between intermediate language and non-intermediate language, with slightly different sets of requirements. For now, it is much easier to simply prove the theorems for the concrete instances (of course this involves many generically proven lemmas). The downside of this is that, in order to verify a compiler using different intermediate languages, one needs to reprove the corresponding transitivity property. Adequacy and modularity, on the other hand, do not need to be reproven as they do not involve the intermediate languages.

To instantiate the generic PILS model and obtain one of the desired similarity relations requires us to provide: (i) the pair of concrete languages, and (ii) the global world for this pair. The latter can be seen as a predefined protocol (in the sense of §3) responsible for fixing calling conventions and data representations. We will discuss global worlds further in §5.

One point that we glossed over so far is that our generic definition is also not entirely generic—as we will see in the next section, it still essentially bakes in our source language’s type structure. Consequently, instantiating PILS as-is with a different source language most likely won’t make much sense. This is fine, because we focus on a single source language in this work. Extending this to multiple source languages, perhaps even allowing interoperability between them, is clearly important but left to future work.

Another point we glossed over is that we actually define two generic models: a typed one and an untyped one. The former is used when the input language is \( S \), the latter is used in all other cases (where no involved language has static types). However, we will continue to refer to them as just “the generic model”, because the untyped one is obtained simply by erasing all the type arguments from the typed one (highlighted in brown in the figures in §5).

4.2 Language Specification

We now describe the language abstraction in terms of which PILS are defined. In the subsequent sections, we then briefly present the concrete languages under consideration (\( S, I, T \)). Common to all

Domains: \( \text{Val, Cont, Conf, Mach, Mod, Anch} \)

Operators and relations:

- \( \text{cloud} \in \text{Mod} \rightarrow \text{Anch} \rightarrow (\text{Lbl} \times \text{Val})^* \rightarrow \mathcal{P} (\text{Conf}) \)
- \( \text{vload} \in \text{Mod} \rightarrow \text{Anch} \rightarrow (\text{Lbl} \times \text{Val})^* \rightarrow \text{Lbl} \rightarrow \mathcal{P} (\text{Val}) \)
- \( \text{cload} \in \text{Conf} \rightarrow \text{Conf} \rightarrow \mathcal{P} (\text{Comm} \text{utative and associative}) \)
- \( \emptyset \in \mathcal{P} (\mathcal{EVT} \times \text{Mach} \times \text{Mach}) \)
- \( \text{real} \in \text{Conf} \rightarrow \mathcal{P} (\text{Mach}) \)
- \( \text{error} := \{ m \in \text{Mach} \mid \forall c. m \notin \text{real}(c) \} \)

**Figure 3. Language specification**

Languages are a set of events and a countably infinite set of labels:

\[
  t \in \text{EVT} ::= \epsilon | ?n | !n \quad F_1, F_2, \ldots \in \text{Lbl}
\]

Events are produced by an executing program; they consist of internal computation (\( \epsilon \)) and I/O operations (reading or writing a number \( n \), respectively). Labels are used to identify module components; in this work, we consider a simplistic notion of module as the unit of compilation.

Figure 3 presents the abstract language in terms of a signature that any concrete language must implement. Keep in mind that we need to account for a very high-level language (\( S \)) on the one extreme and a very low-level language (\( T \)) on the other extreme. A language must come with a set \( \text{Val} \) of values, a set \( \text{Conf} \) of continuations, a set \( \text{Conf} \) of configurations, a set \( \text{Mach} \) of machines, a set \( \text{Mod} \) of modules, and a set \( \text{Anch} \) of anchors (think: load addresses). The core of the operational semantics is given in the form of a transition system \( \rightarrow \) of machines, whose transitions are labelled with events \( t \). Configurations can be thought of as partial machines—they play different roles in different contexts (e.g., they might represent just a heap or just an expression, or even a full machine). If a configuration \( c \) is complete, it is realized by a set of machines \( \text{real}(c) \). In all our instantiations, this is either empty, meaning the configuration is invalid or incomplete, or it contains exactly one machine.) Configurations must form a partial commutative monoid with composition \( \cdot \) and neutral element \( \emptyset \), except that the partiality is implicit via real. (Having \( \cdot \) be total is more convenient for mechanization [16]). We say a machine denotes an error, \( m \in \text{error} \), iff it does not realize any configuration.

Intuitively, modules are sets of labelled function values (the exports), which may refer to external functions (the imports) by their unique labels. In the language specification, the module interface consists of two operations, \( \text{cloud} \) and \( \text{vload} \). The former, \( \text{cloud} \), takes an anchor saying “where” the module is to be loaded and values for each of its imports. It then returns a set of initial configurations in which the module is considered loaded. Given the same inputs and additionally the label of one of the exported functions, \( \text{vload} \) returns the module’s corresponding function value.

4.3 Source Language \( S \)

The source language \( S \) is a standard PCF-like language extended with products, sums, universals, existentials, general recursive types, general reference types, and numeric I/O. Its type and term syntax is given in Figure 4.

Machines consist of a heap \( h \), a read-only environment \( \sigma \) for resolving labels to values, and an expression \( e \). Heaps are either undefined (\( \perp \)) or partial maps from locations to values. We assume the obvious composition operation \( \cdot \) for heaps (overloading notation) that returns the union of two heaps iff both are defined and they don’t overlap (otherwise it returns \( \perp \)). Note that the empty heap \( \emptyset \) is its neutral element. The step relation between machines is a pretty standard substitution-based left-to-right call-by-value reduction and we state only a few rules. If a machine cannot take a successful step (according to the usual rules such as beta reduction), then it steps to an error state by invalidating its heap component.
We define the semantics of a program, i.e., explicitly named and functions never "return". Concretely, we show in Figure 5 that we need to reason about substitutions when verifying optimizations, which evaluate to closures of code and environment. This avoids the need to store a closure stack.

### 4.4 Intermediate Language

#### I

In contrast to the source language, I is defined using an environment-based semantics where continuations and functions evaluate to closures of code and environment. This avoids the need to reason about substitutions when verifying optimizations, which is often a hassle.

In continuation-passing style, every subexpression is implicitly named and functions never "return". Concretely, we distinguish between (i) pure expressions $a$, which are evaluated

---

5 This is merely to keep the module semantics simple. PILS themselves, being a coinductive method, are perfectly compatible with mutual recursion.

6 in let-bindings and always yield a value without any side-effects, and (ii) control expressions $e$. Ignoring conditionals, every control expression is essentially a sequence of bindings ending in a function or continuation call. For instance, let $k y = e_1$ in $e_2$ defines a new continuation $k$ with argument $y$ and body $e_1$, and then executes $e_2$ (which may use $k$). Here $y \in \mathcal{T}Var$ is a term variable, while $k \notin \mathcal{K}Var$ is a continuation variable. Any $x$ in the language syntax stands for either a term variable or a label.

Modules, anchors, configurations, etc. are similar to those in the source language. We define $\text{Cont}$ simply as $\text{Val}$ because continuations are already values in the language.
Reg ⊳ r ::= sp | clo | arg | env | ret | aux | i
Oper ⊳ o ::= n | r | (r ± n) | [r ± n]
Instr ⊳ z ::= jmp o | jnz r o | di r o | sto o r | lpc r |
| bop r o₁ o₂ | input r | output r | alloc r₁ r₂

Val ::= Word
Anc ::= Word
Cont ::= Word
RegFile ::= Reg → Word
Stack ::= (Word → Word)⊥
Heap ::= (Word → Word)
Mach ::= Heap × Stack × RegFile × Word
real(c) := {m | m = c ∧ c.hp ≠ ⊥ ∧ c.hp finite ∧ c.st ≠ ⊥}

5. PILS, Part 2: Key Components

5.1 Worlds

Worlds are the formalism used by KLRs and PILS to incorporate the idea of protocols described in §3. Their shape is shown in Figure 7. The account of worlds and PILS we present here is somewhat simplified for the sake of presentation. For example, we omit an important distinction between public and private transitions in protocols, because it is inherited directly from the prior work on KLRs [6] and PBs [7]. At various points, we will discuss several ways in which our actual model (verified in Coq) is more sophisticated. Full details are given in the supplementary appendix.

We distinguish between global worlds and local worlds. Ultimately we relate programs under a full world, which is the composition of a global and a local one. The global world is a parameter of the definitions and needs to be instantiated together with the language implementations—we define exactly one global world for each language pair of interest (language implementations—we define exactly one global world for each language pair of interest). Let us now analyze the simplified presentation of E in Figure 8, first at a high level and then in more detail. Very roughly, it relates two programs e₀ (the “target” of a transformation) and e₁ (the “source” of a transformation) iff one of three cases holds:

(ERR) e₁ can silently (i.e., without I/O) produce an error.

(RET) e₀ is finished, and e₁ can silently finish returning a related value.

(STEP) e₀ can take a step and e₁ can match it (perhaps after some internal computation), “Match” means that both steps produce the same event and that the remaining computations either (REC) are again related by E, or (CALL) are about to call related “external” functions, i.e., functions related by the unknown relation U. This “external call” case is the characteristic feature of PBs, as discussed in §3. (K, not shown here, relates continuations and is defined in terms of E.)

In the explanation above, we glossed over many details. For a start, E is indexed implicitly by a full world W and explicitly by several parameters including the unknown relation U. We now take a closer look at the definition and discuss the key parameters.

Asymmetric small-step formulation (ASF). In contrast to the symmetric big-step formulation of E in Hur et al. [7], PILS employ an asymmetric small-step formulation. Notice how E’s step case asks us to consider possible steps of the “target” program e₀ in turn (using REC repeatedly), each time demanding us to match it with several steps of the “source” program e₁. Besides being seemingly necessary to properly deal with events (here: I/O), such an asymmetric small-step formulation is also important in the context of compiler verification because it gives the compiler the flexibility to remove erroneous behaviors of the source program and resolve some of its nondeterminism.

In this simplified presentation of E, e₀ is forced to take at least as many steps as e₁. Of course, this is overly restrictive and the actual definition relaxes this—more or less in the usual way (allowing stuttering), but with perhaps unusual implications. We discuss this in §6.2.

Protocol conformance. In order to talk about the execution of e₀ and e₁, we first need to “complete” these configurations and convert and environment that are missing). For T, on the other hand, they will usually just be program counters (i.e., heap, stack, and register file are missing). The missing parts will be provided by the world. In particular, in T, the global part of the world always owns the register file, because it is a global resource accessible by any module. Concretely this means that the state of the two global worlds involving T (for ) contains the current register file R and their crel only allows machine configurations whose register file is precisely R.
them into physical machines. These compositions should not be completely arbitrary; they should adhere to the world’s constraints at the current state $s$. Hence we quantify over $c_0$ and $c_0$, representing the portion of the machine state constrained by the world $W$, and require—in the helper definition $\text{configure}$—that they are indeed related by $W.\text{crel}(U)(s)$. We then attach these to $c_0$ and $c_0$, together with arbitrary frame configurations $v_0$ and $v_0$ representing the rest of the running program state. Finally, we only consider machines $m_0$ and $m_0$ that realize these composed configurations.

The two other occurrences of $\text{configure}$ (in $\text{STEP}$ and $\text{RET}$) are proof obligations. For instance, $\text{STEP}$ requires us to show that, after the step, each resulting machine can again be decomposed into a (possibly new) expression $c_0$, a (possibly new) configuration $c_0$, and the original frame configuration $v_0$, and we should not have touched the frame’s private state. Moreover, $c_0$ and $c_0$ must again satisfy $W$’s constraints, but we may advance $s$ to a future state $s'$ in order to achieve that.

Configuration queries. The $\text{RET}$ case has to assert (in a language-generic way) that the two computations have finished and returned similar values. In our source language, termination is a syntactic property and is trivial to check. But what does it mean for a $T$ computation to be “finished”? We take it to mean that control has reached the original return address. In order to determine whether this is the case, we need to know the return address in the first place.

This is why $E$ takes as arguments initial continuations $k_0$ and $k_0$. The actual check is then performed with the help of the world.

Its global part provides $\text{configuration query handlers}$ $\text{cqha}$ and $\text{cqhb}$ that answer questions such as “in state $s$, does $c_0$ constitute a return of value $v_0$ to continuation $k_0$?”, written $c_0 \in W.\text{cqha}(s)(\text{ret } v_0 k_0)$. For $T$, this amounts to checking that the program counter of $c_0$ (typically its only component) equals $k_0$, and that the return register contains $v_0$. The latter requires inspecting the given state $s$, because it contains the register file.

A different query, app, is used by the $\text{STEP/CALL}$ case to test if both configurations represent function calls. In effect, this means that the global world’s implementation of $\text{cqha}$ and $\text{cqhb}$ determines a major part of the calling conventions.

Value closure and value queries. Both the $\text{RET}$ and $\text{STEP/CALL}$ cases also require that certain values are related, e.g., the returned results: $(v_0, v_0) \in \langle\langle U(s')\rangle\rangle^s(\tau)$. But what is this relation exactly? As in PBs, $U$ is intuitively needed to relate unknown “external” functions passed in from the environment. The value closure operation $\langle\langle \rangle\rangle$ lifts it to other value forms in a straightforward way, e.g., by saying that two pairs are related iff their first projections are related (recursively) and their second projections as well.

When defining the value closure relation generically, we need to have a way of determining how $S$’s value forms are represented by languages $A$ and $B$. Since all modules must agree on the representations of values passed by external functions, it makes sense that the global world governs them. This is exactly the purpose of the global world’s $\text{value query handlers}$ $\text{vqha}$ and $\text{vqhb}$.

As an example, consider pairs. In $T$, we choose to represent a pair $(v_1, v_2)$ such that address $a$ holds the representation of $v_1$ and address $a + 1$ holds the representation of $v_2$. Since pairs are immutable in the source language, we must ensure that $a$ will continue to represent $(v_1, v_2)$ in the future.

We achieve this by having the global worlds involving $T$ maintain as part of their state a database of allocated values. Their query handler for $T$ then checks for a matching entry in this database. Moreover, their $\text{crel}$ requires that each value from the database actually exists in memory, with the expected address and representation. Finally, the associated transition system ensures that the database can only ever grow, thus implying that all registered values stay allocated forever. With this in place, the value closure operation can be defined analogously to the one for PBs (as a least fixed point). Some representative cases are shown in Figure 8.

The reader may wonder how we can show that two functions are related by $\langle\langle U(s')\rangle\rangle^s$ if they were defined by us and not passed in from the outside, i.e., known to be related by $U$. To do so, it suffices to show they are “similar”. This is because PBs (and PLS) impose a “validity” condition on $U$, namely that it relates at least all functions that behave “similarly”. Function similarity is defined essentially via Principle 1 in §3, that is: two functions are similar if they map arguments that are assumed-related (roughly related, by $U$) to results that are behaviorally-related by $E$.

5.3 A Note on the Untyped Model

Since our source language is type-safe and therefore its well-typed programs “don’t go wrong”, neither will correctly produced IL or target programs. One may thus wonder why our model takes faulty programs into account (the $\text{ERR}$ case in $E$). The answer is that this feature is actually crucial for verifying transformations in the untyped version of the model. (Recall that we obtain this version by erasing all the type arguments from the definitions in Figures 7–8.)
To see this, first consider the following optimization at the source level (where \( x \) is a variable of type nat):

\[
\text{fix } f(x). \text{if } x \text{ then } e \text{ else } e' \leadsto \text{fix } f(x). e
\]

In the process of showing that \( \text{fix } f(x). e \) is similar to the original function, we will be given arguments related at some unknown relation \( U \) and state \( s \), \((v_a, v_b) \in \langle U(s) \rangle^a\). Now, by inverting the definition of the value closure, we learn that \( v_a = v_b = n \), for some number \( n \).

Let us ignore the remaining proof steps and instead consider this transformation at the IL level, where we would be working in the untyped version of the model. There, we will still be given related arguments, \((v_a, v_b) \in \langle U(s) \rangle^a\), but this time the type information is missing. Consequently, when inverting the value closure, we don’t end up with the single case above (where \( v_a = v_b = n \)), but must also consider all the other cases, such as \( v_a \) and \( v_b \) being pairs. Now, note two important points: First, the global world \( W_I \)'s value query handlers ensure that whenever \( v_b \) is a number, then so is \( v_a \). In that case we can proceed as we would above in the typed model. Second, if \( v_b \) is not a number, then the original program produces an error and, thanks to err, there is nothing more to show.

6. Recovering Symmetric Big-Step Reasoning

As we explained in §5.2, PILS necessarily employ an asymmetric small-step formulation of \( E \). However, the resulting asymmetric small-step reasoning (ASR) is considerably less convenient than the symmetric big-step reasoning (SBR) supported by PBs. Here we discuss how we recover a (limited) form of SBR for PILS and exploit it in the Pilsner verification.

At a high level, one can show \( m_a \sim m_b \) using SBR by showing:

\[
m_a \leftrightarrow^* m_a' \wedge m_b \leftrightarrow^* m_b' \wedge m_a' \sim m_b'.
\]

Using ASR, on the other hand, one instead shows:

\[
\forall m_a'. m_a \leftrightarrow m_a' \implies \exists m_b'. m_b \leftrightarrow m_b' \wedge m_a' \sim m_b'.
\]

As one can see, SBR allows one to reason about multiple steps of computation in one go, but ASR forces one to reason one “target” step at a time. Ideally, we would like to use SBR whenever it is sound to do so. To achieve this for PILS, we developed three techniques.

6.1 Modelling Non-Interference with Internal Computation

The definition of \( E \) does not make any distinction between internal steps of computation (i.e., local steps whose effects are not visible to the environment) and other steps. As a result, after each step of the target program and matching steps of the source program, we must show that the memory constraints currently imposed by the world are again met. And then, in reasoning about the next step of the target program, we are forced to quantify over completely new configurations yet again.

This is unnecessarily strict. Intuitively, we should not need to show that the world’s conditions are satisfied again until the point where we pass control back to the environment; similarly, we should not need to quantify over new configurations except at points where control is passed to us, because there is no way that the state could have changed in between the internal steps of our local computation.

To solve this problem, we parameterize \( E \) with a boolean flag, signalling whether we are currently engaged in internal computation or not. The idea is that when we start a computation and are given configurations related by the world, we temporarily acquire them by merging them into our own configurations and setting the flag to true. As we continue executing local steps, we are freed from any worldly obligations. However, before we are allowed to use the RET or CALL case—i.e., when we want to give up control to the environment—we will be forced to release our grip on the world by setting the internal flag to false, at which point we must show that all the world’s constraints are again satisfied.

6.2 Allowing Stuttering in a Compositional Way

As mentioned in §5, our naive asymmetric small-step formulation forces the “source” program to take at least as many steps as the “target” program (due to the use of \( \leadsto^* \)). We want to allow the source to “stutter” by replacing \( \leadsto^* \) with \( \leadsto^+ \), but simply doing so would relate a diverging target program to any source program and thus render the model unsound.

To solve this we follow the standard approach [15] of indexing \( E \) by an element of a (proof-local) well-founded partially-ordered set (poset), and then demand that this element be decreased in the case of stuttering.

However, in order to support certain “compatibility” lemmas that are used in our compiler verification (see §7.1), we require a monoid structure on the order (technically speaking, we use non-trivial well-founded positive strictly ordered monoids, or “NWPS monoids” [3]). For example, the compatibility lemma for pairs is roughly as follows:

\[
e_a \sim_n e_b \wedge e_a' \sim_{n'} e_b' \implies (e_a, e_b) \sim_{n+n'} (e_b, e_b').
\]

Here \( e_b \) and \( e_b' \) can stutter at most \( n \) and \( n' \) times respectively, and thus \((e_b, e_b')\) can do so at most \( n + n' \) times. We therefore need not just the well-founded poset structure for \( n \), but also a notion of addition, which NWPS monoids provide.

In order to maximize the user’s flexibility, we provide a way to lift an arbitrary well-founded poset to an NWPS monoid. More specifically, given a well-founded poset \( X \), we construct an NWPS monoid \( X \) with an embedding of \( \tilde{X} \) into \( \tilde{X} \) (i.e., an order preserving and reflecting map from \( \tilde{X} \) to \( \tilde{X} \)). Thanks to this, the user can use an arbitrary well-founded poset without worrying about the additional monoid structure required by our proof framework.

6.3 Exploiting Local Determinism

The final obstacle is non-determinism. If the target language is non-deterministic, one must consider all possible executions that the machine can have, which is in conflict with SBR. The traditional solution to this problem (employed by CompCert) is to simply make the compiler’s intermediate and target language (internally) deterministic. This is quite restrictive in practice. Fortunately, we can do better.

The key idea is to lower determinism from being a property of a language to being a property of a machine configuration. This enables us to allow SBR for deterministic execution steps, even if the language itself is non-deterministic. In other words, we can reason about multiple steps at once but only until we hit a non-deterministic instruction.

7. The Pilsner Compiler

Using PILS, we have proven the correctness of Pilsner, a simple yet non-trivial compiler from \( S \) to \( T \) in Coq. Pilsner’s structure is depicted in Figure 1.

First, it translates the source module from \( S \) to \( T \) via a CPS-transformation. It also takes care to alpha-rename all bound variables such that in the resulting \( T \) module, every variable is bound at most once. This uniqueness condition simplifies the implementation of most of the subsequent optimization passes, as one does not have to worry about accidental variable capturing when rearranging code. Another nice characteristic of the produced intermediate code (not of \( T \) per se) is that continuations are used in an affine fashion [10], i.e., called at most once. This property is preserved by all other transformations at the intermediate level and enables a more efficient
treatment of continuation variables compared to ordinary variables in the code generation pass.

At the intermediate level, Pilsner performs six optimizations. It first inline selected top-level functions. For instance, if a module defines \( F = \text{fix}\, f(y, k), e \), then a call to \( F \) inside a subsequently defined function will be rewritten as follows:

\[
F \times k' \sim e[F/f][x/y][k'/k]
\]

(If the function is recursive, i.e., if \( f \) is used in \( e \), this is essentially just an unrolling.) Since inlining destroys the uniqueness property of bound variables, we immediately follow it with a “freshening” pass that re-establishes uniqueness.

Next comes coification, which we discuss in §7.2. Subsequently, Pilsner performs a simple dead code (and variable) elimination, rewriting let \( x = a \) in \( e \) whenever \( x \) does not occur in \( e \). This is justified because, in our IL, evaluation of an atomic expression \( a \) does not have any observable side effects. In the same manner, it also eliminates unused read operations and unused allocations (but not write operations because that would be unsound). Following DCE, it hoists let-bindings out of function and continuation definitions, subject to some syntactic constraints. For example:

\[
\text{let } f = (\text{fix}\, f(y, k), \text{let } z = x.i \text{ in } e) \text{ in } e' \\
\sim \text{let } z = x.i \text{ in let } f = (\text{fix}\, f(y, k), e) \text{ in } e'
\]

if \( x \) is none of \( f, y, k \). This avoids recomputation of the projection each time \( f \) is called.

Next comes a pass that commutes let-bindings (where possible) in order to group together bindings that assign names to the same expression. For example:

\[
\text{let } x = a \text{ in let } y = b \text{ in let } z = a \text{ in } e \\
\sim \text{let } x = a \text{ in let } y = b \text{ in } e
\]

The last IL transformation, derealization, gets rid of such consecutive duplicate bindings by rewriting the above expression as follows:

\[
\text{let } x = a \text{ in let } y = b \text{ in } e[x/z] \\
\sim \text{let } x = a \text{ in let } y = b \text{ in } e'
\]

This can be seen as a common subexpression elimination.

Code generation, the final pass in the chain, translates to the machine language \( T \). Recall that there are three kinds of “variables” in \( T \): term variables \( x \), continuation variables \( k \), and labels \( F \). Labels are translated to absolute addresses according to the import table. Term variable accesses are translated to lookups (based on position) in a linked list on the heap, pointed to by the env register. Functions are converted to closures, i.e., pairs of environment and code pointer (module-level functions simply have an empty environment), which live on the heap. A closure’s environment is loaded into the env register when the function is called. Finally, continuations are allocated on the stack. Accordingly, continuation variable accesses are translated to lookups (based on position) on the stack, with the side effect that the continuation in question, as well as all more-recently defined ones (above it on the stack), are popped. This is safe because the affinity property mentioned earlier ensures that they won’t be needed anymore.

### 7.1 Infrastructure for IL Transformations

For the local IL transformations in Pilsner, we developed a simple framework of transformations as expression annotations. The idea is to split module-level transformations into two parts: (1) an analysis that is applied to each top-level function and annotates selected subexpressions with to-be-performed micro-transformations (but does not actually rewrite the code); and (2) the micro-transformations themselves, together with their correctness proofs. Given these, we automatically produce a verified module transformation that analyzes the input module and performs transformations according to the generated annotations in a bottom-up manner.

A micro-transformation is a partial function on expressions—it must fail if the preconditions for its correctness do not hold. For instance, here is the (only) micro-transformation used in the commute-pass of Pilsner:

\[
\text{commute } e \mapsto e' \\
\text{commute}(e) := \text{let } y = b \text{ in let } x = a \text{ in } e_0 \\
\text{if } e \text{ is let } x = a \text{ in let } y = b \text{ in } e_0 \text{ and } x \notin \text{FV}(b)
\]

In the case that a micro-transformation fails (for which the analysis is to blame), the subexpression that was being transformed simply stays unchanged, or, alternatively, the whole module transformation (and thus the compiler) fails. In either case, if the module transformation succeeds, the output module is guaranteed to correctly implement the input module. This means that the analysis does not need to be verified—in the worst case, the transformation doesn’t optimize the code.

The concrete correctness property demanded by the framework for each micro-transformation \( f \) is twofold. **Syntactic correctness** says that \( f \) preserves well-formedness (including affinity of continuation variables) and the uniqueness condition. **Semantic correctness** states the following:

\[
f(e) = e' \quad \Gamma \vdash e \quad \text{unique}(\text{BV}(e), \Gamma) \\
\Gamma \vdash e' \not\approx e'\Gamma
\]

The relation in the conclusion is the reflexive transitive closure of \( \not\approx \), a fairly straightforward lifting of \( E \) from configurations to open \( I \) expressions (not to be confused with module similarity \( \not\approx_I \)). Its definition considers the expressions under an arbitrary unknown relation and state, with pointwise-related environments\(^6\) providing values for the term variables and labels in \( \Gamma \) as well as continuations for the continuation variables in \( \Gamma \). The local world that it uses is empty, which suffices for reasoning about Pilsner’s optimizations.

In order to ease the proofs of semantic correctness, we provide typical compatibility lemmas about \( \not\approx_I \). They state that the relation is preserved by each language construct of \( I \), and are very helpful in verifying transformations that leave parts of the module unchanged (even outside our annotation framework). The lemmas are straightforward but tedious to show. Only the one about recursive functions requires a proof by coinduction, as one would expect.

### 7.2 Coification

Pilsner includes a (very simplistic) coification pass. Coification \([5]\) is an optimization that turns a function into a continuation when the function is only ever called with the same continuation argument. This makes control flow more explicit, thus potentially enabling subsequent optimizations. In Pilsner, it has the additional benefit that continuations don’t need to be heap-allocated.

Coification in Pilsner uses the framework described above, i.e., it first runs an untrusted analysis that annotates places in the module where the expression-level coification should happen. This saves a lot of work because only the expression-level coification needs to be verified. This transformation consists of two steps. Given a coifiable function binding

\[
\text{let } f = (\text{fix}\, f(x, k), e) \text{ in } e'
\]

(for simplicity we gloss over the issue of variable uniqueness here), we first extend it with a fresh continuation definition, namely the coification of \( f \):

\[
\text{let } f = (\text{fix}\, f(x, k), e) \text{ in } e' \\
\text{let } k_y = e[y/x]/[k'/k] \text{ in } e'
\]

Here, \( k' \) is the continuation being passed in all invocations of \( f \) within \( e' \). In the second step, these invocations of \( f \) are turned into

\[^6\text{We actually allow variables in the “target” expression to be renamings of those in the “source” expression, as needed e.g. in the proof of deduplication.}\]
calls of the newly added continuation \( k_f \) (e.g., \( f \ z \ k' \sim k_f \ z \)). Note that compartmentalization does not merge the definition of \( f \), but leaves this to the dead code elimination pass. If the analysis is incorrect, then either the expression-level compartmentalization will fail or it will succeed but dead code elimination won’t be able to remove the original function binding.

Regarding verification of the expression-level transformation, obviously the first step is trivially correct as it only introduces an unused binding, and so the hard work lies in dealing with the second stage. The core property we need to show is that calls of \( k_f \) are related to calls of \( f \), i.e., something along the lines of:

\[
\Gamma \vdash k_f \ z \sim_{\text{II}}^* f \ z \ k'
\]

Of course this does not hold in such general form, because the connection between \( k_f \) and \( f \) would be lost. We must restrict attention to environments in which \( k_f \) (in the “target” program) actually maps to the contified version of whatever \( f \) maps to (in the “source” program). To do so, we generalize the \( \Gamma \vdash e' \sim_{\text{II}} e \) judgment to the form \( \Gamma; \Sigma_u; \Sigma_v \vdash e' \sim_{\text{II}} e \), where \( \Sigma_u \) and \( \Sigma_v \) each map a subset of \( \Gamma \) to closure values with concrete code expressions (if these subsets are empty, we obtain the original judgment). The property we then prove roughly looks as follows:

\[
\Sigma_u(k_f) = \lambda y. e[y/x][k'[k]] \quad \Sigma_v(f) = \text{fix } f(x, k). e
\]

\[
\Gamma; \Sigma_u; \Sigma_v \vdash k_f \ z \sim_{\text{II}}^* f \ z \ k'
\]

The generalized judgment is crucial because it supports the same compatibility properties as the original one did (modulo some new side conditions). These compatibility properties enable us to lift the above correctness result, which says we can rewrite \( f \ z \ k' \) to \( k_f \ z \), to a result which says we can rewrite all calls of \( f \) to calls of \( k_f \) inside the expression \( e' \) in which \( f \) is bound.

### 7.3 Verification of Code Generation

Code generation is the most complex transformation in Pilnsr, and so it comes as no surprise that its proof is also the most complex. Here we give a brief overview.

The goal is to show codegen(\( M \)) \( \sim_{\text{II}}^* \) \( M \) (ignoring contexts) for well-formed \( \Gamma \) module \( M \). With two simple inductions following the recursive definition of code generation (one on module well-formedness, one on expression well-formedness), the goal reduces to showing a collection of compatibility-like lemmas. Each of these states that one particular \( \Gamma \) expression form is related to the code generated for it. The hard work lies in defining this relation—let’s call it \( \sim_{\text{II}} \) (not to be confused with module similarity \( \sim_{\text{II}}^* \))—and then proving the lemmas.

In \( \sim_{\text{II}} \), we must ultimately say that the two programs are similar according to \( \text{E} \). But clearly the generated code makes many assumptions about its environment, e.g., where term variables can be looked up, how continuations are laid out on the stack, where temporary results are placed, etc. In order to restrict the environments in which the code is placed, we must therefore express parts of the compiler-internal protocol that the code follows. Naturally, the world plays a critical role here.

While the proofs of all other passes are oblivious to the local world, for the code generation proof it is critical that we can choose a particular one. We construct it such that its state is always a pair of a heap and a stack, representing the memory used internally by Pilnsr. This memory is used for storing (i) code, (ii) variable environment, and (iii) continuations.

Regarding (i), we use our local world in \( \sim_{\text{II}} \) to express the initial assumption that the generated code resides in memory and that the program counter points to the first instruction. This alone is not sufficient, though: when reasoning about an external function call we need to know that, when the function returns, our code is still in memory—otherwise, we wouldn’t even know which instructions get run next. To achieve this, we define our world’s transition relation such that it rules out any mutation of Pilnsr-generated code residing on the heap. (Note that this does not prevent other modules that we link with from using self-modifying code themselves.)

Regarding (ii), we use our local world in \( \sim_{\text{II}} \) to express that the env register points to a linked list in the heap—the variable environment—and that each of the values stored there is related to its counterpart in the \( \text{I} \) program’s environment by (the value closure of) \( U \). When reasoning about code involving a variable lookup, we can then be sure to get the correct value. Of course, when reasoning about code that extends the variable environment (e.g., code for a regular let-binding), we must be able to update the state accordingly. For this reason, while disallowing transitions that mutate the environment list, we do allow transitions that extend it.

Regarding (iii), we use our local world in \( \sim_{\text{II}} \) to express that for each continuation in the \( \text{I} \) program’s environment there is a corresponding continuation (code pointer and environment) on the machine stack. This correspondence is not trivial, as it must in turn describe the assumptions that code generated for continuations makes. For instance, such code assumes that before it gets executed the stack is popped as described at the beginning of the section.

### 7.4 Extraction

Our Coq development contains a script that extracts Pilnsr as OCaml code and couples it with code for parsing command-line arguments as well as a lexer and parser for the source language. In order to execute target machine code, we have implemented a single-step interpretation function in Coq and proved that it conforms to the operational semantics. This function is extracted to OCaml and wrapped in a loop.

Pilnsr provides command-line flags to selectively disable optimizations (more precisely, one flag per IL transformation). Accordingly, the extracted Pilnsr function takes not only a source module as argument but also a selection, a record of booleans indicating for each transformation whether it is to be performed. We have proven that Pilnsr is correct for any such selection and thus for any combination of compiler flags (not shown in Theorem 3).

### 8. Discussion and Related Work

#### Proof of transitivity. Recall the statement of Theorem 4. We actually derive this as a corollary of two properties:

**Lemma 1** (Transitivity, decomposed).

\[
\begin{align*}
[\Gamma'] & \vdash M_T \sim_{\text{II}} M_1 \Gamma' \quad \Gamma' & \vdash M_1 \sim_{\text{II}} M_2 : \Gamma' \\
[\Gamma] & \vdash M_T \sim_{\text{II}} M_2 : \Gamma' \\
[\Gamma] & \vdash M_1 \sim_{\text{II}} M_2 : \Gamma' \\
[\Gamma'] & \vdash M_T \sim_{\text{II}} M_2 : \Gamma' \\
[\Gamma'] & \vdash M_1 \sim_{\text{II}} M_2 : \Gamma'
\end{align*}
\]

Note that we do not need to show transitivity of \( \sim_{\text{II}} \) itself because we can simply iterate (2).

Thanks to our uniform setup, the proofs of (1) and (2) mirror each other. They also closely follow the transitivity proof of the original PB model [8]. However, since our language “in the middle” is untyped, the complexity having to do with abstract types can be avoided. On the other hand, due to our asymmetric small-step formulation of \( \text{E} \) and the possibility of stuttering, the part of the proof dealing with \( \text{E} \) becomes significantly more tricky.

As for PBs, one of the main complexities in the PILS transitivity proof lies in dealing with an ambiguity regarding reference allocation: while in one of the two given proofs, an allocation of the middle stage. The core property we need to show is that calls of \( k_f \) is bound.

### 8. Discussion and Related Work

#### Proof of transitivity. Recall the statement of Theorem 4. We actually derive this as a corollary of two properties:

**Lemma 1** (Transitivity, decomposed).

\[
\begin{align*}
[\Gamma'] & \vdash M_T \sim_{\text{II}} M_1 \Gamma' \quad \Gamma' & \vdash M_1 \sim_{\text{II}} M_2 \Gamma' \\
[\Gamma] & \vdash M_T \sim_{\text{II}} M_2 \Gamma' \\
[\Gamma] & \vdash M_1 \sim_{\text{II}} M_2 \Gamma' \\
[\Gamma'] & \vdash M_T \sim_{\text{II}} M_2 \Gamma' \\
[\Gamma'] & \vdash M_1 \sim_{\text{II}} M_2 \Gamma'
\end{align*}
\]

Note that we do not need to show transitivity of \( \sim_{\text{II}} \) itself because we can simply iterate (2).

Thanks to our uniform setup, the proofs of (1) and (2) mirror each other. They also closely follow the transitivity proof of the original PB model [8]. However, since our language “in the middle” is untyped, the complexity having to do with abstract types can be avoided. On the other hand, due to our asymmetric small-step formulation of \( \text{E} \) and the possibility of stuttering, the part of the proof dealing with \( \text{E} \) becomes significantly more tricky.
of the IL that makes the choice of public vs. private allocation explicit in the program code. It seems to us that this approach only makes sense if one is willing to a priori decide on all subsequent optimizations. The issue is that a later added optimization might, for instance, figure out that some reference is never used and therefore can be removed. Such can only be proven correct if the reference was allocated as private, which it may not have been. For the sake of modularity, we therefore believe it is better to bite the bullet and deal with the ambiguity issue semantically, e.g., in the way we did.

**Parametric bisimulations.** PBs [7] were inspired by a number of different methods for relational reasoning about higher-order stateful languages: notably, Kripke logical relations [4], normal form bisimulations [21], and environmental bisimulations [22, 19]. From Kripke logical relations, PBs adapted the mechanism of possible worlds as state transition systems, enabling the enforcement of protocols on local or global state. From normal form bisimulations, PBs took the central idea of viewing unknown functions as black boxes—in particular, the CALL case in Figure 8 is highly reminiscent of a similar case in normal form bisimulations. From environmental bisimulations, PBs borrowed the treatment of abstract types and polymorphism via type names (which we have glossed over here).

The key advance of PBs was to show how to combine all these mechanisms in a way that supported transitive composition and did not rely on “syntactic” devices employed by the other higher-order simulation methods (e.g., modeling related unknown functions as a common free variable [21], or using context closure operations [22, 19]), because such syntactic devices would preclude a generalization to inter-language reasoning. But it was far from obvious whether PBs would necessarily fare any better in this regard.

In this paper, we have demonstrated through PILS that the claims of Hur et al. [7] were indeed correct, and that PBs do in fact generalize to inter-language reasoning as promised.

**Multi-language semantics.** Motivated by the goal of supporting compiler verification for programs that interoperate between different languages, Perconti and Ahmed [17] propose an approach based on multi-language semantics [14]. In particular, they define a “big-tent” language that comprises the source, target, and intermediate languages of a compiler, and provides “wrapping” operations for embedding terms of each language within the others. They then use logical relations to prove that every source module is contextually equivalent to a suitably wrapped version of the target module to which it is compiled. In this way, their method synthesizes the benefits of logical relations (modularity and different source and target languages) and contextual equivalence (transitivity).

One downside of their approach is that the intermediate languages (ILs) used in a compiler show up explicitly in the statement of compiler correctness. This leads to a loss of modularity: the semantics of source-level linking is not preserved when linking the results of compilers that have different ILs. Another downside concerns flexibility: the approach seems to be restricted to compilers that use typed intermediate and assembly languages, and has only so far been applied to a purely functional source language. On the other hand, their goal is in a different sense more ambitious: they aim to say something about what it means to link arbitrary typed assembly code with ML code, whereas we only focus on linking assembly code that refines some source-level counterpart.

**Compositional verified compilation for C.** Motivated by the goal of compositional compiler verification, Beringer et al. [2, 20] propose an adaptation of the CompCert framework based on a novel operational semantics that differentiates between internal (intra-module) and external (inter-module) function calls. They introduce a notion of “structured simulation” that assumes little about the memory transformations performed by external function calls.

Beringer et al.’s approach is transitive, but lacking somewhat in flexibility. It depends on compiler passes only performing a restricted set of memory transformations—permitting additional transformations can potentially break the transitivity property. In addition, their method appears to be geared specifically toward compilers in the style of CompCert, which employ a uniform memory model across source, intermediate, and target languages. It is not clear how to generalize their technique to support richer (e.g., ML-like) source languages, or compilers whose source and target languages have different memory models.

Wang et al. [23] have also recently explored compositional compiler verification for a restricted C-like language called Cito. Their approach embeds the verification statement within a Hoare logic for assembly; further work is needed to understand the relationship between this approach and traditional refinement-based verification.

**References**


